文章编号：1003-0077(2021)08-0064-09

基于句子选择的关键短语生成

罗益超，李争彦，张 奇

（复旦大学 计算机科学技术学院，上海 200433）

摘 要：关键短语生成是一个能从长文档或者文献中捕获中心思想的实用任务。先前的神经关键短语生成方法基本只注重词级别的信息而忽略文档结构。该文提出了一个句级选择网络（sentence selective network，SenSeNet）用于关键短语生成。该模型重点关注文档的句子结构信息，通过学习句子隐式表示来判断其是否有可能生成关键短语，然后根据判断结果引入对应归纳偏置来辅助解码器生成关键短语。该文使用直通估计量（straight-through estimator）来端到端地训练模型。为了提高句级选择网络性能，该文还提出了一个任务强相关的弱监督信息。实验表明，模型成功地捕获了文档信息，并合理选择了相对较重要的句子，而且模型也更倾向于从这些重要句子中生成关键短语。该文将模型引入到绝大多数序列到序列模型中，在五个数据集中的两个评价指标下，均有显著的性能提升。

关键词：关键短语生成；文档结构；直通估计量；弱监督

中图分类号：TP391 **文献标识码**：A

Neural Keyphrase Generation with Sentence Selective Network

LUO Yichao[1], LI Zhengyan[1], ZHANG Qi[1]

(School of Computer Science, Fudan University, Shanghai 200433, China)

Abstract: Keyphrase Generation (KG) is the task of capturing themes from a document, revealing the key information necessary to understand the content. Existing neural keyphrase generation approaches focus only on the token-level information while ignore sentence-level information such as document structure. In this paper, we incorporate the sentence-level inductive bias into KG and propose a new method named Sentence Selective Network (SenSeNet), which can automatically learn the sentence-level information and determine whether the sentence more likely to generate the keyphrase. We use straight-through estimator to train the model in an end-to-end manner and incorporate a weakly-supervised setting which is helpful for the training of the sentence selection module. Experiments show that our model successfully captures the document structure and reasonably distinguishes the significance of sentences, and consistent improvements achieved on two metrics in five datasets.

Keywords: keyphrase generation; document structure; straight-through estimator; weakly-supervised

0 引言

关键短语生成是一个在自然语言处理中传统并具有挑战的任务。它可以通过生成的关键短语捕获文档的中心思想。当模型生成一个简洁的输出后（关键短语），我们可以很方便地应用于下游任务。例如，文本分类（text categorizing）[1]、文本摘要（summarization）[2-3]和意见挖掘（opinion mining）[4]等。

在关键短语生成中，根据关键短语是否存在于原文中可以分为抽取式（present）关键短语和生成式（absent）关键短语。传统的方法[5]致力于生成抽取式关键短语，所以这类方法也被称为关键短语抽取。最近很多研究致力于同时生成两种关键短语。Meng 等[6]提出了一种基于注意力机制（attention-based）的序列到序列（Seq2Seq）框架[7]，并引入了拷贝机制（copy mechanism）[8]，从而有效地生成稀有词。Chen 等[9]提出了一种结合覆盖机制（coverage

收稿日期：2020-11-20 **定稿日期**：2020-12-28

基金项目：国家重点研发计划（2017YFB1002104）

mechanism)[10]和回顾机制(review mechanism)的模型，从而生成更多样化的关键短语。Chen 等[11]提出了一种利用标题信息去指导生成关键短语的模型。Chan 等[12]利用结合自适应奖励函数的强化学习方法去提升模型效果，并生成更多数量的关键短语。

所有这些方法都基于一个序列到序列(Seq2Seq)的框架，它把整个输入文档当成一个序列，平等地对待文档中的每个单词。然而，文档中信息的重要性是不同的。例如，对于科学性论文的摘要，一般包括目的(purpose)、方法(methodology)、实验发现(findings)、实验价值(value)等几句话，而像“目的”、“方法”中很大概率有关键短语，但“实验发现”、“价值”中很少有关键字。因此，利用文档结构去建模句子级信息可以有效减少无效信息，并且能让模型更加致力于重要的部分。

本文提出了一个新颖的方法，叫作句级选择网络(sentence selective network，SenSeNet)，其可以自动收集每一个句子的信息并用隐变量隐式地表示句子特征，从而判断这个句子是否倾向于生成关键短语，然后根据二值信号(0、1)引入对应的归纳偏置到原本的序列生成框架中辅助模型生成。在句级选择网络中，需要将平滑连续的句子表示值转化成离散型变量(0、1)，这会造成梯度无法回传的问题。所以我们引入直通估计量[13](straight-through estimator，STE)方法来分别处理前向传播和反向传播，从而使模型可以端到端训练。另外，为了保证句级选择网络的性能和效率，本文提出了一个与任务强相关的弱监督信号来监督句级选择网络。

- **Electric Power Steering** System Based on LQR Techniques
- **摘要**：①This paper aims to present a linear quadratic regulator (LQR) employed to improve performance of an **electrical power steering** (EPS) system. ②Generally, EPS is a full electric system having an electrical motor which provides the assist torque on the steering mechanism in order to reduce the workload and to enhance the steering feel of the driver during the steering process. Since the torque sensors are considerably expensive, the authors present a control strategy that eliminates the driver torque sensor by introducing a torque estimator. Three main technical areas are described in this paper. First, the principle and structure of EPS are presented including the dynamic model. Second, LQR and Kalman filter techniques are employed to derive an **optimal controller** for the EPS system. Finally, the simulations and hardware results are depicted.③The combined tools of Matlab/Simulink and dSPACEprovide the environment for modelling the controller in software and applying it to the actual hardware via a digital signal processing board based on the DS1401 MicroAutoBox. The controller is evaluated via simulation results, dSPACEhardware results, and verified on vehicle testing data. ④This paper presents a controller design for an EPS system based on the LQR techniques. Within the controller concept shown, elimination of the driver torque sensor offers advantages in terms of both cost and mechanical performance. Simulations and measured data prove the good functionality of the controller proposed.
- **关键短语**：**electric power steering**, **linear quadratic regulator**, **optimal control**, robust control, electric motors

图 1　关键短语生成的一个例子

注：粗体单词是 present keyphrase，下划线单词是 absent keyphrase。其中摘要被分为四部分：目的(purpose)、方法(methodology)、发现(findings)、价值(value)。

本文在五个数据集上做了实验。实验表明，句级选择网络能够较为准确地选择相对重要的句子，并且在抽取式关键短语和生成式关键短语两部分都有一定的性能提升，尤其是在生成式关键短语部分。为了进一步分析实验结果，本文提出了一个新概念——半抽取式关键短语(semi-present keyphrase)。另外，本文还做了通用性实验，实验表明，该模型在各种序列生成模型中均能发挥有效的作用。

1　相关工作

目前的关键短语生成方法主要分为传统的关键短语抽取(keyphrase extraction)和近期的关键短语生成(keyphrase generation)。

传统的关键短语抽取方法[14-15]直接抽取文中重要片段作为关键短语，它一般分为两步：候选词抽取和排序。在第一阶段，首先使用一些启发式方法提取一些候选词，例如，词性标注(part-of-speech tags)[16]和 N 元模型(N-gram)[17]，然后将这些候选词依据成为关键短语的可能性被有监督[18]或者无监督[19]的方法排序。传统方法的缺点是不能生成文中不存在的短语。

为了解决这个问题，meng 等首先提出了一个基于 Seq2Seq 模型[20]——CopyRNN。它结合了注意力机制[7]和拷贝机制[8]。这个工作提出了一个大规模数据集 KP20k，并将关键短语生成引入深度学习时代。此后，许多方法或者机制被引入到这个框架中用来提高性能。CorrRNN[9]引入覆盖机制[10]和回顾机制来缓解重复和冗余的问题，使模型生成更加多样的关键短语。TG-net[11]充分利用了标题信息并把标题作为一个额外信息去指导原文编码，取得不错的效果。

以上深度学习的方法是基于 one2one 模式的，具体做的时候，模型会将文档和每一个关键短语一一对应地放入模型中训练，最后用 beam-search 进行解码，选取概率最高的 k 个关键短语作为结果，所以这些方法生成的关键短语数量是固定的。Yuan 等[21]提出两个方法：catSeq 和 catSeqD，可以生成不同数量的多个关键短语。它的做法是将一个文档对应的多个关键短语拼接在一起，用分隔符<seq>分开，以结束符<eos>结尾，拼接后作为一个序列。在解码阶段时，生成的关键短语到结束符为止。这种模式被称为 one2many。最近 Chan 等[12]结合一个自适应的奖励函数的强化学习方法提升模

型性能并生成数量更多的关键短语(由于之前的catSeq生成的平均数量小于标签的平均数量)。另外这篇文献中提出了一种结合维基百科语料的评价指标缓解同义词未被合理评价的问题。

本文的工作参考了Yuan等[21]的工作,也是one2many模式,一次生成不同数量的关键短语。和TG-net模型相似,本文的模型也致力于从文中的重要部分抽取关键短语,但是区别于TG-net只关注标题信息,本文的模型更加关注所有重要句子。

2 主要方法

2.1 问题定义

在本节中,将对关键短语生成给出一个正式的问题定义。给定一个源文档 x,我们的目标是生成多个真实标签关键短语 $Y=\{y_1,\cdots,y_{|Y|}\}$。其中,源文档 $x=(x^1,\cdots,x^{|x|})$ 和目标关键短语 $y_i=(y_i^1,\cdots,y_i^{|y_i|})$ 都是单词序列,$|x|$ 和 $|y_i|$ 代表源文档序列 x 和第 i 个关键短语 y_i 的长度。关键短语被分为抽取式关键短语(present keyphrase)和生成式关键短语(absent keyphrase)。本文提出了一个新概念:半抽取式关键短语(semi-present keyphrase)。这类关键短语是所有单词都出现在了一个特定句子中,但不像抽取式关键短语那样可以直接从原文中抽取。本文用 $y^p=\{y_1^p,\cdots,y_{|y^p|}^p\}$,$y^a=\{y_1^a,\cdots,y_{|y^a|}^a\}$ 和 $y^s=\{y_1^s,\cdots,y_{|y^s|}^s\}$ 分别去定义这三种关键短语。根据定义,$y^s\subseteq y^a$。表1给出了三种不同关键短语的具体例子。其中,一个字符代表一个单词,SOURCE代表源文档。

表1 三种不同关键短语的例子

SOURCE	A B C D E F G H I J
present	"A B C","E F G","H I J",etc.
semi-present	"A B D","B C A","A D H",etc.
absent	"A B D","B C A","X Y Z",etc.

2.2 基于句子选择的编码器

传统的编码器(encoder)模块首先将输入序列 $x=(x_1,x_2,\cdots,x_T)$ 做词嵌入(embedding),转换成 $e=(e_1,e_2,\cdots,e_T)$,然后经过一个编码器层,如双向门控循环单元(GRU)[22]或者Transformer[23]得到隐状态表示 $\boldsymbol{H}$,如式(1)所示。

$$\boldsymbol{H}=\text{Encode}(e) \tag{1}$$

在句子选择模块中,本文使用卷积神经网络(CNN)对句子进行分类[24],将之前得到的序列词嵌入表示 e 按照句号进行分离。第 i 个句子 S_i 如式(2)所示。

$$S_i=e_{\pi_i}\oplus e_{\pi_i+1}\oplus\cdots\oplus e_{\pi_i+|S_i|-1} \tag{2}$$

其中,$|S_i|$ 表示句子 i 的长度,$\oplus$ 表示单词信息之间的拼接,π_i 表示第 i 个句子开始的位置。然后,我们用CNN来收集每个句子信息。句子 S_i 的特征被一个窗口长度为 k 的卷积核给压缩。第 i 个句子的第 j 个新特征 c_i^j 计算如式(3)、式(4)所示。

$$c_i^j=\sigma(\boldsymbol{W}\cdot e_{\pi_i+(j-1)k:\pi_i+jk-1}+b) \tag{3}$$

$$\boldsymbol{c}_i=[c^1,c^2,\cdots,c^{|S|-k+1}] \tag{4}$$

其中,b 是一个偏置项,σ 是一个非线性激活函数。然后用max-pooling操作取出 c_i 中的最大值 $c\bar{Y}_i=\max\{c_i\}$ 作为句子表示,利用 $c\bar{Y}_i$ 计算一个二值表示来决定这个句子是否更倾向于生成关键短语,如式(5)~式(7)所示。

$$m_i=\text{MLP}(c\bar{Y}_i) \tag{5}$$

$$\eta_i=\text{sigmoid}(m_i^{\mathrm{T}}\boldsymbol{w}_{m_i}) \tag{6}$$

$$z_i=\begin{cases}1, & \eta>0.5\\ 0, & \eta\leqslant 0.5\end{cases} \tag{7}$$

其中,MLP是多层感知机,sigmoid是非线性激活函数,$\boldsymbol{w}_{m_i}$ 是一个可训练的权重向量。z_i 是一个二进制控制门,决定句子 i 是否重要(1表示重要,0表示不重要)。

为了将这个重要性信息反馈给模型,将被判断重要的句子中每个单词都标记1,判断为不重要的句子中的每个单词都标记0,用向量 $\boldsymbol{g}$(长度 T)表示。然后用一个embedding矩阵 $\boldsymbol{D}\in\boldsymbol{R}^{1\times 2d}$ 对向量 $\boldsymbol{g}$ 做一个词嵌入,然后得到一个和隐状态 $\boldsymbol{H}$ 形状相同的矩阵 $\boldsymbol{G}$,将 $\boldsymbol{G}$ 和 $\boldsymbol{H}$ 进行同纬度相加得到最终的隐状态 $\boldsymbol{F}$,如式(8)、式(9)所示。

$$\boldsymbol{G}=\boldsymbol{g}^{\mathrm{T}}*\boldsymbol{D} \tag{8}$$

$$\boldsymbol{F}=\boldsymbol{H}+\boldsymbol{G} \tag{9}$$

其中,d 是隐状态的维度,因为是双向GRU所以需要 $2d$。

2.3 解码器

在解码器中,用编码器得到的最后隐状态 $\boldsymbol{F}$ 结合注意力机制计算上下文向量 $\boldsymbol{u}$。时间 t 的上下文向量 $\boldsymbol{u}_t$ 根据式(10)计算:

$$\boldsymbol{u}_t=\sum_{j=1}^{T}\alpha_{tj}\boldsymbol{F}_j \tag{10}$$

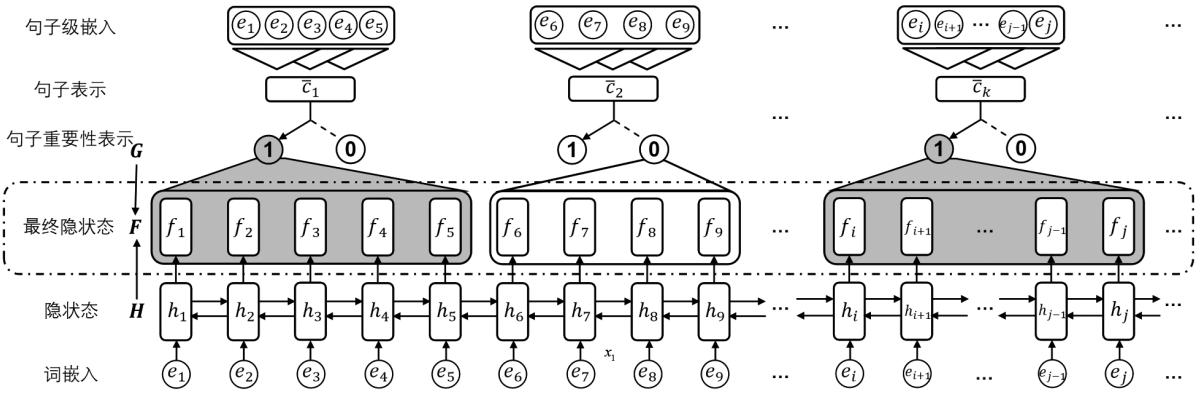


图 2　基于句子选择的编码器模块

注：底部的圆圈表示词嵌入，顶部的圆圈表示按照句子划分后的词嵌入。句子表示为词嵌入经过卷积网络后得到的。句子重要性为句子表示得到的 01 标签。隐状态为源词序列经过 encoder 得到的。最终隐状态为两个同纬度隐状态相加。

其中，α_{tj} 表示源序列中第 j 位置和输出第 t 个位置的相关性。

在上下文 $\boldsymbol{u}_t$ 的帮助下，解码器可以更好地通过传统的语言模型生成单词序列，如式(11)、式(12)所示。

$$\boldsymbol{s}_t = \text{Decode}(y_{t-1}, \boldsymbol{s}_{t-1}, \boldsymbol{u}_t) \tag{11}$$

$$p_g(y_t \mid y_{<t}, \boldsymbol{F}) = \text{Softmax}(y_{t-1}, \boldsymbol{s}_t, \boldsymbol{u}_t) \tag{12}$$

其中，$\boldsymbol{s}_t$ 代表解码器在 t 时刻时的隐状态，Softmax 可以得到生成词表中每个单词的概率，y_t 表示 t 时刻的输出。

由于 out-of-vocabulary(OOV)，模型不能生成稀有词。因此我们引入拷贝机制到解码器中用来直接从源文档中拷贝稀有词。所以最终生成一个单词的概率被分为两部分——生成和拷贝，如式(13)～式(15)所示。

$$p(y_t \mid y_{<t}, \boldsymbol{F}) = p_g(y_t \mid y_{<t}, \boldsymbol{F}) + p_c(y_t \mid y_{<t}, \boldsymbol{F}) \tag{13}$$

$$p_c(y_t \mid y_{<t}, \boldsymbol{F}) = \frac{1}{Z} \sum_{j:\, x_j = y_t} \mathrm{e}^{\omega(x_j)}, y_t \in \chi \tag{14}$$

$$\omega(x_j) = \sigma(\boldsymbol{F}_j^{\mathrm{T}} \boldsymbol{W}_c) \boldsymbol{s}_t \tag{15}$$

其中，χ 代表源文 $\boldsymbol{x}$ 中稀有词的集合，$\boldsymbol{W}_c$ 是一个可学习的参数矩阵，Z 被用于归一化。

2.4　训练过程

2.4.1　标准训练过程和 STE 优化

为了训练模型，最小化负似然对数按照式(16)计算：

$$L_{\text{MLE}} = -\sum \log p(y_t \mid y_{<t}, \boldsymbol{F}) \tag{16}$$

然而，在编码器模块中，模型为句子生成了一个二进制离散型表示[见式(7)]，这样就造成模型不连续和梯度无法回传的问题。为了解决这个问题，一个通用的方法就是利用策略梯度(policy gradient)从语言模型 $p(y_t \mid y_{<t}, x)$ 中采样一组单词，然后环境就会给出一个奖励函数(这里就是评价指标 F_1 值)。然而，策略梯度有众所周知的问题——花费高和方差大，这就导致模型非常难训练。

在这个工作中，受之前训练不连续的神经网络的工作[13,25]的影响，利用直通估计量(straight-through estimator)去估计二值表示的梯度。对于解码器的一个特定参数 θ，估计梯度按照式(17)计算：

$$\begin{aligned} &\frac{\mathrm{d}E[\sum_t \log p(y_t \mid y_{<t}, \boldsymbol{F})]}{\partial \theta} \\ &= \frac{\mathrm{d}E[\sum_t \log p(y_t \mid y_{<t}, \boldsymbol{F})]}{\partial z} \frac{\mathrm{d}z}{\mathrm{d}\eta} \frac{\mathrm{d}\eta}{\partial \theta} \\ &= \frac{\mathrm{d}E[\sum_t \log p(y_t \mid y_{<t}, \boldsymbol{F})]}{\partial z} \frac{\mathrm{d}z}{\mathrm{d}\eta} \partial \theta \end{aligned} \tag{17}$$

用一个连续并且平滑的函数 η 来估计之前不连读的离散变量 z。本方法在正向传播的时候还是传递 0、1 离散值，而在反向传播的时候用 η 的梯度去估计 z 的梯度。尽管这个估计有一个偏差，但是它在估计离散变量时非常高效。

2.4.2　弱监督训练

由于目前提出的模型只接收源文档和目标关键短语来作为训练指导，而其中句子的重要性特征(0、1)需要模型自己来学习，这会使模型训练得非常困难和缓慢。另外，如果缺乏这部分的监督信息，模型学习到最后很有可能将句子选择层全部学成 1 或者全部学成 0，这样句级选择网络就失去了其作用，模型就退化成了标准的 Seq2Seq 模型。调研以往的

工作可以发现,大部分正确的关键短语都是被直接从源文档中抽取出来的,我们认为存在抽取式关键短语的句子相对比较重要,受 Zhou 等[26]将弱监督应用于 NLP 工作启发,我们为每一个句子提出了一个弱监督信号 $a_i \in \{0,1\}$,用它来描述这个句子是否是重要的,其中 1 表示重要,0 表示不重要。信号设定来源为:若这个句子中存在抽取式关键短语或者半抽取式关键短语,则标记为 1,反之标记为 0。然后将这个弱监督信号添加到模型中,并用 BCE loss(Binary Cross-Entropy loss)进行监督,如式(18)所示。

$$L_{\text{BCE}} = -\sum_i a_i \log \eta_i + (1-a_i)\log(1-\eta_i) \tag{18}$$

因此,最后的损失函数如式(19)所示。

$$L = L_{\text{MLE}} + \lambda L_{\text{BCE}} \tag{19}$$

其中,λ 是一个超参。

3 实验部分

3.1 数据集

本文在五个公开科学性关键短语生成数据集上做了实验:Inspec[17]、NUS[18]、Krapivin[27]、SemEval-2010[28]、KP20k[6]。

其中,KP20k 中有约 50 万条数据,比其他数据集规模都大许多,所以下文会以该数据集为重点进行分析。

3.2 实现细节

遵循 Chan 等[12]工作的实验设计,在预处理过程中,将原文中所有数字用<digit>代替,将 present keyphrase 和 absent keyphrase 用<peos>分割,并且将 present keyphrase 按照文中出现顺序进行排序。在生成关键短语时,将重复的关键短语删除。另外,为了控制变量来证明句级选择器的有效性,所有 baseline 参数设置都和 Chan 等人的工作保持一样,如词表大小设为 50 000,隐层大小设置为 150。

另外,还有一些文中所提出模型的独特的细节,如设置 embedding 矩阵 $\boldsymbol{D}$ 中的纬度 d 为 150,这和隐层的维度一样大,符合模型的设计;卷积神经网络的卷积核大小设置为{1,3,5},通道数为 100;BCE 损失函数的权重系数 λ 经过实验调试,在 0.08 附近模型结果最优。

3.3 Baseline 模型和评价指标

由于传统方法性能远远低于深度学习的方法,所以本文只考虑基于 Seq2Seq 的模型。参考 Chan 等人[12]的工作,选取四个前人提出的模型并沿用了他们取的模型名:结合拷贝机制的 catSeq[21]、可生成不同数量的 catSeqD[21]、结合覆盖机制和回顾机制的 catSeqCorr[9]、结合标题信息的 catSeqTG[11],其中,后三个模型都是往第一个模型中添加机制或者方法。本文提出的模型也是在 catSeq 中添加方法,所以其性能也是相对于 catSeq 的。另外,本文尝试使用不同的编码层(如 Transformer、LSTM 等)来验证提出的模型的通用性。

先前的一些工作[6,11]从 beam-search 的结果中切取概率最高固定个数(如 5 或 10)的关键短语作为结果计算评价指标,如 F_1@5 和 F_1@10。为了更好地评价可变数量的关键短语,Yuan 等[21]提出了一个新的评价指标,F_1@M。它通过比较所有预测的关键短语和真实标签来计算 F_1。

本文生成可变数量的关键短语,所以使用 F_1@5 和 F_1@M 两个评价指标来分别评价 present keyphrase 和 absent keyphrase。其中需要注意的是,由于生成可变数量的关键短语有可能不足 5 个,所以计算 F_1@5 时,需要将生成不足 5 个关键短语的例子用随机错误的关键短语补全。

4 实验分析

4.1 Present 和 absent 关键短语预测分析

本文分别评测了模型在 present 和 absent keyphrase 上的性能。不同模型在 present keyphrase 上的评测结果如表 2 所示。可以看到,句级选择网络的性能在几乎所有数据集上都超过了四个基准数据集,除了在 Krapivin 数据集上略低于 catSeqTG 模型。事实上,本文提出的模型相对 catSeq 性能有一定的提升。经统计分析,Krapivin 数据集中关键短语更多地集中于标题中,而 catSeqTG 则更加关注标题信息,所以结果会比本文模型略好一点。

表 2　Present keyphrase 在五个数据集上的预测结果

Model	Inspec		NUS		Krapivin		SemEval		KP20k	
	$F_1@M$	$F_1@5$	$F_1@M$	$F_1@5$	$F_1@M$	$F_1@5$	$F_1@M$	$F_1@5$	$F_1@M$	$F_1@5$
catSep	0.262	0.225	0.397	0.323	0.354	0.269	0.283	0.242	0.367	0.291
catSeqD	0.263	0.219	0.394	0.321	0.349	0.264	0.274	0.233	0.363	0.285
catSeqCorr	0.269	0.227	0.390	0.319	0.349	0.265	0.290	0.246	0.365	0.289
catSeqTG	0.270	0.229	0.393	0.325	**0.366**	**0.282**	0.290	0.246	0.366	0.292
SenSeNet	**0.284**	**0.242**	**0.403**	**0.348**	0.354	0.279	**0.299**	**0.255**	**0.370**	**0.296**

Absent keyphrase 上的结果展示在表 3 中。可以观察到，句级选择网络在所有数据集上都取得了最好的结果，并且性能得到了显著的提升。

表 3　Absent keyphrase 在五个数据集上的预测结果

Model	Inspec		NUS		Krapivin		SemEval		KP20k	
	$F_1@M$	$F_1@5$	$F_1@M$	$F_1@5$	$F_1@M$	$F_1@5$	$F_1@M$	$F_1@5$	$F_1@M$	$F_1@5$
catSep	0.008	0.004	0.028	0.016	0.036	0.018	0.028	0.020	0.032	0.015
catSeqD	0.011	0.007	0.024	0.014	0.037	0.018	0.024	0.014	0.031	0.015
catSeqCorr	0.009	0.005	0.024	0.014	0.038	0.020	0.026	0.018	0.032	0.015
catSeqTG	0.011	0.005	0.018	0.011	0.034	0.018	0.027	0.019	0.032	0.015
SenSeNet	**0.018**	**0.010**	**0.032**	**0.018**	**0.046**	**0.024**	**0.032**	**0.024**	**0.036**	**0.017**

Present 和 absent keyphrase 取得如此显著的成果，证明了我们提出的句级选择网络能有效利用句子信息，过滤不重要信息，使模型更加集中于重要的信息。

4.2　Semi-present 关键短语预测分析

在本节中，为了验证前一节提出的观点，我们分析了句级选择网络在 semi-present keyphrase 上的性能。根据统计，semi-present keyphrase 占所有 keyphrase 的 7.9%，而 absent keyphrase 占所有的 41.3%。又根据定义 semi-present 属于 absent keyphrase，所以 semi-present 占 absent 中的很大一部分。因此，若能改进 semi-present 的准确率，就能很大程度提升 absent 的性能。

表 4 中比较了所有基准模型在 semi-present keyphrase 上的性能。由于现有的模型在生成 absent keyphrase 上 F_1 值非常低，所以我们选取正确预测的关键短语数量和 Recall 值作为评价指标，来分析 semi-present keyphrase 和 absent keyphrase 在除去前者剩下部分(用 absent w/o 表示)中的效果。可以发现，不管是 semi-present keyphrase 还是 absent keyphrase，模型 SenSeNet 的性能都有很大提升。特别是在 semi-present keyphrase 上，SenSeNet 相对于基准模型在召回率上提升了 0.6%，同时在 absent w/o 上提升 0.4%，所以能在 absent 关键短语上取得性能的显著提升。

表 4　Semi-present keyphrase 和 absent w/o keyphrase（除去 semi-present）在 KP20k 数据集上的结果

Model	semi-present		absent w/o	
	# count	Recall	# count	Recall
catSeq	398	0.053	637	0.020
catSeqD	384	0.051	674	0.021
catSeqCorr	408	0.054	676	0.021
catSeqTG	390	0.052	681	0.021
SenSeNet	**440**	**0.059**	**763**	**0.024**

4.3　句子数量对预测的影响

在本节中，重点分析了文档句子数量对 SenSeNet 预测关键短语性能的影响。本文将测试数据集按照文档句子数量分为 5 份，然后计算每一部分中 SenSeNet 相对于 catSeq 提升了多少。评价指标选取了 $F_1@5$ 和 $F_1@M$，present 和 absent 分别展示，数值为提升的百分数，结果如图 3 所示。

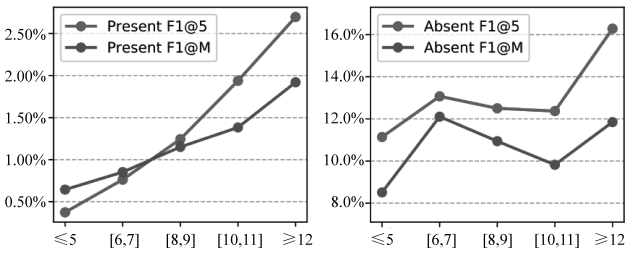


图 3 不同句子数性能相对提升

注：SenSeNet 相对于 catSeq 在不同句子数量的测试数据集中性能提升的比例。

可以看到，不管是 present keyphrase 还是 absent keyphrase，随着句子数量增加，SenSeNet 相对于基准模型提升的百分数不断提升。这说明 SenSeNet 在随着文档句子数量变多时，仍然能有良好的性能，确实具有良好的去噪能力。另外可以看到在 absent keyphrase 中，模型能提升的百分数远远大于在 present keyphrase 中，也说明 SenSeNet 在 absent keyphrase 中特别有效。

本文统计显示：平均句子数量为 7.6，平均“重要”句子数量（弱监督信号标记为 1 的）为 3.9。其中，“重要”的句子占比为 53.4%，也说明如果能从原文中选取重要的句子，会对生成关键短语非常有效。而实际预测过程中，表现最好的模型，句级选择网络选取了 54.1%的句子作为重要句子，这个比例和弱监督信号中的比例十分接近。另外，本模型中抽取的 76%的关键短语都来自重要句子。这说明我们模型成功地捕获了句子重要性的文档结构，并且成功地将这个归纳偏置引入到序列生成模型中。

4.4 用例分析

本节进行了用例分析。图 4 用一个具体的例子比较了 catSeq 和 SenSeNet 生成的关键短语。其中，在真实标签中，粗体单词是 present keyphrase（原文中有对应），下划线单词是 absent keyphrase，特别地，这个下划线单词属于 semi-present keyphrase（原文中灰底字为来源）。原文中第 1 句和第 4 句为模型实际选择的重要句子。

文档: ①Intelligent **decoupling control** of power plant main steam pressure and power output. ②An intelligent decoupling control strategy has been proposed and successfully applied to a 300 MW boilerturbine unit, i.e. Unit 1 of Yuanbaoshan Power Plant in China. ③For the strong couplings between control loops of main steam pressure and power output, a new design for decoupler aimed at decoupling for set-points and unmeasured pulverized coal disturbance of the system at the same time is presented. ④For the variation of operating condition and slowly varying dynamics, an **intelligent control** scheme has been developed by integrating **fuzzy reasoning** with **adaptive control and auto-tuning techniques**. ⑤Satisfactory industrial application results show that such a control system has enhanced adaptability and robustness to the complex process, and better control performance and high economic benefit have been obtained.
真实标签: decoupling control, intelligent control, fuzzy reasoning, adaptive control and auto tuning, power plant control
catSeq 预测: **decoupling control**, power plant, power output, control system
SenSeNet 预测: decoupling control, fuzzy reasoning, adaptive control, power plant control

图 4 对比 SenSeNet 和 catSeq 输出的用例分析

注：第①和第④句是被 SenSeNet 判断为重要的句子。粗体单词表示 present keyphrase，下划线单词表示 absent keyphrase（特别的在这个例子中也是 semi-present keyphrase），方框单词表示错误预测的关键短语但是存在于文档中。第①句中灰色部分单词表示 semi-present keyphrase 的来源。

对比 catSeq 和 SenSeNet 的输出可以发现，catSeq 只对了一个 present keyphrase，而 SenSeNet 则相对于 catSeq 多对了两个 present keyphrase 和一个 absent keyphrase。分析原因，SenSeNet 将这两个单词“fuzzy reasoning”和“power plant control”所在的句子给判断为重要的，所以模型能以更大可能从这两个句子中选出关键短语。相反，catSeq 比 SenSeNet 多预测了一个错误单词“control system”，而这个单词存在的句子被 SenSeNet 判断为不重要的，所以模型降低了从这个句子中选取关键词的概率，最终没有错误预测这个单词。这也说明句子重要性判断对关键短语预测有很大帮助。

4.5 可视化分析

在本节中对模型生成关键短语时的 attention 分布做了可视化分析。如图 5 所示，三张热力图分别表示同一句话在 catSeq、SenSeNet 加入重要的偏置后（1）、SenSeNet 加入不重要的偏置后（0）三种情况下，每个单词在解码时 attention 数值之和。可以理解为数值越高，被抽取为关键短语的概率越高。其中，粗体的几个词是真实关键短语标签。

在 catSeq 中，真实关键短语中的几个单词概率虽然相对于同句中其他单词高一点，但是还是很低，

个别单词如“time”甚至为 0，所以这几个单词不容易被抽出来作为输出。而当句级选择网络判断这句话为重要，并加入了 1 的偏置后，可以发现，这几个词的概率变得非常高，很容易被抽取出来作为关键短语。在图 5 的第 3 行 Sen SeNeto 中，当加入 0 的偏置后，这个句子中，连原本概率相对比较高的两个单词“short”和“fractional”的数值也变得很低，变得更加不容易被抽取。图 5 有效说明了 SenSeNet 对句子加入“重要”的偏置和“不重要”的偏置处理都十分有效。

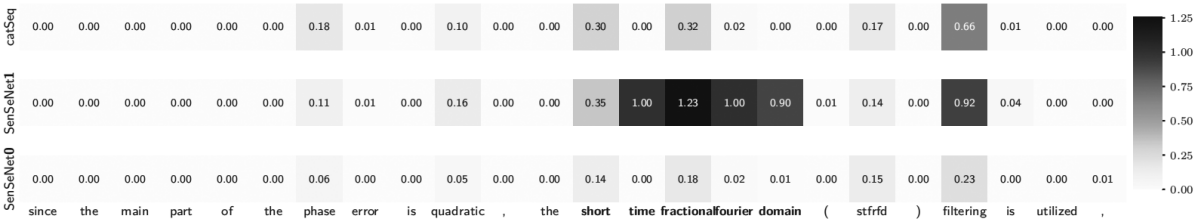


图 5　同一句话中所有单词被预测的概率在三种不同情况下的可视化分析

4.6　通用性分析

在本节中，重点对模型的通用性做了分析，比较了多个 Seq2Seq 框架在 KP20k 数据集上有无添加 SenSeNet 的结果，如表 5 所示。

表 5　通用性对比实验

Model	present		absent	
	w/o SSN	w/SSN	w/o SSN	w/SSN
TF	0.273	0.278	0.015	0.016
LSTM	0.289	0.293	0.015	0.016
GRU	0.291	0.296	0.015	0.017

表 5 中，评测指标是 F_1@5，其中“TF”表示 Transformer。从表 5 中可以看出，无论选用哪种 Seq2Seq 框架，SenSeNet 在 present 和 absent keyphrase 上都有显著的提升。这说明本文提出的模型是十分通用的。我们将 SenSeNet 能非常容易地移植入不同的框架并十分有效的原因总结为以下三点：①句级选择网络是一个独立运营的模块，可以单独收集句子信息并给出判断；②使用“1”和“0”的词嵌入表示作为句子级的归纳偏置，而不是使用比较“硬”的“0、1”门控（直接将不重要的信息舍去），这样能让模型更加平滑，并且能保留更多的原始语义信息，从而能使模型更好地运行；③加入配套且和任务强相关的弱监督信号能更好地指导模型进行句子选择。

表 5 中有一个奇怪的现象，就是当今比较流行且具有并行性的 Transformer 框架的结果不如相对传统的 GRU 和 LSTM。事实上，不仅关键字生成中有这个问题，其他的任务中也同样，如命名实体识别（NER）[29]。Yan 等[30]分析了这个现象，并且总结了两个原因：①标准 Transformer 不关注方向和距离信息；②Transformer 的注意力分布其实相对比较平滑，但是在 NER 中，注意力分布比较稀疏，因为只有少数词是比较重要的。这两点在关键短语生成中也同样存在。

5　结语

本文提出了一个新颖的模型句级选择网络（SenSeNet）用于关键短语生成。它可以自动评估文档中的每个句子是否更倾向于生成关键短语并向模型中引入偏置。本文使用直通估计量（STE）解决模型不连续问题，并引入弱监督信号指导模型更好地选择重要的句子。实验表明，句级选择网络能有效地捕获文档结构信息，同时有效地选择倾向于生成关键短语的句子，最后在抽取式关键短语和生成式关键短语中都有显著效果。

参考文献

[1] Hulth A, Megyesi B. A study on automatically extracted keywords in text categorization[C]//Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, 2006: 537-544.

[2] Wang L, Cardie C. Domain-independent abstract generation for focused meeting summarization[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, 2013: 1395-1405.

[3] 马亮，何婷婷，李芳，等. 以关键词抽取为核心的文摘句选择策略[J]. 中文信息学报，2008，22(6)：50-54.

[4] Berend G. Opinion expression mining by exploiting keyphrase extraction[C]//Proceedings of the IJCNLP, 2011.

[5] Liu Z, Huang W, Zheng Y, et al. Automatic keyphrase extraction via topic decomposition[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2010: 366-376.

[6] Meng R, Zhao S, Han S, et al. Deep keyphrase generation[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017: 582-592.

[7] Luong M T, Pham H, Manning C D. Effective approaches to attention-based neural machine translation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2015: 1412-1421.

[8] Gu J, Lu Z, Li H, et al. Incorporating copying mechanism in sequence-to-sequence learning[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016: 1631-1640.

[9] Chen J, Zhang X, Wu Y, et al. Keyphrase generation with correlation constraints[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2018: 4057-4066.

[10] Tu Z, Lu Z, Liu Y, et al. Modeling coverage for neural machine translation[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016: 76-85.

[11] Chen W, Gao Y, Zhang J, et al. Guided encoding for keyphrase generation[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33: 6268-6275.

[12] Chan H P, Chen W, Wang L, et al. Neural keyphrase generation via reinforcement learning with adaptive rewards[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 2163-2174.

[13] Bengio Y, Léonard N, Courville A. Estimating or propagating gradients through stochastic neurons for conditional computation[J]. arXiv preprint arXiv: 1308.3432, 2013.

[14] 周宁,石雯茜,朱昭昭. 基于粗糙数据推理的 TextRank 关键词提取算法[J]. 中文信息学报, 2020, 34(9): 44-52.

[15] 索红光,刘玉树,曹淑英. 一种基于词汇链的关键词抽取方法[J]. 中文信息学报, 2006, 20(6): 27-32.

[16] Wang M, Zhao B, Huang Y. PTR: Phrase-based topical ranking for automatic keyphrase extraction in scientific publications[C]//Proceedings of the International Conference on Neural Information Processing. Springer, Cham, 2016: 120-128.

[17] Hulth A. Improved automatic keyword extraction given more linguistic knowledge[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2003: 216-223.

[18] Nguyen T D, Kan M Y. Keyphrase extraction in scientific publications[C]//Proceedings of the International Conference on Asian Digital Libraries. Springer, Berlin, Heidelberg, 2007: 317-326.

[19] Mihalcea R, Tarau P. Textrank: Bringing order into text[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2004: 404-411.

[20] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]//Proceedings of Advances in Neural Information Processing Systems, 2014: 3104-3112.

[21] Yuan X, Wang T, Meng R, et al. One size does not fit all: Generating and evaluating variable number of keyphrases[J]. arXiv preprint arXiv: 1810.05241, 2018.

[22] Cho K, Merriënboer B, Gulcehre C, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 1724-1734.

[23] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of Advances in Neural Information Processing Systems, 2017: 5998-6008.

[24] Kim Y. Convolutional neural networks for sentence classification[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2014: 1746-1751.

[25] Courbariaux M, Hubara I, Soudry D, et al. Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1[J]. arXiv preprint arXiv: 1602.02830, 2016.

[26] Zhou Z H. A brief introduction to weakly supervised learning[J]. National Science Review, 2018, 5(1): 44-53.

[27] Mikalai Krapivin, Aliaksandr Autayeu, Maurizio Marchese. Large dataset for keyphrases extraction, DISI-09-055[R/OL]. http://eprints.biblio.unitn.it/1671/1/disi09055-krapivin-autayeu-marchese.pdf.2008.

[28] Kim S N, Medelyan O, Kan M Y, et al. Semeval-2010 task 5: Automatic keyphrase extraction from scientific articles[C]//Proceedings of the 5th International Workshop on Semantic Evaluation, 2010: 21-26.

(下转第 81 页)

[19] 王子牛，姜猛，高建瓴，等.基于 BERT 的中文命名实体识别方法[J].计算机科学，2019，46(S2)：138-142.
[20] 尹学振，赵慧，赵俊保，等.多神经网络协作的军事领域命名实体识别[J].清华大学学报(自然科学版)：2020，60(8)：35-42.
[21] 谢云. 面向中文法律文本的命名实体识别研究[D]. 南京：南京师范大学硕士学位论文，2018.
[22] 王礼敏. 面向法律文书的中文命名实体识别方法研究[D].苏州：苏州大学硕士学位论文，2018.
[23] 林义孟. 面向司法领域的命名实体识别研究[D].昆明：云南财经大学硕士学位论文，2019.
[24] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

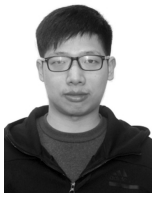
李春楠(1984—)，硕士研究生，主要研究领域为自然语言处理、关系抽取。
E-mail：lichunnan@mail.dlut.edu.cn

孙媛媛(1979—)，通信作者，博士，教授，主要领域为自然语言处理，复杂网络理论。
E-mail：syuan@dlut.edu.cn

王雷(1977—)，博士，三级高级检察官，主要研究领域为刑事司法。
E-mail：18804002266@163.com

(上接第 72 页)

[29] Guo Q, Qiu X, Liu P, et al. Star-Transformer[C]// Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019: 1315-1325.
[30] Yan H, Deng B, Li X, et al. Tener: Adapting transformer encoder for name entity recognition[J]. arXiv preprint arXiv: 1911.04474, 2019.



罗益超(1995—)，硕士研究生，主要研究领域为自然语言处理。
E-mail：ycluo18@fudan.edu.cn

张奇(1981—)，博士，教授，主要研究领域为自然语言处理、信息检索。
E-mail：qz@fudan.edu.cn

李争彦(1997—)，硕士研究生，主要研究领域为自然语言处理。
E-mail：lizy19@fudan.edu.cn