计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(2):242−249 [doi: 10.15888/j.cnki.csa.008913]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于多特征融合的 TextRank 新闻自动摘要模型①

徐　飞[1], 彭佳佳[1], 刘　军[2], 杨　博[1]

[1](西安工业大学 计算机科学与工程学院, 西安 710021)
[2](63768 部队, 西安 710021)
通信作者: 彭佳佳, E-mail: 1448344552@qq.com

摘　要: 随着互联网的发展, 如何快速地从海量新闻中获取核心信息, 减少浏览负担, 是信息部门目前急需解决的问题. 现有的 TextRank 及其改进算法在新闻摘要抽取任务中, 考虑文本特征不全面. 在摘要句选择时, 只考虑到摘要的冗余度, 忽略了摘要的多样性及可读性. 针对上述问题, 本文提出了融合多特征的文本自动摘要方法 MF-TextRank (multi-feature TextRank). 根据新闻的结构、句子和单词总结了更全面的文本特征信息用于改进 TextRank 算法的权重转移矩阵, 使句子权重计算更准确. 采用 MMR 算法更新句子权重, 通过集束搜索得到候选摘要集, 在 MMR 得分的基础上选择内聚性最高的候选摘要集作为最终的摘要输出. 实验结果表明, MF-TextRank 算法在摘要抽取任务中摘要 Rouge 得分优于现有改进的 TexRank 算法, 有效提高了摘要抽取的准确性.

关键词: TextRank; MMR; Word2Vec; 新闻摘要; 多特征融合; 自动摘要

引用格式: 徐飞,彭佳佳,刘军,杨博.基于多特征融合的 TextRank 新闻自动摘要模型.计算机系统应用,2023,32(2):242–249. http://www.c-s-a.org.cn/1003-3254/8913.html

Automatic News Summarization Model Based on Multi-feature TextRank

XU Fei[1], PENG Jia-Jia[1], LIU Jun[2], YANG Bo[1]

[1](School of Computer Science and Engineering, Xi'an Technological University, Xi'an 710021, China)
[2](Unit 63768, Xi'an 710021, China)

Abstract: With the development of the Internet, how to quickly obtain core information from massive news and make browsing easy has become an urgent problem for information departments. The existing TextRank and its improved algorithm fail to consider text features comprehensively in extracting news summaries. In selecting summaries, they only focus on the redundancy and ignore the diversity and readability of the summaries. In order to solve the above problems, this study proposes a multi-feature automatic text summarization method, namely, MF-TextRank. A more comprehensive text feature information is summarized according to the structure, sentences, and words of news, which is used to improve the weight transfer matrix of the TextRank algorithm and make the sentence weight calculation more accurate. Furthermore, an MMR algorithm is used to update sentence weight, and the candidate summary set is obtained by beam search. According to the MMR score, the candidate summary set with the highest cohesion is selected as the final summary for output. The experimental results show that the MF-TextRank algorithm outperforms the existing improved TextRank algorithm in extracting summaries and effectively improves the accuracy in this regard.

Key words: TextRank; MMR algorithm; Word2Vec; news summary; multi-feature fusion; automatic summary

随着互联网及移动设备的普及, 信息呈现爆炸式增长. 新闻网页数量巨大、内容繁杂, 需要大量的时间阅读和整理, 相关部门人员如何高效地从新闻中获取需要的信息, 成为目前急需解决的问题. 自动文本摘要

① 基金项目: 新型网络与检测控制国家地方联合工程实验室基金 (GSYSJ2018006); 陕西省教育厅专项科研计划 (18JK0399)
收稿时间: 2022-06-14; 修改时间: 2022-07-12; 采用时间: 2022-07-20; csa 在线出版时间: 2022-09-14
CNKI 网络首发时间: 2022-11-15

技术通过对原文内容进行理解和深层挖掘, 运用计算机技术自动生成覆盖面广、冗余度低的文本摘要, 从而减少阅读时间. 目前的文本自动摘要技术分为抽取式和生成式. 抽取式摘要选择原文本中信息量最大的句子, 按照一定规则将所选句子在不做任何更改的情况下生成摘要. 生成式摘要则是利用神经网络对文本进行编码, 然后通过解码器对特征进行解码, 生成新的摘要. 在没有较大数据集的情况下, 生成式文本自动摘要模型研究难度较大且研究质量不够. 因此, 在不依赖训练数据的情况下, 完善抽取式算法获得高质量的新闻摘要仍然是一个研究重点. 本文以 TextRank 算法作为研究开展的基础, 旨生成一个最大限度覆盖原文本的内容, 最小化冗余的摘要, 同时保证其可读性与内聚性.

1 相关工作

本节将介绍 TextRank 算法, MMR 算法及其改进算法在抽取式文本摘要领域的一些研究成果.

无监督的 TextRank 算法[1] 实现过程简单, 适用于多文本和单文本. 该算法将文本单元构成图的顶点, 利用句子相似性将顶点进行连接, 通过迭代计算, 选取得分较高的句子组成文摘. 但在计算过程中受词频影响较大, 生成摘要准确性较低, 因此一些学者对其进行改进. 文献 [2] 提出通过大量语料进行关键词扩展, 强化关键词对文本摘要抽取进行指示, 从而提高新闻摘要的质量. 文献 [3] 通过融入标题、特殊段落的句子位置和长度信息对 TextRank 算法进行改进. 文献 [4] 在计算句子权重时加入句子位置信息、线索词, 不仅包含了句子的整体信息, 还包含了句子本身的信息, 因此提高了文摘的准确性. 文献 [5] 提出一种基于主题的情感摘要方法 SE-TextRank, 利用 LDA 算法进行主题抽取, 获得相关主题句子分组, 加入句子位置特征、关键字特征、句子长度特征, 改进句子权重计算公式. 文献 [6] 利用 BM25 计算句子相似度, 选择词频和句子位置作为文本特征进行摘要抽取. 文献 [7] 提出一种新的 DK-TextRank 算法, 采用 K-means 算法进行相似句子聚类, 通过句子的位置、句子与标题的相似性对句子权重进行优化, 挑选每个簇类中权重最高的句子作为摘要. 文献 [8] 利用 TF-IDF 算法计算句子之间的相似性, 通过句子位置、句子与标题相似性、特殊句子计算句子权重, 生成文本摘要候选句群, 利用 MMR 算法对候选句群做冗余处理. 文献 [9] 利用 Word2Vec 算法进行句子表示, MMR 算法对 TextRank 算法生成的候选摘要集去除冗余. 文献 [10] 首先通过 LDA 模型计算新闻主题, BM25 计算句间相似度. 根据句子位置、句子长度、句子与标题相似度 3 部分改进 TextRank 打分函数. 文献 [11] 提出 SW-TextRank 算法, 利用 Word2Vec 算法进行句子表示, 通过句子的位置、句子与标题相似性、关键词覆盖率、关键句子、线索词改进状态转移矩阵. 利用余弦相似度进行冗余处理. 文献 [12] 将 TextRank 算法当作一个过滤器, 基于分层结构提出了一种双层单文档摘要地提取算法. 文献 [13] 一方面结合词频逆句频相似度与词向量余弦相似度共同计算句子得分, 另一方面采用最大边缘相关度算法将抽取到的摘要去除冗余. 文献 [14] 提出了改进的 MMR 的新闻摘要方法. 使用改进的 MMR 模型对支持向量机算法分类结果进行二次选择生成摘要. 该算法平均准确率提高了 0.148, 0.104. 文献 [15] 在 MMR 算法的基础上, 利用 Word2Vec 模型进行句子表示, 并根据关键词与位置信息对句子重要性的影响对句子进行排序, 得到一个高质量的摘要. TextRank 算法及其改进算法在文本摘要生成领域已经有了很多突破性进展, 改进的方向主要是通过加入文本特征信息使句子权重更准确, 减少摘要冗余来更好地解决文本摘要生成任务. 基于上述文章的启发, 本文从两方面对该算法进行改进. 在句子权重计算时, 为了使句子权重更准确, 构建了以新闻结构、句子和单词为文本特征信息的权重转移矩阵, 包括句子段落位置、句子标题相似性、关键句子、句子长度、线索词与转折词、关键词与专有名词. 一个好的摘要应是互相联系的[16], 传统的 MMR 算法只能解决摘要冗余问题, 未考虑摘要的内聚性[17] 及多样性. 因此本文在 MMR 算法的基础上采用集束搜索选择候选摘要集, 输出内聚性最高的摘要集作为最终摘要.

2 文本的网络图构造

文本预处理在解决本文摘要的任务中是必不可少的, 本文将文本表示为图结构, 引入 Word2Vec 进行词向量训练, 使用余弦相似度计算句子之间的相似性.

2.1 文本预处理

根据句子分隔符“, .; ！ ？ ”对文本进行分句处理. 为保证无意义短句对摘要提取不产生影响, 本文删除小于 7 个字的短句. 利用中文开源分词包 Jieba (结巴) 进行分词、去停用词. 为了提高分词精度, 在分词时导入专有名词词表. 最终得到由词项序列构成的句子集合.

2.2 图的构建

TextRank 算法以图模型为基础, 如图 1 所示, 将文本单元构成图的顶点. 其中图的顶点表示原文本中的句子, 两个顶点用边进行连接. 文本可以表示为一个有权图 $G=(V, E)$, V 表示句子顶点的集合, E 表示为边的集合, E 是一个 $V\times V$ 的子集.

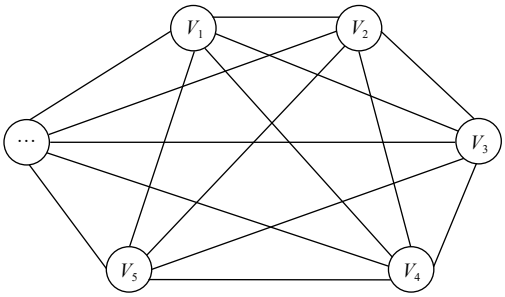

图 1 TextRank 图模型

2.3 权重设置

边的权重能够度量句子之间的语义相似度. 传统的句子相似度计算往往采用 TF-IDF 算法, 未考虑到词语间的关系对句子相似度的影响. 由于同义词在自然语言处理中也是一个非常值得考虑的因素. 句子的相似度建立在词语的相似度之上, 即词语的语义表达, 通过引入词语的相似度必然会提高句子相似度的准确性. 因此, 引入 Word2Vec 词向量模型[18] 进行词向量表示. Word2Vec 的基本思想是通过训练将每个词映射成k维的实数向量后, 通过词之间的距离来判断他们之间的语义相似度. 本文使用 Gensim 自带的 Word2Vec 包进行词向量训练从而将句子间的相似度计算转化为向量运算. $V=\{s_1,s_2,\cdots,s_n\}$表示句子顶点的集合, 集合中的每个句子s_i可以表示为$s_i=\{w_{i1},w_{i2},\cdots,w_{ir}\}, i=1,2,\cdots,n$. r表示句子中包含的词语数量. w_{ij}为k维词向量. 使用向量之间的余弦相似度近似表示句子相似度. 句子s_i, s_j之间的相似度计算公式如下:

$$sim(s_i,s_j)=\frac{\frac{1}{n}\sum_{r=1}^{n}w_{ir}\times\frac{1}{m}\sum_{r=1}^{m}w_{jr}}{\sqrt{\sum_{r=1}^{n}w_{ir}^2}\times\sqrt{\sum_{r=1}^{m}w_{jr}^2}} \tag{1}$$

句子之间的相似度作为图 G 边的权重, 图 G 的邻接矩阵 M 表示如下:

$$M=\begin{bmatrix} 0 & sim(s_1,s_2) & \cdots & sim(s_1,s_n) \\ sim(s_2,s_1) & 0 & \cdots & sim(s_2,s_n) \\ \vdots & \vdots & \ddots & \vdots \\ sim(s_n,s_1) & sim(s_n,s_2) & \cdots & 0 \end{bmatrix} \tag{2}$$

其中,n表示原文本中句子的数量, $sim(s_i,s_j)$表示句子s_i,s_j之间的相似性. 对于无向图而言, 句子之间没有边, 因此邻接矩阵主对角线为 0, 矩阵关于主对角线对称.

3 文本特征计算

文本的特征信息对句子权重的计算有很大的影响. 本节根新据闻的特殊结构从句子和单词层面总结和改进得到了更全面的文本特征信息, 包括句子段落位置、句子标题相似性、关键句子、句子长度、线索词与转折词、关键词与专有名词.

3.1 句子与标题的相似性

从新闻的结构上来说, 标题一般是文章的概括和总结. 句子与标题的相似性越高, 则说明该句子更有可能接近新闻的主题, 那么该句子成为摘要句的可能性就越高. 将新闻标题记为s_0本文采用余弦相似度计算新闻句子与标题的相似度记为$T(s_i)$, 计算公式同式 (1).

3.2 位置信息

根据科学研究成果, 在人工摘要提取中, 选取首段作为摘要的比例为 85%[19]. 新闻结构一般为总分结构, 首段概括性的交代文章的主旨内容. 因此, 本文根据段落和段落中句子的位置对句子进行加权. 首段中越靠前的句子权值越高, 末端中越靠后的句子权值越低. 提出句子位置特征计算公式如下:

$$POS(s_i)=\frac{n-p_i+1}{n}\times\frac{m-p_i+1}{m} \tag{3}$$

其中, $POS(s_i)$表示句子s_i的位置特征权值, n, m分别表示段落数和每段句子数. p_i表示该句子在该段落的位置.

3.3 关键句子

在中文文章中, 如果一个句子自成一段, 那么这个句子一般都是起着承上启下、过渡句或者小标题的作用. 这些句子由于其高度的概括性、精炼性往往是摘要句的首选. 因此需要对此类句子的权重进行提高. 加权规则如下:

$$KS(s_i)=\begin{cases} 1, & s_i\text{是关键句} \\ 0, & s_i\text{不是关键句} \end{cases} \tag{4}$$

3.4 句子长度

在摘要选择中, 句子过长或者过短都不合适. 过长的句子语言描述过于冗余, 过短的句子内容描述不够清晰. 研究发现, 当句子的长度特征大于 0.3 小于 3 时, 符合摘要句条件. 句子的长度特征计算公式如下[10], 其中$L(s_i)$表示句子s_i的长度:

$$l(s_i)=\frac{|L(s_i)-\frac{\sum_{i=1}^{n}L(s_i)}{n}|}{\frac{\sum_{i=1}^{n}L(s_i)}{n}} \tag{5}$$

句子的长度特征值$l(s_i)$越靠近区间 [0.3, 3], 则该句子权重越大, 反之越小, 因此本文定义加权公式如下:

$$L(l(s_i))=\begin{cases} e^{l}-e^{0.3}+1, & 0<l<0.3 \\ 1, & 3\geqslant l\geqslant 0.3 \\ e^{3.3-l}-e^{0.3}+1, & l>3 \end{cases} \tag{6}$$

3.5 线索词与转折词

线索词和转折词通常可以引出具有总结性或者强调性的句子, 见表 1. 在新闻中, 如果一个句子包含线索词或者转折词, 那么该句子更能表达新闻的主要内容. $AC(s_i)$表示句子s_i的线索词和转折词权重, 该权重计算公式如下[15]:

$$AC(s_i)=\begin{cases} 1, & s_i\text{中包含ClueWord中的词语} \\ 0, & s_i\text{中不包含ClueWord中的词语} \end{cases} \tag{7}$$

表 1 线索词转折词表 (部分)

序号	线索词	序号	转折词
1	总之	5	但是
2	常常	6	突然
3	并不是	7	以至于
4	与此同时	8	以上

3.6 关键词和专有名词

关键词和专有名词是一组能够代表文章主要内容的词. 因此, 新闻关键词和专有名词对于提取摘要来说是必不可少的. 一个句子包含的关键词和专有名词越多, 该句子与文章主旨相关程度就越高. 本文采用萨尔顿提出的词频-逆向文件频率 (TF-IDF) 算法[20] 获取新闻的关键词以及关键词的权重. 导入专有名词词典. 基于关键词和专有名词句子权重计算如下:

$$K.PN(s_i)=\frac{\sum_{i=1}^{n}kw\times w}{\sum_{i=1}^{m}KW\times w} \tag{8}$$

$K.PN(s_i)$表示新闻中第i个句子的权重, $kw=ks\cap kos$, $KW=ks\cup kos$, ks表示第k个句子中包含的关键词和专有名词, kos表示除第k个句子外其他句子中包含的关键词和专有名词. w表示词的权重, 本文设置专有名词的权重为 0.5, 关键词权重由 TF-IDF 算法得出.

4 MF-TextRank 算法

传统的 TextRank 算法将句子的相似度矩阵作为权重转移矩阵, 只考虑到句子之间的相似性, 而没有考虑文本自身的特征. 因此本文基于 TextRank 算法提出 MF-TextRank 算法. 在相似度矩阵的基础上融入文本特征来构建一个新的权重转移矩阵, 提高了句子权重计算的准确度. 在进行摘要选择时, 在 MMR 算法的基础上采用集束搜索选择摘要集合, 同时将内聚性最高的候选摘要作为最终摘要.

4.1 构造权重转移矩阵

在 TextRank 算法中, 句子的重要程度是由句子本身所得到的其他句子的“投票”数量和质量决定的. 文本特征信息也会影响句子权重. 如果一个句子的文本特征得分越高, 那么其他句子将会传递更大的权重给它, 也就是说该句子会得到更高的投票, 同时与之关联的句子也会获得更大的权值. 句子最终的计算结果将会更加精准. 基于上节提出文本特征, 定义定句子特征计算公式如下:

$$W_{\mathrm{MF}}(s_i)=\frac{(T+POS+L+AC+K.PN)}{6} \tag{9}$$

以相似度矩阵作为权重转移的基础融入句子的多维度特征, 构建新的权重转移矩阵 W. 等式左边第 1 部分为图 G 的相似度矩阵 M, 第 2 部分为句子特征得分矩阵, 矩阵的第i行表示句子i的特征得分$W_{\mathrm{MF}}(s_i)$:

$$W=M+\begin{bmatrix} W_{\mathrm{MF}}(s_1) & W_{\mathrm{MF}}(s_1) & \cdots & W_{\mathrm{MF}}(s_1) \\ W_{\mathrm{MF}}(s_2) & W_{\mathrm{MF}}(s_2) & \cdots & W_{\mathrm{MF}}(s_2) \\ \vdots & \vdots & \ddots & \vdots \\ W_{\mathrm{MF}}(s_n) & W_{\mathrm{MF}}(s_n) & \cdots & W_{\mathrm{MF}}(s_n) \end{bmatrix} =\begin{bmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{n1} & \cdots & w_{nn} \end{bmatrix} \tag{10}$$

4.2 句子重要性计算

设图 G 中每个句子节点的初始权重$B=[b_1,b_2,\cdots,b_n]$, 其中$b_n=1/n$, 权重转移矩阵为 W. 则句子节点s_i的权重计算公式如下:

$$Score(s_i)=(1-d)+d\sum_{s_j\in In(s_i)}\frac{w_{ij}}{V_k\in Out(s_j)w_{jk}}Score(s_j) \tag{11}$$

其中, $In(s_i)$表示指向节点s_i的句子的集合, $Out(s_i)$表示

节点s_i指向的句子节点的集合, 其中d为阻尼系数, 取值范围为 0–1, 一般取值为 0.85. $Score(s_j)$表示上一步迭代后句子s_j的权重. 经过若干次迭代计算得到$B_i = WB_{i-1}$, 收敛后的B_i包含各个句子节点的权重值.

4.3 摘要句选择

通过第 4.2 节我们得到了每一个句子的重要性权重, 传统方法是对所有句子分数进行降序排序, 选取前 Top-N 个句子作为摘要[13], 该方法忽略了摘要的冗余度. 采用 MMR 算法进行摘要句选择虽然可以对摘要进行冗余处理, 但未考虑到摘要的内聚性以及多样性. 一个好的摘要应该是互相联系的, 低冗余, 可读性高的. 内聚因素 (cohesion factor) 是用来度量摘要中的句子是否在讨论相同的内容[17]. 内聚越高, 摘要可读性就会高. 传统的 MMR 算法通过贪心算法进行摘要选择. 本文采用集束搜索 (beam search) 进行摘要选择从而得到摘要备选集, 最后将摘要备选集内聚性最高的摘要作为最终摘要输出.

MMR 算法由两部分组成, 第 1 部分是第 4.2 节计算得到的句子权重, 用来衡量摘要覆盖原文的程度. 第 2 部分计算所选摘要句之间的相似度, 用来衡量生成摘要的冗余度. 如果当前句子与摘要集句子之间的相似度过大, 那么该句子的 MMR 得分就会越低.

$$\phi_{\text{cov-red}}(s_j) = \sum_{i=1}^{n} \lambda Score(s_i) - (1-\lambda)(sim(s_i, s_j)) \quad (12)$$

一个好的摘要包括紧密耦合的句子. 内聚因素用来衡量摘要句之间的关联程度, 对于备选摘要集 S, CF 值计算如下[21]:

$$CF = \frac{\log(9C_s + 1)}{\log(9M + 1)} \quad (13)$$

$$C_s = \frac{\sum \forall_{s_i, s_j \in summary} sim(s_i, s_j)}{N_s} \quad (14)$$

其中, $N_s = \dfrac{s(s-1)}{2}$, 表示具有s个节点的摘要子图的边数. 采用文献 [17] 公式归一C_s, 得到归一后的 CF 值. 其中M为$\max_{i,j<N} sim(s_i, s_j)$. 因为$C_s \leqslant M$, 所以$CF \leqslant 1$.

在选择摘要句时, 采用集束搜索. 首先对B_i中各个句子节点权重排序, 选择 Top-2 分别放入候选摘要S_A、S_B中, 判断摘要是否已满. 若不满, 则计算剩余句子的 MMR 得分, 选择得分靠前的两个句子放入候选摘要S_A、S_B中生成新的候选摘要, 见图 2. 直至候选摘要已满, 计算 MMR 得分靠前的候选摘要的内聚性, 内聚性最高的摘要按原文顺序输出, 作为最终摘要.

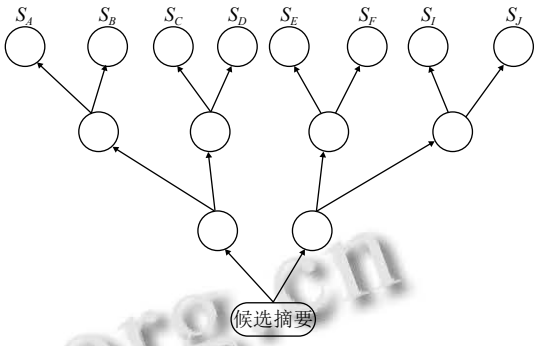

图 2 候选摘要集合

4.4 算法实现

FM-TextRank 算法流程图见图 3.

5 实验与分析

为了验证 FM-TextRank 算法的有效性, 本节设计了 3 组实验分别进行证明. 1) 采用消融分析, 在 TextRank 算法的基础上分别加入不同的特征, 以验证单一特征的有效性. 同时对比本文提出的 MF-TextRank 算法. 2) 将传统的 TextRank 算法及其改进的主流算法与本文提出的 FM-TextRank 算法产生的结果进行对比. 3) 在摘要选择阶段, 使用传统的 MMR 算法以及本文提出的算法进行摘要选择, 同时对比 SW-TextRank 算法. 抽取不同数量的句子作为摘要, 用于研究提取摘要句数量对摘要结果的影响. 采用 Rouge-1、Rouge-2、Rouge-L 三种评价指标对各个算法生成的摘要进行评估.

5.1 数据集与评价标准

本文选取 2017 年 NLPCC 比赛 Task3 提供的 nlpcc2017 摘要数据集. 该数据集包含 50 000 个样本. 其中摘要平均字数为 44, 正文平均字数 990. 随机选取 500 篇作为测试文档.

实验采用 Lin 提出的 ROUGE 评价方法[22]. 其本质思想是将模型产生的系统摘要和参考摘要进行对比, 计算它们之间重叠的基本单元内数目来评价系统摘要的质量. 常用的评价指标为 Rouge-1, Rouge-2, Rouge-L, 其中 1, 2, L 分别表示基于一元词、二元词和最长子字串[23]. 使用工具包 pyrouge 计算 ROUGE 分数. 为了直观地观察实验结果, Rouge 得分取均值.

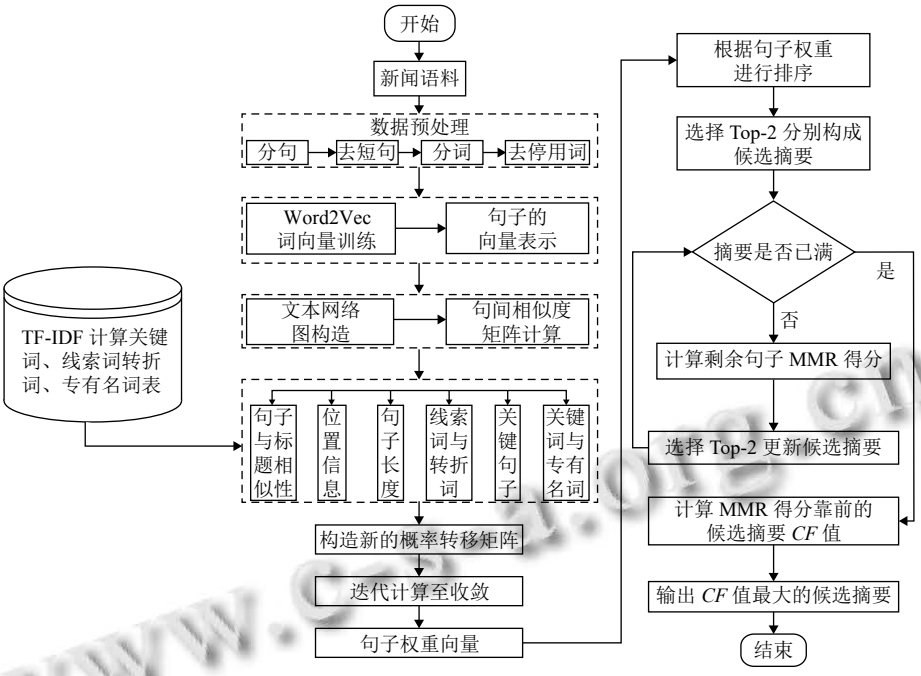

图 3 FM-TextRank 算法流程图

5.2 实验步骤

5.2.1 数据预处理

对正文按照标点符号进行分割, 过滤字数小于 7 的句子. 使用 Jieba 分词包对句子分词、去停用词, 得到由词项构成的句子集合. 该数据集不包含新闻标题, 因此将正文第 1 句话设置为标题句.

5.2.2 句子向量化

输入剩余的 45 000 篇报道, 用于训练 Word2Vec 模型. 选取 DBOW 模型, 实验语言为 Python 3.7, 环境为 Anaconda 数据处理环境. 词向量维数设为 300 维.

5.3 实验与结果分析

5.3.1 实验 1

实验首先验证不同文本特征对摘要抽取的影响, 在经典的 TextRank 算法的基础上分别单独加入句子段落位置、句子标题相似性、关键句子、句子长度、线索词与转折词、关键词与专有名词特征进行摘要抽取, 并将实验结果与本文提出的结合以上 6 种特征的 MF-TextRank 实验结果进行对比. 不同特征值结合的摘要 Rouge 测评如表 2 所示.

实验结果显示, 在只考虑句子之间相似度因素 TextRank 算法的基础上, 单独加入任何一种特征改变权重计算方式的实验结果均有所提升. 单一特征算法效果提升程度有所不同, 融合句子与标题的相似度, 关键句子后 Rouge 的分较高, 标题、关键句子中包含大量新闻信息, 对于摘要抽取有很大的参考意义. 句子位置及句子长度相较标题信息对摘要抽取效果影响较小. 在自然语言处理中语义相较于物理结构更具有参考性. 关键词和专有名词, 线索词和转折词对实验结果影响较小, 词级信息对摘要抽取影响较小. 本文提出的同时结合 6 种特征的 MF-TextRank 方法实验结果最佳, 该方法将 6 种不同的特征组合调整句子权重得到更加精确的句子权重, 从而生成更符合人类阅读的摘要.

表 2 结合不同特征抽取摘要结果 Rouge 值对比

摘要方法	Rouge-1	Rouge-2	Rouge-L
TextRank	0.412	0.196	0.337
+T (标题)	0.44	0.243	0.354
+POS (位置)	0.426	0.231	0.34
+KS (关键句)	0.431	0.237	0.352
+L (句子长度)	0.428	0.214	0.353
+AC (线索词转折词)	0.423	0.220	0.34
+K.PN (关键词和专有名词)	0.421	0.218	0.343
MF-TextRank	0.472	0.280	0.372

5.3.2 实验 2

为了验证 FM-TextRank 算法的有效性, 在相同数据集的基础上, 将本文提出的算法实验结果与的 TextRank 摘要抽取算法以及近些年基于 TextRank 算法提出的一些改进算法 iTextRank、DK-TextRank、联合打分算法的实验结果进行对比, 得到 Rouge-1, Rouge-2、Rouge-L 结果如表 3、图 4.

表 3 不同摘要抽取算法 ROUGE 值对比

算法名称	Rouge-1	Rouge-2	Rouge-L
TextRank	0.412	0.196	0.337
iTextRank	0.434	0.231	0.353
DK-TextRank	0.442	0.258	0.349
联合打分	0.427	0.219	0.352
MF-TextRank	0.472	0.280	0.372

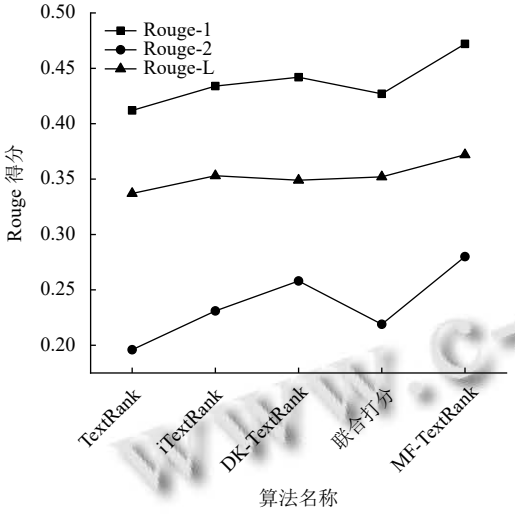

图 4 不同摘要抽取算法 ROUGE 值对比

由表 3 数据可以看出, TextRank 算法在进行摘要抽取任务时表现最差, 它只考虑了句子之间的相似度, 而没有考虑到文本自身的特征. 基于 TextRank 算法的联合打分算法在计算句子相似性时, 不仅使用余弦相似度来衡量句子之间的相似度, 同时采用改进的 TF-IDF 计算句子之间的相似度, 该算法生成的摘要相对 TextRank 算法摘要结果更高, 说明句子之间的相似度会对实验结果产生影响. iTextRank 算法与 DK-TextRank 算法 Rouge 得分非常相近, 这两个算法在构建权重转移矩阵时, 考虑了句子与标题、特殊段落的句子位置, 句子长度信息. DK-TextRank 在考虑文本特征的基础上对相似句子进行聚类, 在不同类别中进行摘要抽取, 减少了摘要的冗余程度. 所以文本特征会对句子的权重计算产生影响, 减少摘要冗余度使得摘要在有限的字数覆盖面更广, 更接近人工摘要. 本文提出的 MF-TextRank 算法生成摘要的 Rouge 得分整体上明显优于上述 4 种算法. 本文提出 6 种特征的结合能更加精确的计算句子权重, 得到较优的摘要.

5.3.3 实验 3

为了进一步验证本文提出的算法的有效性, 以及提取摘要句数目对摘要质量的影响. 在摘要选择时以 MMR 算法作为基准算法、对比本文提出的改进的 MMR 算法, 文献 [11] 提出的 SW-TextRank 算法 . 该算法通过余弦相似度对摘要句群进行冗余处理, 选取适量排序靠前的句子作为摘要. 分别抽取的 4、5、6、8 句作为摘要进行对比, 表 4 给出了对比实验结果.

表 4 不同摘要句算法实验对比

摘要句数目	算法	Rouge-1	Rougr-2	Rouge-L
4	MMR	0.341	0.159	0.271
	SW-TextRank	0.331	0.143	0.258
	MF-TextRank	0.382	0.174	0.293
5	MMR	0.362	0.179	0.290
	SW-TextRank	0.326	0.169	0.284
	MF-TextRank	0.406	0.237	0.315
6	MMR	0.432	0.240	0.357
	SW-TextRank	0.385	0.198	0.318
	MF-TextRank	0.475	0.271	0.369
8	MMR	0.439	0.248	0.352
	SW-TextRank	0.369	0.201	0.320
	MF-TextRank	0.473	0.279	0.375

根据表 4 实验数据可知, SW-TextRank 算法 Rouge 得分最低, 该算法在摘要句选择时仅通过余弦相似度对候选句群进行冗余处理, 对相似度较高的且得分较后的句子进行删除处理. 只考虑到摘要的冗余度, 未考虑摘要覆盖全文的程度. 而 MMR 算法通过引入惩罚因子, 对冗余度较高的句子进行惩罚, 所抽取到的摘要覆盖面广、冗余度小. 本文提出一种改进的 MMR 算法用于摘要句选择, 不仅具有 MMR 算法本身的优点, 同时提高了摘要的灵活性与内聚性, 在相对多样的摘要候选集中选择一个紧密度最高的摘要, 因此在该数据集上有较高的 Rouge 得分.

从图 5 可以清晰地看出不同数量的摘要句提对摘要质量有一定的影响. 随着摘要句的数量的增加, 摘要的 Rouge 得分也在增加. 句子数量在 6–7 句时, Rouge 得分相差不多, 说明该算法抽取摘要句在 6 句左右质量最好. 如果句子较少, 那么覆盖原文内容不全面, 如果句子较多则会产生冗余.

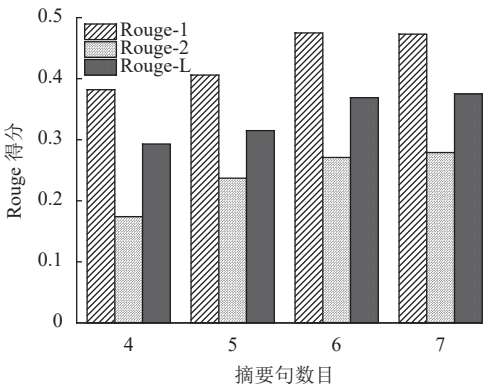

图 5 FM-TextRank 算法不同摘要句数目的 Rouge 值对比

6 结论与展望

自动摘要生成是目前自然语言处理领域的研究重点. TextRank 算法及其改进算法在摘要生成任务上已有一些研究成果, 但仍存在不足. 在句子权重打分时, 考虑的文本特征不够全面, 粒度较粗. 在摘要句选择时, 忽略了自然语言的灵活性以及语言的内聚性. 基于此, 本文提出了一个面向新闻领域的文本自动摘要算法. 根据新闻的结构、句子和单词总结了更全面的文本特征信息用于改进 TextRank 算法的权重转移矩阵, 包括句子段落位置、句子标题相似性、关键句子、句子长度、线索词与转折词、关键词与专有名词. 为了保证摘要的可读性及灵活性, 改进了 MMR 算法的选择方式, 同时考虑了文本的内聚性. 从实验结果来看, 本文提出的算法相较主流的摘要生成模型 Rouge 得分有明显的提高, 模型生成的摘要效果较好. 但抽取式摘要是从原文本中抽取关键的文本单元然后组成摘要, 这种方法生成的摘要阅读起来通常不够顺畅, 因此下一步工作将会考虑将传统的抽取式方法与深度学习融合起来进行摘要模型的探索.

参考文献

1 Mihalcea R, Tarau P. TextRank: Bringing order into text. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. Barcelona: ACL, 2004. 404–411.

2 李峰, 黄金柱, 李舟军, 等. 使用关键词扩展的新闻文本自动摘要方法. 计算机科学与探索, 2016, 10(3): 372–380. [doi: 10.3778/j.issn.1673-9418.1509085]

3 余珊珊, 苏锦钿, 李鹏飞. 基于改进的 TextRank 的自动摘要提取方法. 计算机科学, 2016, 43(6): 240–247. [doi: 10.11896/j.issn.1002-137X.2016.06.048]

4 曹洋. 基于 TextRank 算法的单文档自动文摘研究 [硕士学位论文]. 南京: 南京大学, 2016.

5 刘志明, 于波, 欧阳纯萍, 等. 基于主题的 SE-TextRank 情感摘要方法. 情报工程, 2017, 3(3): 97–104.

6 李楠, 陶宏才. 一种新的融合 BM25 与文本特征的新闻摘要算法. 成都信息工程大学学报, 2018, 33(2): 113–118.

7 徐馨韬, 柴小丽, 谢彬, 等. 基于改进 TextRank 算法的中文文本摘要提取. 计算机工程, 2019, 45(3): 273–277. [doi: 10.19678/j.issn.1000-3428.0051615]

8 李娜娜, 刘培玉, 刘文锋, 等. 基于 TextRank 的自动摘要优化算法. 计算机应用研究, 2019, 36(4): 1045–1050. [doi: 10.19734/j.issn.1001-3695.2017.11.0786]

9 罗飞雄. 基于 TextRank 的自动文摘算法的研究与应用 [硕士学位论文]. 西安: 西安电子科技大学, 2020.

10 罗芳, 汪竞航, 何道森, 等. 融合主题特征的文本自动摘要方法研究. 计算机应用研究, 2021, 38(1): 129–133.

11 汪旭祥, 韩斌, 高瑞, 等. 基于改进 TextRank 的文本摘要自动提取. 计算机应用与软件, 2021, 38(6): 155–160. [doi: 10.3969/j.issn.1000-386x.2021.06.025]

12 何春辉, 李云翔, 王孟然, 等. 改进的 TextRank 双层单文档摘要提取算法. 湖南城市学院学报 (自然科学版), 2017, 26(6): 55–60.

13 朱玉佳, 祝永志, 董兆安. 基于 TextRank 算法的联合打分文本摘要生成. 通信技术, 2021, 54(2): 323–326.

14 程琨, 李传艺, 贾欣欣, 等. 基于改进的 MMR 算法的新闻文本抽取式摘要方法. 应用科学学报, 2021, 39(3): 443–455. [doi: 10.3969/j.issn.0255-8297.2021.03.010]

15 余传明, 郭亚静, 朱星宇, 等. 基于最大边界相关度的抽取式文本摘要模型研究. 情报科学, 2021, 39(2): 34–43.

16 Mitra M, Singhal A, Buckley C. Automatic text summarization by paragraph extraction. Proceedings of the ACL/EACL-97 Workshop on Intelligent Scalable Text Summarization. Madrid: ACL, 1997. 31–36.

17 Qazvinian V, Hassanabadi LS, Halavati R. Summarising text with a genetic algorithm-based sentence extraction. International Journal of Knowledge Management Studies, 2008, 2(4): 426–444. [doi: 10.1504/IJKMS.2008.019750]

18 Mikolov T, Chen K, Corrado G, *et al*. Efficient estimation of word representations in vector space. Proceedings of the 1st International Conference on Learning Representations. Scottsdale: ICLR, 2013. 1–12.

19 Baxendale PB. Machine-made index for technical literature—An experiment. IBM Journal of Research and Development, 1958, 2(4): 354–361. [doi: 10.1147/rd.24.0354]

20 Wu HC, Luk RWP, Wong KF, *et al*. Interpreting TF-IDF term weights as making relevance decisions. ACM Transactions on Information Systems, 2008, 26(3): 13.

21 Chatterjee N, Mittal A, Goyal S. Single document extractive text summarization using genetic algorithms. 2012 3rd International Conference on Emerging Applications of Information Technology. Kolkata: IEEE, 2012. 19–23.

22 Lin CY. ROUGE: A package for automatic evaluation of summaries. Proceedings of the Workshop on Text Summarization Branches Out. Barcelona: Association for Computational Linguistics, 2004. 74–81.

23 李金鹏, 张闯, 陈小军, 等. 自动文本摘要研究综述. 计算机研究与发展, 2021, 58(1): 1–21. [doi: 10.7544/issn1000-1239.2021.20190785]

(校对责编: 牛欣悦)