文章编号：2095-2163(2023)02-0020-09 中图分类号：TP391.1 文献标志码：A

融合位置特征的关键短语集合抽取模型

于子健，孙海春，李 欣
（中国人民公安大学 信息网络安全学院，北京 100038）

摘 要：关键短语抽取任务是文本知识抽取任务的基础性工作，存在关键短语抽取边界不清晰、抽取结果重复率较高等问题，导致抽取结果准确性不佳。本文针对关键短语出现在文章中的位置特征建模，基于 Transfomer 编码器-解码器结构，结合位置特征与预训练模型对关键短语进行预测，提出一种端到端的关键短语预测模型；在模型训练过程中，采用了基于匈牙利算法对预测值与真实值进行序列对应的交叉熵损失函数，使关键短语预测过程，排除序列生成方法中预定排序的影响，并以集合的方式抽取关键短语。分别在 Inspec、SemEval2017、KP20k 数据集进行了实验验证，与现有方法相比较，本文模型 F_1 值均有所提升，有助于提升文本信息的关键短语抽取效果。

关键词：关键短语抽取；位置特征；知识抽取；编码器-解码器；预训练模型

Key phrase set extraction model based on position feature fusion

YU Zijian, SUN Haichun, LI Xin
(College of Information Network Security, People's Public Security University of China, Beijing 100038, China)

【Abstract】Key phrase extraction is a fundamental task in text knowledge mining, but the current task still suffers from unclear boundaries of key phrase extraction and high repetition rate of extraction results, resulting in poor accuracy of extraction results. An end-to-end key phrase prediction model based on Transfomer encoder-decoder structure backbone is proposed, which combines location feature and pre-trained model to predict key phrase. A cross-entropy loss function using Hungarian algorithm for permutation between predictions and ground truth is applied in training process to enable the key phrase prediction process to exclude the effect of predetermined ordering in sequence generation methods and to extract key phrases as a set. The model is validated on Inspec, SemEval2017, and KP20k datasets respectively. The F_1 - scores of the model are all improved compared with existing methods, which helped to improve the key phrase extraction of textual information.

【Key words】key phrase extraction; position feature; knowledge extraction; encoder-decoder; pre-trained model

0 引 言

关键词(Keyword)是篇章内容的高度概括，关键短语(Keyphrase)是关键词的拓展，内容包含关键词，能够简洁地表达更多主题信息，在英文领域中关键短语抽取是更常见任务。关键短语摘要作为自然语言处理的一项基础任务，是文本检索、文本摘要等文本挖掘任务的基础性的工作，可分为关键短语抽取技术(Extractive Keyphrase Extraction)与关键短语生成技术(Abstractive Keyphrase Generation)[1]。准确的专业领域文献关键短语简洁的呈现了文章涉及的领域和关键技术点，不仅有利于文献快速阅读，而且对相关文献推荐和领域研究现状掌握也能起到促进作用。

目前关键短语抽取常用的方法包括基于无监督的特征建模方法和监督序列标注任务方法等。其中特征建模方法需要专家知识确定候选词、打分方式，不同领域的迁移成本较高；序列标注任务方法在训练过程中侧重考虑关键短语的整体信息，边界处与内部权重相同，导致关键短语边界处容易出现预测错误，限制了抽取效果。

为了提高关键短语抽取效果，针对边界抽取准确率较低问题，应强化位置特征，增加针对关键短语边界的训练，并且依据全文位置对文本中每个词增加全局特征，从而增加低频词、关键字的特征表示。据此提出一种基于预训练语言模型的编码器-解码器(encoder-decoder)关键短语预测模型，该模型根

基金项目：国家重点研发计划(2020AAA0107700)；国家自然科学基金(62076246)；公安部技术研究计划项目(2020JSYJC22)。
作者简介：于子健(1997-)，男，硕士研究生，主要研究方向：自然语言处理、知识抽取、知识图谱；孙海春(1985-)，女，博士，讲师，主要研究方向：知识图谱、知识抽取；李 欣(1977-)，男，博士，教授，CCF 会员，主要研究方向：自然语言处理、网络安全。
通讯作者：李 欣 Email：lixin@ ppsuc.edu.cn
收稿日期：2022-06-29

据位置信息预测关键短语中关键字的位置，并将关键字的位置特征与全局语义信息融合，通过提示学习微调预训练语言模型完成文档关键短语的抽取。

本文提出融合预训练语言模型与位置特征的关键短语抽取模型，强化关键短语边界预测，提高边界抽取准确率，且增加低频词的全局位置特征信息，缓解低频词训练样本、语义表示不丰富的问题；通过序列到序列模型实现端到端的从原文本预测关键短语，减少对专家知识的依赖，在针对新领域、新文本风格时能够通过机器学习的方式自动调整迁移模型；以集合方式得到预测关键短语，通过无序的方式对模型预测值与真实值进行对应训练，使模型在训练过程中排除预定序列顺序的影响。

在 Inspec、SemEval2017、KP20k 数据集上的 $F_1@5$、$F_1@10$、$F_1@M$ 结果平均提升 1.2%，4.7%，1.5%，验证了位置特征在优化此任务的可行性。

1　相关工作

关键短语抽取常用的方法可分为两大类：基于无监督或有监督的特征建模方法、基于深度学习模型的关键短语抽取算法。

1.1　无监督方法

无监督方法通过量化表示词的重要度抽取关键词，无须标注语料并具有较高普适性，分为基于统计的方法、基于主题模型的方法和基于图的方法[2]。无监督方法常对文本特征建模，先通过词性、词频—逆文档频率（Term Frequency－Inverse Document Frequency，TF-IDF）等规则从文章筛选出候选词；之后根据定义的指标得到每个候选词的分数，选择高分预测作为模型的输出结果[3]。此类候选关键短语-排序方法，需对文档中的关键短语特征进行充分调查研究后，制定筛选候选词的规则与对候选词打分排序的方法，此过程会使用较多的专家知识，效果相较于传统的统计方法有一定的提升。

Luhn[4-5] 在 1957 年提出最早的基于统计思想的关键短语抽取方法，并在 1958 年指出利用位置特征抽取关键信息的可行性；Ricardo Campos 等[6] 提出了一种基于文章词频、出现位置等多种统计文本特征的无监督关键短语抽取方法，取得了更优的效果。在基于图思想的 TextRank 算法基础上，Xiong 等[7] 针对常见关键短语提取算法中低频词易被忽略的问题，在 TextRank 算法的基础上，根据单词之间的语义差异进行聚类分析，并利用聚类结果计算词图中边缘的权重并调整过渡概率矩阵，迭代计算单词的最终权重，并执行排序以获得关键短语；Wu[8] 将词的频率特征和位置特征合并为字节点的初始权重，对 TextRank 进行改进。

无监督方法基于特征建模，制定候选短语筛选规则、打分规则，对关键短语的不同特征充分建模，可解释性较强，但对专家知识依赖程度较高，且针对不同领域需要相应调整规则，模型迁移的成本较高。

1.2　深度学习方法

随着深度学习模型的发展，相关研究将深度学习方法应用于文本关键短语抽取任务。利用深度学习模型提取关键短语，首先得到文本段的语义向量表示，输入定义的深度学习模型，根据模型预测的关键短语结果与真实关键短语的差异对模型调整优化[3]。利用深度学习模型可以减少对专家知识的依赖，让定义的模型根据数据样本，自动学习关键短语在文本中的隐藏含义，端到端预测文本关键短语。基于深度学习方法实现关键短语抽取有两种经典方法，一是将文本关键短语抽取视为序列标注任务，二是构建原文的词字典，逐词抽取文本关键短语。

序列标注模型能够充分获取文档上下文的相邻语义，在命名实体识别任务上取得了较好效果，将其应用到关键短语抽取任务，提高了抽取效果，但序列标注模型存在边界错误问题，与命名实体相比，关键短语定义标准不统一，不具有明显自然边界特征，因此序列标注模型在抽取关键短语时边界处准确率较低，限制了整体抽取效果[9]。而且关键短语预测需要综合全局语义，对上下文语义依赖距离较长，应用序列标注模型预测时，会将文本每个单词分割，标注时关注局部语义信息，降低全局的语义信息在预测关键短语时的权重，影响关键短语抽取效果。

逐词抽取文本关键短语方法将关键短语抽取任务转换为文本生成任务。此类方法构建文档的总词表，通过对文本整体建模，采用序列到序列（seq2seq）等模型从总词表中逐词抽取，训练过程中将每一次预测结果作为约束条件，动态生成、组合得到关键短语，提升关键短语预测效果，此方法打破了原文的固定顺序，能够预测原文中不存在的关键短语，解决了传统技术只能抽取原文存在的关键短语这一问题。Meng[10] 首先提出 CopyRNN 模型，将此方法应用到此任务中；Zhang[11] 在此基础上将 RNN 替换为训练速度更快的 CopyCNN 模型；Chen[12] 在 CopyRNN 基础上增加限制机制，提出 CorrRNN 模型，降低预测重复率，提升了模型效果。随着预训练语言模型的发展，Ding[13] 将 Bert 预训练模型和对抗生成网络结合，应

用在关键短语抽取任务中,通过预训练模型 Bert 获得高质量的文本表示,对抗性神经网络的应用缓解了监督算法需要大量注释数据的缺点;Wu[14] 基于提示学习方法,充分利用预训练模型的优势,降低训练成本的同时提升了关键短语预测效果。

将关键短语抽取任务视为文本生成任务,提高了预测效果,但是由于候选词表维度较大,难以预测到低频的专业词汇;且训练生成模型所需要的资源较多,训练时间较长。在模型逐词抽取的过程中会产生一个预定顺序的关键短语序列,即训练过程中模型预测的结果与真实值的损失会受到关键短语预定顺序的影响。关键短语应是无序的集合,关键短语的顺序不应成为必须施加于模型的限制,而传统的生成方法并未去除此影响因素。

2 文本关键短语抽取模型

关键短语应具备以下特征:

(1)关键短语出现的位置具有一定规则,可以通过位置特征抽取关键短语;

(2)关键短语的内容中更容易包含低频词,且原文中会明确出现该特定词汇;

(3)关键短语彼此间有联系,但无固定顺序,是无序集合的形式。

根据以上特征,将抽取关键短语任务拆解为以下步骤:

(1)根据文本语义特征预测文档中关键短语中的关键字所在位置,关键短语为待预测的最终结果,关键字为关键短语中的特征边界位置,如首字、尾字等;

(2)根据关键字前后的位置视野,融合其在全文的位置特征与语义信息,判断该关键字所在关键短语的范围,通过预训练语言模型对关键短语进行预测输出。

模型结构如图 1 所示,首先通过位置特征对预训练语言模型生成的语义向量表示进行注意力加权,得到融合位置与语义的文本向量表示,将其输入 seq2seq 模型中 transformer 编码器层,得到文本的隐含状态,再将全文隐含状态输入 transformer 解码器层,端到端得到文本关键字位置;最后基于预训练语言模型的提示学习,利用得到的位置特征以及对应的关键字文本语义信息,构建提示学习模板,对原文本内容进行注意力加权,预测对每个关键字位置所对应的关键短语内容,通过微调训练优化预训练语言模型,完成关键短语预测。

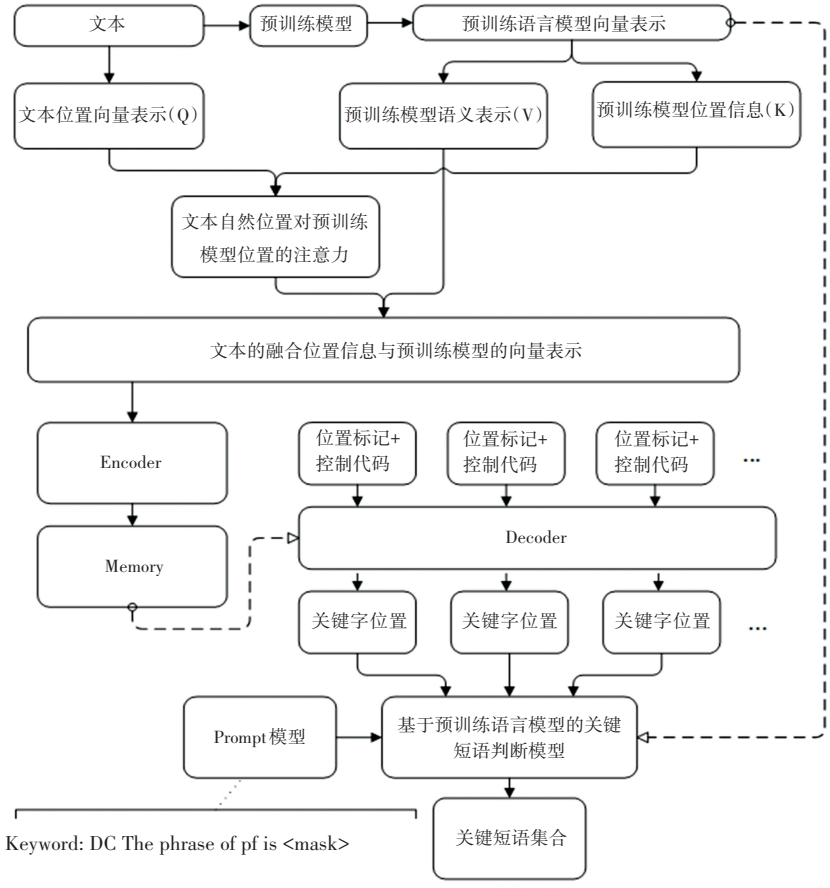


图 1 模型结构图

Fig. 1 Model structure

2.1　文本位置特征表示

预训练语言模型在对文本建模时会将输入文本中带有连字符的词、部分过长专业词汇以及超出预设词典的词汇，进行分解拆分，破坏原有的位置结构，例如：将“unwanted”分解成“un”、“##want”以及“##ed”，并不是按照自然分词输入，得到的语义表示长度往往会大于原文，导致每个词的语义表示不能与原文的词汇一一对应，无法直接得到对于原文每个位置单词的语义表示。

由于专业领域文本中存在较多专业特定词汇，经典预训练语言模型的训练语料中该样本数量较少，导致领域专业词汇的特征表示效果不理想，甚至超出其预设字典。而专业词汇更能反映文本的领域特征，因此关键短语中更倾向于出现这些专业词汇、低频词。针对专业词汇表示不充分问题，模型通过融合位置特征增强预训练模型对文本的向量表示。

本文使用 Beltagy [15] 在科学文献数据集预训练过的 Sci-Bert 模型，提供 768 维的预训练词向量，该预训练模型在专业领域更具针对性、更广泛的词汇字典，减少了模型未知词的数目，对于专业领域低频词的表示比经典模型更好。为了保证文本从预训练语言模型中得到的预训练语义表示适应此任务，模型训练过程微调（Fine-tuning）预训练语言模型，提供的文本的语义表示记为 emb_{plm}。

在此基础上增加基于全文位置的可学习位置特征，具体对文本中的每个词以及每个词在全文中的位置特征进行建模，将未经预训练的自然分词后的词向量 emb_{word} 与位置特征 $position_{emb}$ 结合，得到文本中每个位置的全文特征信息，记为 emb_{doc}，式（1）：

$$emb_{doc} = emb_{word} + position_{emb} \tag{1}$$

通过构建的全文位置特征，利用该位置向量对预训练语言模型提供的语义信息进行注意力加权结合，得到融合语义与位置特征的全文位置特征。首先，融合位置特征的 emb_{doc} 与预训练语言模型分词得到的 emb_{plm}，利用 Vaswani A [16] 提出的注意力机制获得全文位置特征对 Bert 等预训练向量的注意力，从而得到经过位置注意力加权的向量表示，获得融合预训练语言模型的语义特征的原文本的位置向量，通过公式（2）~公式（5）获得的文本位置向量表示 $emb_{position}$，融合了位置特征及预训练语言模型的语义，且文本每个位置的向量表示，不仅会得到预训练模型中对应位置的注意力加权表示，还会获得语义相似位置的向量表示信息，拓展了全局的位置特征、语义特征，从而增加每个位置的特征关注视野，能够获取更长的语义依赖信息。

$$Q = \boldsymbol{emb}_{doc}\boldsymbol{W}_{query} \tag{2}$$

$$K = \boldsymbol{emb}_{plm}\boldsymbol{W}_{key} \tag{3}$$

$$V = \boldsymbol{emb}_{plm}\boldsymbol{W}_{value} \tag{4}$$

$$\boldsymbol{emb}_{position} = \mathrm{Softmax}\left(\frac{QK^{\top}}{\sqrt{d}}\right)V \tag{5}$$

其中，$\boldsymbol{W}_{query}$，$\boldsymbol{W}_{key}$，$\boldsymbol{W}_{value}$ 为待学习的参数矩阵，生成的 $\boldsymbol{emb}_{position}$ 与预训练语言模型向量 $\boldsymbol{emb}_{plm}$ 的维度相同，d 为词向量维度。

2.2　序列到序列预测模型

将预测关键短语起始位置视为序列生成任务，将此任务视为序列标注问题，采取如 BiLSTM-CRF 模型对文本中元素进行逐个标注，往往存在边界不准确的问题，即整体的损失值较低，在边界处错误导致预测与实际不符。基于对此问题，将该任务视为类似阅读理解、摘要抽取任务，预测关键短语出现的起始位置与结束位置，将总体的损失转为具有边界针对性的损失。

构建序列到序列预测模型，输入为文档的向量表示 $\boldsymbol{emb}_{position}$，输出为文档的关键字位置。有 3 种位置采样规则：关键短语的首字、尾字以及中位字。为简化模型采用首、尾字的策略进行实验，模型接收 $\boldsymbol{emb}_{position}$ 后输出预测关键短语起始、终止位置集合，输出的取值范围为总文本长度空间。

以首字为例，将预测关键短语结束位置任务转换为根据关键短语起始位置预测关键短语长度任务。生成关键短语起始位置后，基于预训练语言模型，融合关键字位置特征与语义信息，从原文中预测关键字对应的整体关键短语。

预测关键字位置任务采用 seq2seq 模型，由编码器与解码器两层组成，由编码器层获得完整文本的隐含状态，将隐含状态传递给解码器层；解码器层根据前面隐含状态的向量表示与当前输出情况，得到预测的关键短语起始位置或终止位置。

模型编码器层由 6 层 transformer 组成，隐藏层维度为 512，分为 8 头注意力机制，输入为原文本的 $\boldsymbol{emb}_{position}$，该向量表示经过全连接层转化为隐藏层维度，融合可学习的位置权重，将融合预训练语言模型语义与位置特征的向量表示输入编码器层，得到文本隐藏状态。

解码器层主体与编码器层结构相同，在 transformer 层的输出后增加全连接层，用于得到预测值。对于解码器层第 i 时刻的输入 d_i 公式（6）：

$$d_i = \boldsymbol{emb}_{position}^{\begin{cases} <start>, i=0 \\ y_{i-1}, i\neq 0 \end{cases}} + code_{position}^{i} + code_{control}^{n} \quad (6)$$

其中，$\boldsymbol{emb}_{position}^{\begin{cases} <start>, i=0 \\ y_{i-1}, i\neq 0 \end{cases}}$ 为上一时刻解码器层的输出的向量表示；$code_{position}^{i}$ 为可学习的位置特征节点，$code_{control}^{n}$ 为第 n 个可学习的全局控制节点，用于减少重复预测。

seq2seq 模型得出的结果为有序输出，该顺序代表了模型预测的先后逻辑。传统方法中将 seq2seq 模型输出的关键短语直接对应匹配原文的关键短语顺序、或原文中关键短语出现的顺序，得到预测的损失，以此进行模型的反向传播学习。Ye[17] 指出文章的关键短语应是无序集合，目前通过序列生成模型预测关键短语额外要求机器学习该真实值的序列顺序，令模型的预测增加了预定顺序的影响，不符合关键短语为无序集合的特征，可能影响模型预测结果。针对此问题，训练过程中使用模型预测值与真实值的对应关系序列中最优期望序列，以此对应序列计算模型预测值与真实值的损失，更新模型参数。

首先，利用定义的 seq2seq 模型生成一组关键短语起始点的概率分布 *Prediction*。构建图 $G(V, E)$，其中 V 由两个独立的空间组成，分别为预测值 *Prediction* 与真实值 Y，E 代表每个预测值与真实值对应的代价矩阵，对任意 $p_n \in Prediction, y_n \in Y$ 的关系代价记为 $Cost(p_n, y_n)$，维度为预测数目 N ×真实数目 $|Y|$，式(7)：

$$E = \{Cost(p_n, y_n)\}_{N\prod|Y|} \quad (7)$$

目标是在预测值 *Prediction* 在全部排列情况 *Prediction*（记为 P）中，对于真实值 Y 的最小代价的排列情况，实际意义为找到模型预测值的集合，$\hat{P}^i$ 为最接近的真实答案集合分布，取 $\hat{P}^i$ 分布时此时预测值与真实值对应的分布对应的代价总计最小，式(8)：

$$\hat{P}^i = \underset{t\in\prod(N)}{\operatorname{argmin}} \sum_{n=1}^{N} Cost(P_n^t, y_n) \quad (8)$$

其中，P_n^t 为预测值 *Prediction* 在 P^t 排列下第 n 个取值。

计算过程采用匈牙利算法，匈牙利算法是一种在多项式时间内求解任务分配问题的组合优化算法。

构建匈牙利算法时，结合 CrossEntropy 损失函数的计算方式，该损失 L 经典计算公式为公式(9)、公式(10)：

$$L(Prediction, y) = L = \{l_1, \cdots, l_N\}^{\top} \quad (9)$$

$$l_n = -\sum_{c=1}^{C} \log \frac{\exp(x_{n,c})}{\sum_{i=1}^{C} exp(x_{n,i})} y_{n,c} \quad (10)$$

其中，$Prediction = \{p_1 \cdots p_N\}$，表示模型预测的输出；$y = \{y_1, \cdots y_N\}$，表示每个输出值对应的真实值；$C$ 表示每个为类别总数；N 表示预测总数。

得到真实值 Y 的全部空间 $C_Y\{c \mid Y\}$ 分布概率，预测值 P_n 在 P^t 的排列下记为真实值空间，对每个真实值空间 $C_Y\{c \mid Y\}$ 的概率分布记为 $\hat{P}_n^t$，式(11)：

$$\hat{P}_n^t = P_n^t\{n \mid n \in C_Y\} \quad (11)$$

损失矩阵中每个预测值 p_n 在 P^t 的排列下对应每个真实值 y_n 的损失 $Cost(p_n, y_n)$ 的值为 $\hat{P}_n^t$ 对 y_n 的概率以 e 为底取负对数值，式(12)：

$$Cost(p_n, y_n) = -\log(\hat{P}_n^t(y_n)) \quad (12)$$

通过匈牙利算法得到总代价最小的预测值对应情况，将预测值排序，排序得到的顺序变化通过矩阵变换方法逆向应用于真实值，得到最接近预测值的真实值序列，计算起始点的损失函数。预测数目多于真实数目时，将真实值补零填充至两者相同，令每个预测值对补零填充的真实值的 *Cost* 远大于平均值，使模型优先分配非补零填充的真实值，获得最优分布；真实数目多于预测数目时，对预测值补零填充至两者相同，令每个补零填充的预测值对应每个真实值的概率相同，即 *Cost* 值相同，对补零填充的预测值额外增加远大于平均值的损失，则模型在计算分布序列 *Cost* 时，排除多余预测值对总体的影响。

Ye[17] 率先提出了以集合方式预测文本的关键短语。本文基于其预测分布序列计算方法，根据交叉熵损失函数计算方式，在计算模型预测分布序列的 *Cost* 值之前增加了 Softmax 操作，计算时对每个预测与真实的损失值求负对数，使分布序列 *Cost* 值能更好地体现该序列在交叉熵损失函数中的表现，令每个控制节点对应下的预测值与真实值有更高的期望，在迭代过程中始终保持对应，缓解了模型训练初期的预测值与真实值对应混乱问题，提高了模型训练的收敛速度。本方案对于模型预测数目没有要求，预测的数目可以少于真实数目，减少特异样本的干扰，更灵活地控制模型预测的关键短语数目，减少模型训练时间、降低学习成本。

2.3 提示学习模型

模型的训练过程采取了多任务训练思路，分为两部分：第一部分是预测关键字位置，第二部分是每

个关键字位置对应关键短语的预测。

对于前文得到的文本关键字，可以获取每个关键字在全文中的位置，通过此位置特征对每个关键字对应的关键短语预测。本模型基于提示学习(Prompt Learning)的思想，将前文得到的关键字位置作为特征，构建提示学习模板(Prompt)，通过预训练语言模型的微调，优化获得每个关键字对应的预测关键短语。

Wu[14]通过 Bert 类预训练语言模型的 MLM(Mask Language Model)任务构建预测关键短语模板，即“phrase of kw is [MASK] [MASK] kw [MASK] [MASK]”，其中 kw 为输入的关键信息。本模型基于 Liu[18]提出的离散提示学习模板，采用 Raffel[19]提出的更适合文本生成任务的 T5 预训练模型，该模型为编码器-解码器结构的文本到文本(text-to-text)模型。融合 Wu [14]和 Gao[20]构建的提示学习模板策略，构建本模型提示学习模板为“Keyword : DOC The phrase of pf is <mask>”，其中 pf 为预测的文本关键字位置，DOC 为原文本内容，筛选其中存在 pf 对应文字的句子作为输入，以此构建生成预测结果的模板。设置最大预测长度为 6，解码时不考虑特殊符号，对预测结果去除停用词并进行词干提取，得到输出内容。在以上条件下构建提示学习模板，对预训练模型 T5-base 进行提示学习微调，得到最优预测结果。

3　实验

3.1　数据集

实验使用的数据集为 Inspec，SemEval2017，KP20K 等，实验数据集为互联网上公开获取。Inspec 数据集由2 000篇期刊论文摘要及其关键词组成，包含 1998 年~2002 年中计算机与控制、信息技术学科等领域论文；KP20K 由567 830篇计算机科学领域的论文组成，选取其中20 000篇作为验证集，20 000篇作为测试集；SemEval2017 由 493 篇科学领域论文组成，为国际语义评测大赛(Semantic Evaluation)2017 年任务 10 提供的关键短语识别数据集。

数据集均可分为两部分，完整文档以单个字符串存储，对应的若干关键短语字符串以列表形式存储，关键短语中包含原文本中直接存在的以及原文本中不存在的关键短语，本文仅对原文本中直接存在的关键短语进行抽取。每个数据集均筛选存在原文出现关键短语的样本，随机采样 50%计算数据集特征，数据集的基本情况见表 1。

表 1　数据集情况

Tab. 1　Dataset information

数据集	数目	平均每篇的关键短语数目	关键短语的平均词长度	原文中明确存在的关键短语数目
Inspec	2 000	14.03	2.19	7.76
SemEval2017	493	17.30	3.01	16.51
KP20K	570 809	5.42	2.04	2.53

3.2　数据处理与评估标准

本文使用 NLTK(Natural Language Toolkit)工具集对数据集进行预处理，具体包括：去除占位符等无实意符号，对文章进行分词，分句，英文文本全部转换为小写等。

针对完整文本段与关键短语，利用 NLTK 分词工具得到单词级别的原文本内容，利用正则表达式匹配等方法得到关键短语在文章中单词级别的位置特征；将关键字位置与对应关键短语存于抽取关键短语列表中，得到分词后的完整文本段、抽取关键短语列表(关键字位置与对应关键短语内容)。

数据预处理过程参照 Meng[21]的处理方式，将文本中数字统一用 < digit > 替换、使用 Porter Stemming 策略对结果提取词干，在评估矩阵(Evaluation Metrics)上选择 $F_1@5$、$F_1@10$、$F_1@M$ 作为评估标准，$F_1@5$ 与 $F_1@10$ 分别为计算前 5 个预测结果、前 10 个预测结果的 Micro F_1 - *score*，当预测数不足 5 或 10 时填充错误答案，$F_1@M$ 为计算所有预测结果的 Micro F_1 - *score*。Micro F_1 - *score* 的计算如式(13) ~ 式(15)：

$$Recall = \frac{\sum_{c=1}^{C} TP_c}{\sum_{c=1}^{C} (TP_c + FN_c)} \tag{13}$$

$$Precision = \frac{\sum_{c=1}^{C} TP_c}{\sum_{c=1}^{C} (TP_c + FP_c)} \tag{14}$$

$$Micro - F_1 = 2\frac{Recall \times Precision}{Recall + Precision} \quad (15)$$

其中，TP 为将正例预测为正类的数目；FN 为将正例预测为负类的数目；FP 为将负例预测为正类的数目；C 为所有待预测关键短语集合。

本文模型输出的关键短语为无序结果，在计算 $F_1@5$、$F_1@10$ 时，按照控制节点依次选取作为预测结果顺序。

3.3 对比实验

实验环境 GPU 为 Rtx 3060 12 G，Pytorch 版本 1.7.0。批处理大小设置为 16，最大训练步设置为 100 000步，热身学习步数设置为 3 个 *Epoch*，学习率为从 10^{-7} 增加至 10^{-4} 后开始降低，优化器为 AdamW，本模型采用关键短语首字位置特征，控制节点数目设置为 20。

KP20k 训练集经过去除重复样本、去除超过预训练语言模型允许文本长度样本，去除空关键短语样本，训练集共421 970 条，模型随机选取其中 100 000条数据进行训练，在各个数据集取验证结果最好的模型进行测试。

KP20k 验证集为15 849条，测试集为15 937条；Inspec 数据集去重、去除空关键短语样本后共1 956 条，选取1 500条作为验证集，456 条作为测试集；SemEval2017 数据集经处理共 478 条，选 300 条作为验证集，178 条作为测试集。测试时对输入文本中超出预训练模型最大允许长度的文本内容，采取截断措施。

论文的对比模型为深度学习模型 CatSeq、CatSeqD [22] 与 ExHiRD-h、ExHiRD-s[23]，对比模型与本文模型均在相同条件下进行训练，训练集、验证集与测试集设置相同。

对比实验结果见表 2，本模型相较于目前的关键短语抽取模型，在 Inspec、SemEval2017、KP20k 数据集上的 $F_1@5$、$F_1@10$、$F_1@M$ 结果平均提升 1.2%，4.7%，1.5%。

表 2 实验结果

Tab. 2 Experimental results

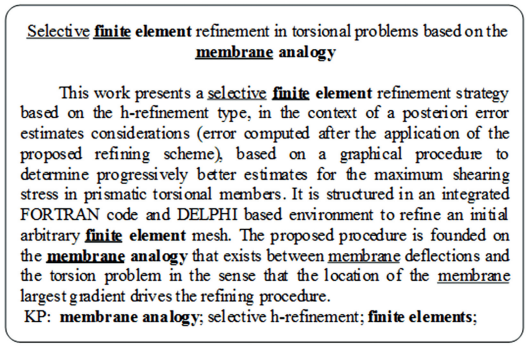
模型	Inspec			SemEval2017			KP20K		
	$F_1@5$	$F_1@10$	$F_1@M$	$F_1@5$	$F_1@10$	$F_1@M$	$F_1@5$	$F_1@10$	$F_1@M$
CatSeq	0.185	0.127	0.242	0.138	0.104	0.162	0.224	0.140	0.301
CatSeqD	0.157	0.107	0.217	0.136	0.102	0.166	0.208	0.130	0.288
ExHiRD-h	0.225	0.152	0.262	0.175	0.132	0.195	0.250	0.157	0.304
ExHiRD-s	0.207	0.144	0.252	0.157	0.119	0.178	0.238	0.162	0.296
Our Model	**0.248**	**0.208**	**0.286**	**0.176**	**0.159**	**0.207**	**0.263**	**0.222**	**0.312**

3.4 位置特征选择与全局控制节点数目

关键字位置表达了关键短语位置特征，全局控制节点能够提供额外限制特征减少预测重复，同时也能控制预测输出的数目。实验过程中对不同位置特征的选择与全局控制节点数目进行测试，以确定更优设置。

数据集中的某样例的节选如图 2 所示，文中加粗的内容为关键短语，下划线为关键短语的首字在文中出现的情况。由图 2 可知关键短语中的首字、尾字在原文中出现的位置不仅是关键短语中，且该关键字出现的位置周边为关键短语相关内容的概率较大，因此通过对全文所有关键字的位置特征进行特征建模，能够构建全文中关键短语的位置、相邻语义等特征信息，更好的获得关键短语的隐含特征向量。

实验过程中对关键短语的首字与尾字位置抽取效果进行对比，针对全局控制节点的数目进行实验，考虑到更多控制节点会引入更多错误预测结果，根据数据集中每篇文章关键短语数量选择控制节点候选数目为 10 与 20，评价指标为关键字位置抽取的召回率，召回率越高，则代表此时位置特征的获取效果越好。


Selective **finite element** refinement in torsional problems based on the **membrane analogy**

This work presents a selective **finite element** refinement strategy based on the h-refinement type, in the context of a posteriori error estimates considerations (error computed after the application of the proposed refining scheme), based on a graphical procedure to determine progressively better estimates for the maximum shearing stress in prismatic torsional members. It is structured in an integrated FORTRAN code and DELPHI based environment to refine an initial arbitrary **finite element** mesh. The proposed procedure is founded on the **membrane analogy** that exists between membrane deflections and the torsion problem in the sense that the location of the membrane largest gradient drives the refining procedure.

KP: **membrane analogy**; selective h-refinement; **finite elements**;

图 2 关键字在文章中的分布(首字)

Fig. 2 Key word in document(First word)

在 KP20K 数据集下以不同参数设置实验，寻找

最佳模型，模型训练过程中对 KP20K 验证集进行关键字位置抽取的结果如图 3 所示，示例 S-10、E-20 中，S 代表首字，E 代表尾字，数字代表全局控制节点数目，S-10 表示以首字为关键字，全局控制节点数目设置为 10。由图 3 可知，全局控制节点数目对于预测召回率影响较大，能显著提升抽取效果；在控制节点数目取 20 时，验证集中首字与尾字的抽取效果差距不大；虽然 KP20K 验证集差距不大，但测试集中效果差距明显，可知尾字特征的普适性不如首字。

再针对不同数据集对关键短语预测模型进行测试，实验结果见表 3，$R@10$ 为控制节点数目为 10 时对应关键字位置抽取结果的召回率，$R@20$ 为控制节点数目为 20 时相应的召回率，召回率反映了覆盖全文关键位置特征的程度。3个数据集在对首字作位置特征抽取、控制节点数目取 20 时关键字位置抽取结果召回率最高，此时能够获取更广泛覆盖全文的位置特征，从而更好的抽取文中存在的关键短语。

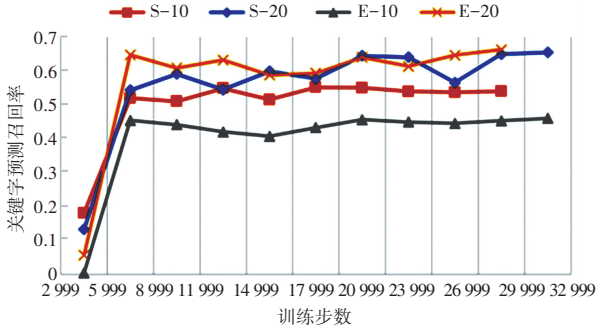


图 3　KP20K 验证集关键字位置预测

Fig. 3　Key word position prediction in KP20K Valid Dataset

表 3　关键字位置特征抽取结果

Tab. 3　Key word position feature extraction result

模型	Inspec		SemEval2017		KP20K	
	$R@10$	$R@20$	$R@10$	$R@20$	$R@10$	$R@20$
首字	0.390	0.479	0.297	0.406	0.530	0.680
尾字	0.347	0.473	0.265	0.405	0.459	0.608

4　结束语

目前文本关键短语抽取任务的结果受限于边界准确率，且模型训练过程受到真实值预定序列影响。本文模型融合预训练语义信息与位置特征，构建针对关键短语边界关键字的编码器-解码器模型，强化对关键短语边界的抽取训练，缓解了边界抽取效果限制问题，提升了整体准确率。以集合的方式抽取关键短语，通过匈牙利算法获得预测值-真实值键值对的无序集合，排除了预定序列对抽取结果的影响。在与 CatSeq、ExHiRD-h 等模型的对比实验中，本模型抽取结果 F_1 值有提高，验证了将位置特征与预训练语言模型结合进行关键短语抽取方法的有效性。

实验过程中发现，基于位置特征的关键短语预测模型在面对长文本数据时效果不佳，在面对词数超过 1 000 的文本时准确率降低明显。经分析，在提取位置特征时，长文本会增加过多关键字相邻语义与关注的信息，降低位置特征的信息密度，且预训练模型允许的长度有限，不能完整获取过长文本的语义信息，导致影响特征获取质量。下一步将研究长文本位置特征表示方法，以提高模型对长文本数据的抽取效果；研究自动化构建关键短语抽取任务的提示学习模板以提升模型的可迁移性。

参考文献

[1] 胡少虎，张颖怡，章成志. 关键词提取研究综述[J]. 数据分析与知识发现，2021，5(3)：45-59.

[2] 王昊，刘丹，刘硕. 基于句法分析及主题分布的关键词抽取模型[J]. 计算机应用研究，2022：1-6.

[3] 于强，林民，李艳玲. 基于深度学习的关键词生成研究综述[J]. 计算机工程与应用，2022：1-16.

[4] LUHN H P. A statistical approach to mechanized encoding and searching of literary information[J]. IBM Journal of Research and Development，1957，1(4)：309-317.

[5] LUHN H P. The automatic creation of literature abstracts[J]. IBM Journal of Research and Development，1958，2(2)：159-165.

[6] CAMPOS R，MANGARAVITE V，PASQUALI A. YAKE! keyword extraction from single documents using multiple local features[J]. Information Sciences，2020，509：257-289.

[7] XIONG A，LIU D，TIAN H. News keyword extraction algorithm based on semantic clustering and word graph model[J]. Tsinghua Science and Technology，2021，26(6)：886-893.

[8] WU C，LIAO L，AFEDZIE KWOFIE F. TextRank keyword extraction method based on multi-feature fusion[C]. YANG X-S，SHERRATT S，DEY N，et al.，eds.//Proceedings of Sixth International Congress on Information and Communication Technology. Singapore：Springer，2022：493-501.

[9] HE Z，WANG Z，WEI W. A survey on recent advances in sequence labeling from deep learning models[J/OL]. ArXiv：

2011.06727 [Cs], 2020[2022-04-22]. http://arxiv.org/abs/2011.06727.
[10] MENG R, ZHAO S, HAN S. Deep keyphrase generation[C/OL]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver, Canada: Association for Computational Linguistics, 2017: 582-592[2021-12-15]. http://aclweb.org/anthology/P17-1054.
[11] ZHANG Y, FANG Y, WEIDONG X. Deep keyphrase generation with a convolutional sequence to sequence model[C/OL]//2017 4th International Conference on Systems and Informatics (ICSAI). Hangzhou: IEEE, 2017: 1477-1485[2021-12-15]. http://ieeexplore.ieee.org/document/8248519/.
[12] CHEN J, ZHANG X, WU Y, et al. Keyphrase generation with correlation constraints[J/OL]. ArXiv: 1808.07185 [Cs], 2018[2022-01-03]. http://arxiv.org/abs/1808.07185.
[13] DING T, YANG W, WEI F, et al. Chinese keyword extraction model with distributed computing[J]. Computers & Electrical Engineering, 2022, 97: 107639.
[14] WU H, MA B, LIU W, et al. Fast and constrained absent keyphrase generation by prompt-based learning[J]. 2022.
[15] BELTAGY I, LO K, COHAN A. SciBERT: a pretrained language model for scientific text[J/OL]. arXiv: 1903.10676 [cs], 2019[2022-05-13]. http://arxiv.org/abs/1903.10676.
[16] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J/OL]. ArXiv: 1706.03762 [Cs], 2017[2022-04-19]. http://arxiv.org/abs/1706.03762.
[17] YE J, GUI T, LUO Y, et al. One2Set: generating diverse keyphrases as a set[J]. 2021.
[18] LIU P, YUAN W, FU J, et al. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing[J/OL]. ArXiv: 2107.13586 [Cs], 2021[2022-01-30]. http://arxiv.org/abs/2107.13586.
[19] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer: arXiv: 1910.10683[Z/OL]. arXiv, 2020(2020-07-28)[2022-05-25]. http://arxiv.org/abs/1910.10683.
[20] GAO T, FISCH A, CHEN D. Making pre-trained language models better few-shot learners[J/OL]. ArXiv: 2012.15723 [Cs], 2021[2022-03-31]. http://arxiv.org/abs/2012.15723.
[21] MENG R, YUAN X, WANG T, et al. Does order matter? an empirical study on generating multiple keyphrases as a sequence[J/OL]. ArXiv: 1909.03590 [Cs], 2022[2022-04-21]. http://arxiv.org/abs/1909.03590.
[22] YUAN X, WANG T, MENG R, et al. One size does not fit all: generating and evaluating variable number of keyphrases[J/OL]. ArXiv: 1810.05241 [Cs], 2020[2022-01-14]. http://arxiv.org/abs/1810.05241.
[23] CHEN W, CHAN H P, LI P, et al. Exclusive hierarchical decoding for deep keyphrase generation[J/OL]. ArXiv: 2004.08511 [Cs], 2020[2022-01-14]. http://arxiv.org/abs/2004.08511.

(上接第 19 页)

3 结束语

本文针对遥感影像背景复杂多样、检测目标小且不清晰等问题进行了试验研究。在 YOLOv4 框架基础上,提出了一种改进的 YOLOv4 的目标检测算法。由实验结果可知,改进的 YOLOv4 目标检测算法,满足实时性检测的需求。但本文算法在遥感飞机小目标检测上仍有不足之处,一是训练样本的不足,所用的样本背景环境不够复杂,导致其检测率高于实际应用到遥感领域中真实值,二是在面对图像中飞机目标排列密集的情况下,仍存在漏检现象。

参考文献

[1] 史瑞鹏,蒋丹妮,方青. 基于 YOLOv4 的遥感影像飞机目标检测[J]. 测绘通报,2021(S1):134-138.
[2] 王瑶,胥辉旗,姜义,等. AI 目标检测网络应用研究[J]. 兵器装备工程学报,2021,42(6):236-242.
[3] 秦伟伟,宋泰年,刘洁瑜,等. 基于轻量化 YOLOv3 的遥感军事目标检测算法[J]. 计算机工程与应用,2021,57(21):263-269.
[4] 公明,刘妍妍,李国宁. 改进 YOLOV3 的遥感图像舰船检测方法[J]. 电光与控制,2020,27(5):102-107.
[5] 国腾飞,张则言,付宏财,等. 融合注意力机制的轻量级红外高压套管识别算法[J]. 计算机与现代化,2022(1):70-76.