

基于多粒度语义交互的抽取式多文档摘要^①



田媛, 郝文宁, 陈刚, 靳大尉, 邹傲

(陆军工程大学 指挥控制工程学院, 南京 210001)

通信作者: 郝文宁, E-mail: hwnbox@foxmail.com

摘要: 信息爆炸是信息化时代面临的普遍性问题, 为了从海量文本数据中快速提取出有价值的信息, 自动摘要技术成为自然语言处理 (natural language processing, NLP) 领域中的研究重点. 多文档摘要的目的是从一组具有相同主题的文档中精炼出重要内容, 帮助用户快速获取关键信息. 针对目前多文档摘要中存在的信息不全面、冗余度高的问题, 提出一种基于多粒度语义交互的抽取式摘要方法, 将多粒度语义交互网络与最大边界相关法 (maximal marginal relevance, MMR) 相结合, 通过不同粒度的语义交互训练句子的表示, 捕获不同粒度的关键信息, 从而保证摘要信息的全面性; 同时结合改进的 MMR 以保证摘要信息的低冗余度, 通过排序学习为输入的多篇文档中的各个句子打分并完成摘要句的抽取. 在 Multi-News 数据集上的实验结果表明基于多粒度语义交互的抽取式多文档摘要模型优于 LexRank、TextRank 等基准模型.

关键词: 多文档摘要; 抽取式; 多粒度语义交互; MMR; 排序学习

引用格式: 田媛, 郝文宁, 陈刚, 靳大尉, 邹傲. 基于多粒度语义交互的抽取式多文档摘要. 计算机系统应用, 2022, 31(7): 186-193. <http://www.c-s-a.org.cn/1003-3254/8562.html>

Extractive Multi-document Summarization Based on Multi-granularity Semantic Interaction

TIAN Yuan, HAO Wen-Ning, CHEN Gang, JIN Da-Wei, ZOU Ao

(Command & Control Engineering College, Army Engineering University of PLA, Nanjing 210001, China)

Abstract: Information explosion is a common problem in the information age. In order that valuable information can be extracted rapidly from massive text data, automatic summarization technologies have become a research priority in the field of natural language processing (NLP). The purpose of multi-document summarization is to refine important content from a group of documents on the same topic and thereby help users get key information quickly. To address the problems of incomplete information and high redundancy in multi-document summarizations, this study proposes an extractive summarization method based on multi-granularity semantic interaction that combines the multi-granularity semantic interaction network with maximal marginal relevance (MMR). Semantic interaction with different granularities is used to train sentence representation and key information with different granularities is captured to ensure the comprehensiveness of the summarization. In addition, modified MMR is employed to ensure the low redundancy of the summarization. The sentences in the input documents are scored by learning to rank, and summary sentences are then extracted. Experimental results on the Multi-News dataset show that the proposed extractive multi-document summarization model based on multi-granularity semantic interaction outperforms some baseline models such as LexRank and TextRank.

Key words: multi-document summarization; extractive; multi-granularity semantic interaction; maximal marginal relevance (MMR); learning to rank

^① 基金项目: 国家自然科学基金 (61806221)

收稿时间: 2021-10-05; 修改时间: 2021-11-08; 采用时间: 2021-11-12; csa 在线出版时间: 2022-06-02

随着网络信息的爆炸式增长,信息过载成为不可避免的问题,为了帮助用户从海量数据中快速挖掘出有价值的信息,提高获取和利用信息的效率,自动文本摘要技术受到越来越多的关注.自动摘要技术是计算机通过人为制定的算法和输入的文章自动生成摘要的技术^[1],其目的是找到输入文本的概括性关键信息.自动文本摘要可分为几种不同的类型,根据输入文档的数量可将文本摘要分为单文档摘要和多文档摘要;根据摘要的目的可将其分为一般的文本摘要和面向查询的文本摘要,一般的文本摘要要求在摘要中包含输入文档的全部关键信息,而面向查询的文本摘要要在摘要中仅包含输入文档中与特定的用户查询相关的关键信息.

本文研究的对象是一般的多文档摘要,旨在针对单一主题下的多个文本文档生成一个简洁的摘要.多文档摘要的方法通常分为两类:抽取式和生成式.生成式方法需要在理解源文档的基础上生成新的词和句子,Fabrizi等人^[2]将输入的多篇文档拼接成一篇长文档作为模型的输入,然后将多文档摘要转换成一个序列到序列的单文档摘要任务;为了避免过长的输入导致摘要退化的问题,Liu等人^[3]提出了一个层次编码器,使用注意力机制表示跨文档之间的潜在关系,允许文档之间共享信息,而不是简单的将文档拼接.生成式的方法相对复杂,由于自然语言生成技术的限制,其生成的摘要通常存在语法错误、可读性较差等问题.抽取式方法是从源文档中直接抽取具有代表性的句子构成摘要,由于在很大程度上保持了原意,不会出现语法上的错误且相对简单而被广泛使用,常见的有基于质心的方法、基于图的方法以及有监督的方法等.抽取式摘要的关键问题就是要保证抽取句子的主题覆盖度以及多样性,即摘要中包含各个方面的关键信息,且其中重复内容较少.

本文提出一种基于多粒度语义交互的抽取式多文档摘要模型(multi-granularity semantic interaction extractive multi-document summarization model, MGSI),使用多头注意力机制进行词语、句子以及文档3种粒度之间的语义交互,使得学习到的句子表示能包含不同粒度的关键信息,以保证在计算句子重要程度时充分考虑其针对主题内容的全面性;同时结合改进的MMR算法通过排序学习对输入文档中的各个句子打分,该得分同时考虑句子的主题覆盖度以及与其他句子之间的重复度,选取Top-K个句子作为最终的摘要

句并按照在原文中的位置对其进行排列.

1 相关研究

近年来,多文档摘要技术已经成为了NLP中的研究热点,其相关研究能帮助用户快速筛选出关键信息.由于生成式方法需要理解并重新组织输入文档中的信息,相对复杂,当前的主流方法依然是抽取式.Radev等人^[4]将基于质心的方法应用到多文档摘要中,将文档中的重要信息浓缩成几个关键词,根据聚类中心与簇中句子的相似度以及句子的位置信息来识别重要的句子;文献^[5]在此基础上进行改进,提出用句向量表示代替词向量表示,并通过对句子内容相关性、新颖度和位置3个指标的线性结合来改进评分函数.基于图的方法可以利用整个文本的信息来进行排序,TextRank^[6]和LexRank^[7]是两种常见的图排序算法,Alzuhair等人^[8]提出将多种基于图的方法相结合,在计算边的权重时,对4种不同的相似度计算方法进行线性组合,此外,结合两种不同的图排序算法:PageRank^[9]和HITS^[10];张云纯等人^[11]提出了一种聚类和图模型相结合的方法,首先使用基于密度的两阶段聚类方法为全部句子划分主题,然后在各个子主题下建立图模型完成摘要句的抽取.

神经网络随着其不断发展已经被广泛应用于自动文本摘要中,且被证实能有效提高文本摘要的质量,特别地,神经抽取式方法关注学习源文档中句子的向量表示.Cao等人^[12]使用卷积神经网络训练文本分类模型,然后将文档通过分类模型进行分布式表示,利用表示向量来连接文本分类和摘要生成,解决了训练数据不充足的问题;Yasunaga等人^[13]提出使用图卷积网络获取句子嵌入,通过句子关系图来对句子进行重要性评估;Wang等人^[14]构建了一个超图网络进行摘要抽取,在句子级节点的基础上,增加更多的语义单元作为图中额外的节点以丰富句子之间的关系;Cho等人^[15]将行列式点过程(determinantal point process, DPP)应用于抽取式多文档摘要中,并使用胶囊网络^[16]对DPP中句子对之间的相似度计算方法进行改进,以保证摘要中句子的高度多样性;Narayan等人^[17]通过强化学习对ROUGE度量进行全局优化,完成抽取式摘要模型的训练,在训练期间,将最大似然交叉熵损失与强化学习的奖励相结合,直接优化与摘要任务相关的评估指标.上述方法在对句子编码时,

只考虑了句子级的语义信息,缺少对句子中单词、句子所在文档等结构化语义整合的研究.本文通过将不同粒度的语义信息融合,使得获取的句子表示包含更丰富的语义信息,从而保证抽取出的摘要句包含的关键信息更加全面.

2 多粒度语义交互抽取式多文档摘要

在抽取式多文档摘要中,文本之间的交互对于关键信息的抽取有着重要影响,为此本文提出一种基于多粒度语义交互的神经抽取模型,我们的模型由一个多粒度编码器和一个改进的 MMR 模块构成.整体概述如图 1 所示,首先构建单词、句子和文档 3 种粒度的语义交互图,在同一文档的各个句子中使用多头自注意力机制 self-attention 捕获语义关系,句子与句子中的单词之间使用多头交叉注意力机制 cross-attention 捕获语义关系,句子所在文档的语义信息通过 duplicate 进行传递,使用融合机制将多粒度交互信息融合,从而完成句子表示的更新,使得学习到的句向量具有更丰富的特征;然后使用改进的 MMR 算法通过排序学习对输入文档中的全部句子进行排序,完成摘要句的抽取.

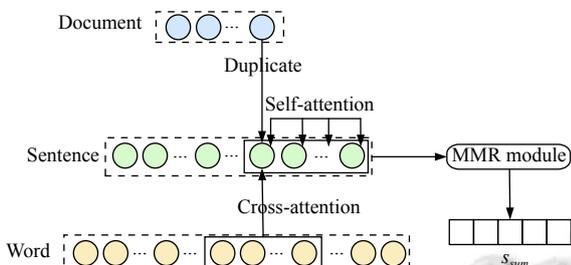


图 1 多粒度语义交互抽取式多文档摘要概述图

2.1 多粒度编码器

使用多粒度编码器获取更新的句子表示.对于输入文档集,首先构建多粒度语义交互图,多粒度编码器的每一层包含两个部分:第 1 部分是一个注意力层,使用多头注意力机制捕获句子与句子、以及句子与单词之间的语义关系,然后使用一个融合门融合不同粒度之间的语义交互信息;第 2 部分是一个全连接的前馈网络,完成多粒度语义信息的进一步转换.

如图 2 所示, $d_i, i = 1, 2, \dots, N$ 表示第 i 个文档, s_{ij} 表示文档 d_i 中的第 j 个句子, w_{ijk} 表示文档 d_i 中第 j 个句子的第 k 个单词, s_{i*} 表示文档 d_i 中的各个句子, w_{ij*} 表示句

s_{ij} 中的各个单词,词 w_{ijk} 的初始向量记为 e_{ijk} ,为了表明输入单词的位置信息,我们使用与 Transformer^[18] 中一致的特殊位置编码,如式 (1) 所示.

$$PE_{(pos,2t)} = \sin(pos/10000^{2t/d})$$

$$PE_{(pos,2t+1)} = \cos(pos/10000^{2t/d}) \quad (1)$$

其中, pos 表示位置索引, t 表示维度索引, d 表示向量的维度.在我们的多粒度编码器中,需要考虑 3 个位置编码:文档位置编码 PE_i 、文档中句子位置编码 PE_j 和句子中单词的位置编码 PE_k ,最终的位置编码和输入层的词向量分别如式 (2)、式 (3) 所示,将初始编码与位置编码相加作为单词级别编码器的输入.

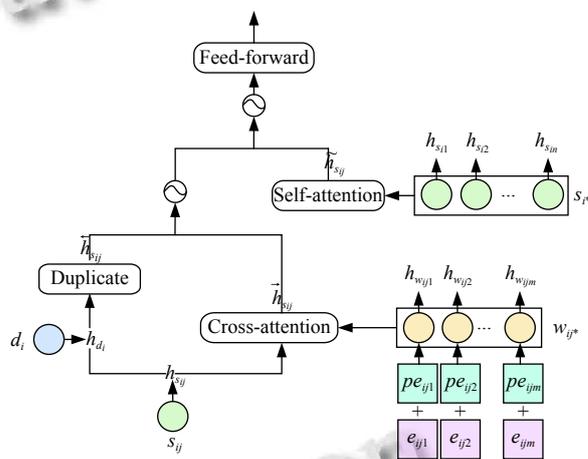


图 2 多粒度编码器

$$pe_{ijk} = [PE_i; PE_j; PE_k] \quad (2)$$

$$h_{w_{ijk}}^0 = e_{ijk} + pe_{ijk} \quad (3)$$

输入层的句子表示 $h_{s_{ij}}^0$ 和文档表示 $h_{d_i}^0$ 均初始化为零,句向量通过对 3 种粒度的语义交互信息融合进行更新:首先是同一文档中句子之间通过多头自注意力机制捕获的上下文表示 $\overleftarrow{h}_{s_{ij}}^l$,如图 2 中所示通过 self-attention 模块获取 d_i 中各个句向量 $h_{s_{i*}}$ 之间的交互信息;然后是句子和句中单词之间通过多头交叉注意力机制 cross-attention 捕获的词粒度的语义信息 $\overrightarrow{h}_{s_{ij}}^l$;最后是句子所在文档传递的文档粒度的语义信息 $\overleftarrow{h}_{s_{ij}}^l$,分别如下所示:

$$\overleftarrow{h}_{s_{ij}}^l = MHA_{tt}(h_{s_{ij}}^{l-1}, h_{s_{i*}}^{l-1}) \quad (4)$$

$$\overrightarrow{h}_{s_{ij}}^l = MHA_{tt}(h_{s_{ij}}^{l-1}, h_{w_{ij*}}^{l-1}) \quad (5)$$

$$\overleftarrow{h}_{s_{ij}}^l = h_{d_i}^{l-1} \quad (6)$$

h^{l-1} , $l=1,2,\dots,L$ 表示编码器第 l 层的输入, $h_{s_{i*}}$ 表示文档 d_i 中各个句子对应的向量, $h_{w_{ij*}}$ 表示句 s_{ij} 中的各个单词对应的向量. *MHAtt*即为 Vaswani 等人^[18]提出的多头注意力机制, 在式(4)中, 文档 d_i 中某一句子的输入向量 $h_{s_{ij}}^{l-1}$ 作为注意力中的*query*, 各个句子的输入向量 $h_{s_{i*}}^{l-1}$ 则作为*keys*和*values*, 在式(5)中, $h_{s_{ij}}^{l-1}$ 同样作为注意力中的*query*, 而此时*keys*和*values*为句中各个单词的输入向量 $h_{w_{ij*}}^{l-1}$.

我们使用两个融合门将多粒度语义信息融合, 从而获取更新的句子表示 $f_{s_{ij}}^l$, 如式(7)所示, 首先将句子与单词之间的交互和句子与文档之间的交互融合, 然后再将其与句子之间的交互进行融合, *Fusion*即表示融合门, 原理如式(8)和式(9)所示, 其中参数 $W \in \mathbb{R}^{2d \times 1}$, $b \in \mathbb{R}^{2d \times 1}$, σ 为 Sigmoid 激活函数.

$$f_{s_{ij}}^l = \text{Fusion}(\text{Fusion}(\overrightarrow{h}_{s_{ij}}^l, \overleftarrow{h}_{s_{ij}}^l), \overleftarrow{h}_{s_{i*}}^l) \quad (7)$$

$$\text{Fusion}(x, y) = zx + (1 - z)y \quad (8)$$

$$z = \sigma(W[x; y] + b) \quad (9)$$

为了进一步转换多粒度语义信息, 将更新后的句子表示 $f_{s_{ij}}^l$ 与该层输入句向量 $h_{s_{ij}}^{l-1}$ 相加, 然后将其传入使用 ReLU 作为激活函数的两层前馈网络中, 归一化后输出, 如式(10)所示, 其中, $W_1 \in \mathbb{R}^{d_{\text{hidden}} \times d}$, $W_2 \in \mathbb{R}^{d \times d_{\text{hidden}}}$, d_{hidden} 为前馈网络隐藏层的大小, 最终获取句向量 h_s^l 用于后续的排序学习.

$$g_{s_{ij}} = W_2 \text{ReLU}(W_1(h_{s_{ij}}^{l-1} + f_{s_{ij}}^l)) \quad (10)$$

$$h_{s_{ij}}^l = \text{LayerNorm}(h_{s_{ij}}^{l-1} + g_{s_{ij}}) \quad (11)$$

2.2 MMR 模块

MMR 算法最初用于文档检索, 通过计算查询与待检索文档之间的相似度以及文档与文档之间的相似度对文档进行打分, 然后对其排序. Carbonell 等人^[19]首次提出将 MMR 用于文本摘要中, 基于与原文的相关度和冗余度为候选句打分, 根据得分进行排序从而完成句子抽取, 目标函数如式(12)所示, 其中, R 表示所有候选句子集, S 表示已经选择了的句子, $R \setminus S$ 表示未被选择的句子, Q 是查询, 在实际摘要中, 通常使用源文档或者输入文档对应的真实摘要作为这里的查询, 每次选取 MMR 得分最高的句子作为摘要句.

$$\text{MMR} = \arg \max_{s_i \in R \setminus S} [\lambda \text{Sim}_1(s_i, Q) - (1 - \lambda) \max_{s_j \in S} \text{Sim}_2(s_i, s_j)] \quad (12)$$

直观上来看, 结合 MMR 算法能帮助选择出与输入文档密切相关且彼此之间重复内容较少的句子, 通过多粒度编码器, 我们已经获取了输入多文档中各个句子的向量表示 h_s^l (L 表示最后一层输出), 由于该向量已经包含了输入的多篇文档中不同粒度的关键信息, 所以对于句子的重要程度, 考虑使用基于句子本身的特征的方法代替计算句子与源文档之间的相似度, 以避免相似度计算过程中丢失句子的相关重要信息, 改进后的 MMR 得分计算公式如式(13)所示, 式中的前半部分通过一个线性转换层的计算来表示句子本身的重要性, 其中, $W_s \in \mathbb{R}^{d \times 1}$, $b_s \in \mathbb{R}^{d \times 1}$, h_s^l 表示输入文档中任意一个句子的向量; 后半部分计算句 s 与源文档中除该句以外的其他所有句子的相似度的最大值, 这里使用余弦相似度函数, 以保证最终抽取的句子包含尽可能少的重复信息, 其中 s' 表示源文档中除 s 以外的其他所有句子.

$$\text{MMR}_s = \lambda(W_s h_s^l + b_s) - (1 - \lambda) \max_{s' \in D \setminus s} \text{Sim}(h_s^l, h_{s'}^l) \quad (13)$$

然后再添加一个 Sigmoid 激活函数对 MMR 得分进行归一化处理, 如式(14)所示, σ 表示 Sigmoid 激活函数.

$$\overline{\text{MMR}}_s = \sigma(\text{MMR}_s) \quad (14)$$

将使用多粒度编码器获取的句子特征向量输入到 MMR 模块中, 通过排序学习为每个句子打分, 得到最终的排序列表, 使用交叉熵作为损失函数, 如式(15)所示, 其中 y_s 为句子的真实得分.

$$L = -\frac{1}{N} \sum_{n=1}^N (y_s^{(n)} \log \overline{\text{MMR}}_s^{(n)} + (1 - y_s^{(n)}) \log(1 - \overline{\text{MMR}}_s^{(n)})) \quad (15)$$

3 实验

实验中分别使用自动评估和人工评估的方法在公开的 Multi-News 数据集上对提出的基于多粒度语义交互的抽取式多文档摘要模型进行评估, 并与一些基准模型进行比较以验证本文的方法的有效性.

3.1 数据集

Multi-News 数据集是用于多文档摘要的第一个大规模数据集, 其中的每个样本由一个人工摘要及其对

应的多个源文档组成,其中,训练集包含 44 972 个样本,验证集和测试集各包含 5 622 个样本.每个摘要平均有 264 个单词,对应的同一主题的源文档平均有 2 103 个单词,摘要对应源文档个数的信息如表 1 所示.数据集中的摘要均为生成式摘要,为了满足本文抽取式模型的训练,我们使用 Jin 等人^[20]通过计算与人工摘要的 Rouge-2 得分构建的标签序列.

表 1 源文档个数分布

# of source	Frequency	# of source	Frequency
2	23 894	7	382
3	12 707	8	209
4	5 022	9	89
5	1 873	10	33
6	763		

3.2 基准模型

实验通过将本文提出的多粒度语义交互抽取式模型与一些经典的基准模型以及近几年中的一些强基准模型进行比较,来验证其对摘要质量改进的有效性,本节对这些基准模型分别作简要的介绍.

LexRank 是一种无监督的基于图的抽取式摘要方法,将文档中的句子作为图中节点,节点之间的连线表示句子间的相似度,通过对句子的相似性进行投票打分以确定句子的重要程度;TextRank 也是一种无监督的基于图的排序方法,句子的重要性得分通过加权图中特征向量的中心性进行计算;MMR 计算句子与原始文档的相关性以及与其他句子之间的相似度,基于相关度和冗余度对候选句子打分,根据得分排名选择句子生成摘要;PGN^[21]是一种基于循环神经网络的生成式摘要模型,该模型使用注意力机制,允许通过指针从源文档中复制单词,也允许根据固定词汇表生成单词,有效缓解了未登录词(out of vocabulary, OOV)的问题;CopyTransformer^[22]对 Transformer 进行扩展,使用一个内容选择器从源文档中筛选出应成为摘要中内容的短语,并将该选择器作为自底向上的注意力机制步骤来对模型进行约束;Hi-MAP 对指针生成网络进行扩展,将其扩展成层次网络,在摘要生成的过程中,结合 MMR 模块对句子打分.

3.3 实验设置

通过初步实验对参数进行设置,将词汇量设为 50 000,词向量维度和隐藏层单元数设为 512,前馈层单元个数设为 2 048,使用 8 头注意力机制,输入时,在不同文档

以及同一文档的不同句子之间分别引入特殊符号,以便于模型对不同粒度进行区分.模型训练时,丢弃率^[23]设为 0.1,Adam 优化器的初始学习 $\alpha = 0.0001$,动量 $\beta_1 = 0.9$, $\beta_2 = 0.999$,权重衰减 $\varepsilon = 10^{-5}$,batch-size 设为 10,超参数 $\lambda = 0.5$,在抽取句子生成摘要时,按照排序抽取 Top-5 个句子作为最终的摘要句.

3.4 实验结果及分析

实验中首先使用 ROUGE 得分^[24]对本文的模型以及基准模型进行自动评估,基准模型中同时包含抽取式模型和生成式模型,通过对比以更好验证本文提出的方法的有效性.ROUGE 基于摘要中 n 元词的共现信息来评价摘要,参考 Lebanoff 等人^[25]的工作,实验中分别使用 ROUGE-1、ROUGE-2 和 ROUGE-SU4 得分作为多文档摘要自动评估的指标,ROUGE-N 主要统计 N-gram 上的召回率,计算预测出的摘要与参考摘要中所共有的 N-gram 个数占参考摘要中总 N-gram 个数的比例;ROUGE-SU4 与 ROUGE-N 不同的是它允许跳词,在对预测出的摘要与参考摘要进行匹配时,不要求 gram 之间必须连续,可以跳过几个单词,考虑了所有按词序排列的词对,能更深入的反映句子级词序.实验结果如表 2 所示,其中 MGSI 表示本文提出的基于多粒度语义交互的抽取式多文档摘要模型.

表 2 Multi-News 数据集测试评估 (%)

模型	ROUGE-1	ROUGE-2	ROUGE-SU4
LexRank	38.27	12.70	13.20
TextRank	38.44	13.10	13.50
MMR	38.77	11.98	12.91
PGN	41.85	12.91	16.46
CopyTransformer	43.57	14.03	17.37
Hi-MAP	43.47	14.89	17.41
MGSI	43.85	15.98	19.62

对于抽取式基准模型,三者 Multi-News 数据集上的表现相差很小,其中 MMR 的 ROUGE-1 得分比 LexRank 和 TextRank 分别高 0.5 个百分点和 0.33 个百分点,而 ROUGE-2 和 ROUGE-SU4 得分则均低于 LexRank 和 TextRank.生成式基准模型普遍比抽取式基准模型表现好,我们认为这可能是因为 Multi-News 数据集中的参考摘要更倾向于使用新的单词和短语来对源文档进行总结.在几个生成式基准模型中,CopyTransformer 比 PGN 在 ROUGE-1、ROUGE-2 和 ROUGE-SU4 三个指标上分别提升了近 4%、9% 和 6%,表明 Transformer 框架在文本摘要任务中优于指

针网络; Hi-MAP 则比 PGN 在 3 个指标上分别提升了近 4%、15% 和 6%, 表明在指针网络的基础上添加 MMR 模块能有效提高文本摘要的质量。

我们的基于多粒度语义交互的抽取式模型在 ROUGE 三个指标上的得分分别是 43.85、15.98 和 19.62, 优于所有的基准模型. 与 MMR 相比, 在 ROUGE-1 上提升了 13.1%, 在 ROUGE-2 上提升了 33.4%, 在 ROUGE-SU4 上提升了 52.0%, 这说明将多粒度语义交互网络与改进的 MMR 相结合抽取的摘要相对于仅用 MMR 模型抽取的摘要有很大的改进, 我们将其归结为多粒度交互网络的有效性, 使用该网络能够捕获到包含不同粒度关键信息的句子表示, 从而提高文本摘要的质量; 从表中实验结果还可以看出, 即使与一些生成式的强基准模型相比, 我们的抽取式多文档摘要模型表现也不差, 比 CopyTransformer 在 3 个指标上分别提升了 0.6%、13.9% 和 13%, 比 Hi-MAP 分别提升了 0.9%、7.3% 和 12.7%, 表明不同粒度之间的语义交互能帮助充分利用全局信息, 例如跨文档、跨句子之间的语义交互, 从而使得更新后的句向量包含不同的关键信息, 即在多文档摘要任务中使用层次编码框架能够有效提升摘要的质量。

为了对摘要的质量做进一步评估, 我们还进行了人工测评. 人工测评要求关注 3 个指标: 相关性、非冗余性和语法性. 其中相关性用来度量摘要是否覆盖源文档中的全部关键信息; 非冗余性用来度量摘要是否包含重复信息; 语法性用来度量摘要的语法是否通顺. 我们从 Multi-News 数据集的测试集中随机选择 20 个样本, 邀请 3 名软件工程专业的研究生对每一个样本对应的摘要依照 Likert scale 就 3 个评估指标分别打分, 使用五级量表, 分值为 1-5, 1 表示最差, 5 表示最好, 每个指标取所有样本得分的平均值作为最终结果. 我们从基准模型中分别选择一个表现较好的抽取式模型和一个生成式模型作为代表, 来与本文提出的 MGSI 模型进行比较。

评估结果如图 3 所示, 本文的 MGSI 模型比其他两种基准模型在 3 种指标上表现都好, 在相关性上, 达到了 3.50 的最高分, 表明多粒度语义交互网络确实能够挖掘句子的深层语义, 从而在计算句子重要性时能考虑到各个方面的关键信息; 在非冗余度方面, MGSI 比 LexRank 和 Hi-MAP 分别高出了 0.91 分和 0.69 分, 表明结合改进的 MMR 算法能够有效减少摘要中的重

复信息, 降低其冗余度; 在语法上, Hi-MAP 模型的得分最低, 这可能是因为生成式的方法需要生成新的词和句子, 往往会造成语法错误, 而抽取式的方法由于直接从原文中抽取句子, 在很大程度上保留了原意, 从评估结果可以看出, MGSI 模型的语法性得分虽然比基准模型略有提高, 但是相对于其他两个指标来说比较低, 这可能是因为我们对抽取的句子直接按照其在原文中出现的位置进行排序, 没有进一步考虑句子之间的逻辑关系, 导致生成的摘要整体上语义连贯性较差, 可读性不高, 这也是后续研究中需要改进的问题。

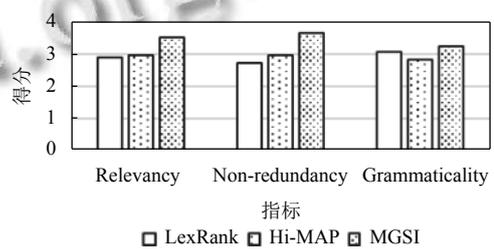


图 3 人工评估

3.5 实例分析

表 3 中展示了本文的 MGSI 模型抽取摘要的一个实例, 从抽取出的摘要本身来看, 句子之间的重复内容较少, 具有较低冗余度, 此外, 与人工摘要进行对比发现, 二者在内容上有较高的重叠, 表中划线部分即为本文的模型抽取的摘要与人工摘要中完全重合的部分. 这进一步说明我们提出的基于多粒度语义交互的抽取式模型能针对多文档生成一个信息较全面且重复内容少的摘要。

表 3 抽取式摘要示例

摘要类型	内容
人工摘要	now the internet is going insane over a hidden panda, buzzfeed reports. illustrator gergely dudás, pen name dudolf, shared an image on facebook last wednesday showing what appears to be a bunch of snowmen, along with the caption: "there's a panda amongst them! can you find it?" though some may spot the panda easily, many apparently have had trouble, with the more than 27, 000 comments including ones like, "there is no panda but it was nice" and "been looking for 15 mins still can't find it lol." the image has been shared nearly 153 000 times so far.
抽取摘要	your reaction? sorry, but you can only react up to 3 times! it looks like you've already used that reaction on this post. <u>there's a panda amongst them! can you find it?</u>

4 结论与展望

本文提出了一种基于多粒度语义交互的抽取式多文档摘要模型,将单词、句子和文档3种粒度的语义关系图与MMR模块结合,以解决多文档摘要中存在的信息主题覆盖度低、冗余度高的问题。通过多粒度编码器获取同一主题下多个文档中全部句子的向量表示,然后使用改进的MMR算法通过排序学习为这些句子打分,从而完成摘要句的抽取。在Multi-News数据集上的实验结果表明,我们的模型优于LexRank、TextRank等基准模型。

在当前的工作中,对于抽取出的句子,我们按照其在文档中出现的先后顺序对其进行排列,但由于抽取的句子来自于不同文档,往往会导致生成摘要的语义连贯性较差,未来我们将考虑通过学习句子的前后逻辑概率对抽取出的句子进行进一步的排序,而不是简单按照其在源文档中出现的顺序进行排列,以保证最终生成的摘要的语义连贯性,增加其可读性。

参考文献

- 张随远,薛源海,俞晓明,等.多文档短摘要生成技术研究.广西师范大学学报(自然科学版),2019,37(2):60-74.
- Fabbri AR, Li I, She TW, *et al.* Multi-News: A large-scale multi-document summarization dataset and abstractive hierarchical model. Proceedings of the 57th Conference of the Association for Computational Linguistics. Long Papers: ACL, 2019. 1074-1084.
- Liu Y, Lapata M. Hierarchical transformers for multi-document summarization. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019. 5070-5081.
- Radev DR, Jing HY, Styś M, *et al.* Centroid-based summarization of multiple documents. Information Processing & Management, 2004, 40(6): 919-938.
- Lamsiyah S, El Mahdaouy A, Espinasse B, *et al.* An unsupervised method for extractive multi-document summarization based on centroid approach and sentence embeddings. Expert Systems with Applications, 2021, 167: 114152. [doi: 10.1016/j.eswa.2020.114152]
- Mihalcea R, Tarau P. TextRank: Bringing order into text. Proceedings of 2004 Conference on Empirical Methods in Natural Language Processing. Barcelona: ACL, 2004. 404-411.
- Erkan G, Radev DR. LexRank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research, 2004, 22: 457-479. [doi: 10.1613/jair.1523]
- Alzuhair A, Al-Dhelaan M. An approach for combining multiple weighting schemes and ranking methods in graph-based multi-document summarization. IEEE Access, 2019, 7: 120375-120386. [doi: 10.1109/ACCESS.2019.2936832]
- Brin S, Page L. The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems, 1998, 30(1-7): 107-117. [doi: 10.1016/S0169-7552(98)00110-X]
- Kleinberg JM. Authoritative sources in a hyperlinked environment. Journal of the ACM, 1999, 46(5): 604-632. [doi: 10.1145/324133.324140]
- 张云纯,张琨,徐济铭,等.基于图模型的多文档摘要生成算法.计算机工程与应用,2020,56(16):124-131. [doi: 10.3778/j.issn.1002-8331.1905-0456]
- Cao ZQ, Li WJ, Li SJ, *et al.* Improving multi-document summarization via text classification. Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco: AAAI, 2017. 3053-3059.
- Yasunaga M, Zhang R, Meelu K, *et al.* Graph-based neural multi-document summarization. Proceedings of the 21st Conference on Computational Natural Language Learning. Vancouver: ACL, 2017. 452-462.
- Wang DQ, Liu PF, Zheng YN, *et al.* Heterogeneous graph neural networks for extractive document summarization. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2020. 6209-6219.
- Cho S, Lebanoff L, Foroosh H, *et al.* Improving the similarity measure of determinantal point processes for extractive multi-document summarization. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019. 1027-1038.
- Hinton GE, Sabour S, Frosst N. Matrix capsules with EM routing. Proceedings of the 6th International Conference on Learning Representations. Vancouver: ICLR, 2018. 1-15.
- Narayan S, Cohen SB, Lapata M. Ranking sentences for extractive summarization with reinforcement learning. Proceedings of 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans: ACL, 2018. 1747-1759.
- Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: ACM,

2017. 6000–6010.
- 19 Carbonell J, Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summaries. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Melbourne: ACM, 1998. 335–336.
- 20 Jin HQ, Wang TM, Wan XJ. Multi-granularity interaction network for extractive and abstractive multi-document summarization. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2020. 6244–6254.
- 21 See A, Liu PJ, Manning CD. Get to the point: Summarization with pointer-generator networks. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver: ACL, 2017. 1073–1083.
- 22 Gehrmann S, Deng YT, Rush AM. Bottom-up abstractive summarization. Proceedings of 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: ACL, 2018. 4098–4109.
- 23 Srivastava N, Hinton G, Krizhevsky A, *et al.* Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 2014, 15(1): 1929–1958.
- 24 Lin CY. ROUGE: A package for automatic evaluation of summaries. Proceedings of the Workshop on Text Summarization Branches Out. Barcelona: ACL, 2004. 74–81.
- 25 Lebanoff L, Song KQ, Liu F. Adapting the neural encoder-decoder framework from single to multi-document summarization. Proceedings of 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: ACL, 2018. 4131–4141.

(校对责编: 孙君艳)