DOI:10.13718/j.cnki.xsxb.2023.05.011

基于 MMR 和 WordNet 的新闻文本摘要生成研究①

张琪, 范永胜, 金独亮

重庆师范大学 计算机与信息科学学院,重庆 401331

摘要:针对新闻文本摘要提取过程中,传统抽取式算法存在对文本内容概括不全面、摘要内容冗余、关键词提取时未考虑异词同义等问题,提出了一种基于最大边界相关算法(MMR)和词汇语义网(WordNet)的新闻文本摘要生成算法——WMMR. 该算法综合考虑文本相似度、关键词、句子位置信息、线索词等特征对句子权重的影响,从而优化 MMR 算法中的句子得分,并在计算关键词得分时引入 WordNet 合并同义词. 在 NLPCC2017 公开数据集上验证本文算法的有效性,结果表明 WMMR 算法的 *ROUGE* 值相较于 TextRank 算法提升 4 个百分点,相较于 MMR 算法提升 7 个百分点. 在神策杯 2018 与 SogouCS 公开数据集上验证本文算法的普适性,结果表明 WMMR 算法的 *ROUGE* 值相较于传统 TextRank,MMR 等算法均有提升,证明 WMMR 算法有效提升了生成摘要的质量.

关 键 词:新闻文本摘要;抽取式算法;最大边界相关算法;词汇语义网;异词同义

中图分类号:TP391.1 **文献标志码**:A **文章编号**:1000-5471(2023)05-0077-10

Research on News Text Summarizations Generation Based on MMR and WordNet

ZHANG Qi, FAN Yongsheng, JIN Duliang

College of Computer and Information Science, Chongqing Normal University, Chongqing 401331, China

Abstract: In the process of extracting news text summarizations, traditional extraction algorithms have some problems, such as incomplete summarization of text content, redundancy of summary content and synonyms of different words are not considered in keyword extraction. An algorithm WMMR based on Maximal Marginal Relevance (MMR) and WordNet is proposed to generate news text summarizations. In order to optimize the sentence score in MMR algorithm, this algorithm comprehensively considers the influence of text similarity, keywords, sentence position information, clue words and other features on sentence weight. Among them, WordNet is introduced to merge synonyms when calculating the score of keywords. The effectiveness of the proposed algorithm is verified on NLPCC2017 public dataset. The results show that the ROUGE value of WMMR algorithm increases by 4 percentage points compared with TextRank algorithm and 7 percentage points compared with MMR algorithm. The universality of the proposed algorithm is verified on Shence Cup 2018 and SogouCS public datasets. The results show that the ROUGE value of the WMMR algorithm is improved compared with the traditional TextRank and MMR algorithms,

① 收稿日期:2022-06-24

基金项目:重庆师范大学(人才引进/博士启动)基金项目(17XCB008);教育部人文社会科学研究项目(18XJC880002);重庆市教育委员会科技项目(KJQN201800539).

作者简介:张琪,硕士研究生,主要从事自然语言处理研究.

which proves that the WMMR algorithm effectively improves the quality of generated summaries.

Key words: news text summarization; extraction algorithm; maximal marginal relevance algorithm; WordNet; synonyms of different words

近年来随着移动互联网的兴起，各种新闻文章、科学论文等文本数据量爆炸式增长[1]，如何让用户快速、准确地在海量互联网信息中获取具有代表性的内容已经成为一个急需解决的问题. 基于此，各种文本摘要算法应运而生.

文本摘要生成是从原始文本中获得最重要的部分并呈现给用户的过程，其目的是减少文本数量，提取出最相关的信息来简化文本内容，节省用户时间. 从文本摘要的获取方式上来看，可以将其分为抽取式和生成式[2]：前者是对原始文本中的句子进行权重计算并排序，最终选择靠前的适量句子来组成摘要；后者是由模型根据文章大意生成新句子，摘要内容可以包含原始文本中不存在的词语或句子.

文献[3]首次提出“自动摘要”概念，开创了文本摘要的先河. 文献[3]认为文章中的词频和单词在句子中的相对位置是衡量词语是否重要的有效指标，最重要的句子就是包含重要词语的句子，而摘要则是将最重要的句子拼合起来. 目前，国内外学者针对抽取式文本摘要任务做了进一步研究[4-9].

WordNet是一个英语词汇数据库[10]，基于同义和反义来描述词语和概念间的语义关系类型. 文献[11]用GAN生成文本摘要，引入WordNet增强判别器的作用. 文献[12]提出了一种结合统计特征和阿萨姆语WordNet的单文档阿萨姆语文本摘要. 文献[13]利用基于WordNet的Lesk算法分析单词语义，改进句子排序算法，利用Seq2Seq双注意模型进行联合训练.

在传统抽取式算法中，TextRank算法只考虑文本的相似度，忽略文本的语义特征，从而导致摘要内容过度冗余. MMR算法则提出一种惩罚机制来解决冗余问题，但其抽取的文本摘要存在对原文概括能力不足的问题，且并未考虑诸多因素对摘要内容的影响. 本文基于此提出了一种MMR和WordNet的新闻文本摘要生成算法，有效解决了文本内容概括不全面、摘要内容冗余、关键词提取时出现异词同义的问题，该方法提高概括摘要内容能力的同时降低摘要内容的冗余度，提升了生成摘要的质量.

1 算法介绍

1.1 TextRank算法

文献[14]借鉴谷歌的PageRank算法[15]，提出了TextRank算法. 其基本思想是将新闻文本划分为词语或句子来构建图模型，迭代各个节点的权重直至收敛，并通过投票机制对这些词语或句子的重要性进行排序. TextRank算法的公式如(1)式所示：

$$W_s(V_i)=(1-d)+d\times\sum_{V_j\in In(V_i)}\frac{w_{ji}}{V_k\sum_{\in Out(V_j)}w_{jk}}W_s(V_j) \tag{1}$$

其中：$W_S(V_i)$代表节点V_i的权重，W_{ji}代表两个节点V_i和V_j之间的相似程度，$W_S(V_j)$代表上一个节点V_j的权重，$In(V_i)$为指向V_i的节点集合，$Out(V_j)$为V_j指向的节点集合，求和运算代表节点V_i在新闻文本中总的权重[16]. d为阻尼系数，用于做平滑处理，代表某一节点指向其余节点的概率，通常取0.85.

1.2 *TF-IDF*算法

*TF-IDF*算法分为*TF*和*IDF*，其中*TF*表示词频，*IDF*表示逆向文件频率[17]. 该算法反映了词语在文本中的重要性，也反映了词语在数据集中的重要性[18]. *TF-IDF*算法的公式如(2)式所示，具体计算过程如(3)式,(4)式所示：

$$TF\text{-}IDF=TF\times IDF \tag{2}$$

$$TF_{ij}=\frac{n_{i,j}}{\sum_k n_{k,j}} \tag{3}$$

$$IDF_i=\log\frac{|D|}{|\{j:t_i\in d_j\}|+1} \tag{4}$$

其中：$n_{i,j}$表示在文本j中词语i的次数，$\sum_k n_{k,j}$表示文本j中所有词语的总次数，$|D|$表示数据集中的

文本数量，$|\{j:t_i \in d_j\}|$ 表示含有词语 i 的文本数量.

1.3 MMR 算法

最大边界相关法(maximal marginal relevance，MMR)[19] 的基本思想是在保证句子与新闻文本之间相似性的同时，使文本摘要更加全面和多样. MMR 算法公式下

$$V_{\mathrm{MMR}}=\arg\max_{V_i\in D\setminus S}[\lambda\cdot sim_1(V_i,D)-(1-\lambda)\cdot\max_{V_i\in S}sim_2(V_i,V_j)] \tag{5}$$

其中：D 代表整篇新闻文本，S 代表候选摘要句集，V_i 代表当前待抽取的句子，V_j 代表目前已经抽取出的摘要句，λ 是控制 MMR 算法摘要多样性的超参数，第一个相似度 $sim_1(V_i,D)$ 表示句子 V_i 与整篇新闻文本的相似度，第二个相似度 $sim_2(V_i,V_j)$ 表示句子 V_i 和 V_j 之间的相似度.

2 本文算法

本文算法流程图如图 1 所示：

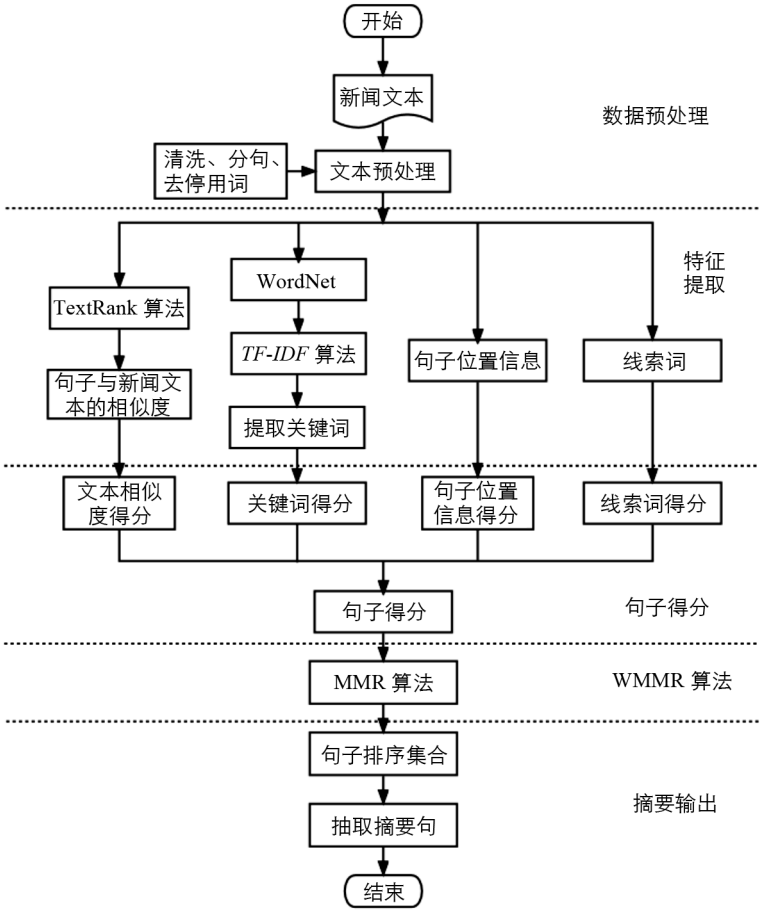

图 1　本文算法的流程图

2.1 数据预处理

数据预处理的主要步骤为：

1) 清洗数据中的多余空字符、网页标记和图片标记，对重复的新闻文本进行去重处理；

2) 将新闻文本和参考摘要分开保存；

3) 首先以句号、感叹号和问号为结束符对新闻文本进行分句，其次采用 jieba 分词的精确模式对句子进行分词，最后进行去停用词操作.

2.2 特征提取

特征提取主要是提取句子与新闻文本的相似度、关键词、句子位置信息和线索词这 4 部分特征.

2.2.1 句子与新闻文本的相似度提取

文本摘要的提取在很大程度上取决于句子与新闻文本中其他句子的相似度. 如果一个句子与其他句子的相似度越高，那么这句话越能概括新闻的大致内容，因此对相似度高的句子赋予更高的权重.

2.2.2 关键词提取

从新闻的关键词中大致可以了解该篇新闻的总体内容，含有关键词的句子相较于新闻中的其他句子更能体现新闻的内容，因此对含有关键词的句子赋予更高的权重.

2.2.3 句子位置信息提取

新闻中句子的重要性通常与句子位置有关，位于新闻首段和尾段的句子一般是对整篇新闻的总结，因此对这些句子赋予更高的权重.

2.2.4 线索词提取

本文将标注新闻来源的词语或具有概括性的词语统称为线索词. 标注新闻来源的词语包括“……网”“……报”“……社”等，具有概括性的词语包括“总之”“综上所述”“总而言之”“归根结底”等. 含有线索词的句子通常具有较强的指向性或总结性，更能概括新闻内容，因此对含有线索词的句子赋予更高的权重.

2.3 特征得分

2.3.1 文本相似度得分

本文通过(1)式所示 TextRank 算法计算每个句子与新闻文本的相似度，最终得出该句的文本相似度得分.

2.3.2 关键词得分

计算关键词得分前先使用 WordNet 合并同义词，输入分词后的中文词语，输出该词语对应的所有英文同义词序列，统计英文单词的词频，将该词频替换(2)式所示 *TF-IDF* 算法中的 *TF* 分数.

使用替换后的 *TF-IDF* 算法公式来提取关键词. 抽取每篇新闻文本中前 20 个关键词，将其作为关键词表，并使用 kw 表示关键词集合. 将单个关键词的权重置为 0.1，如果句子中含有 a 个关键词，关键词权重值 $W_{kw}(V_i)$ 置为 $0.1 \cdot a$；不包含关键词的句子 $W_{kw}(V_i)$ 置为 0. 权重计算如(6) 式所示：

$$W_{kw}(V_i)=\begin{cases}0.1\cdot a, & V_i \in kw,\ 1\leqslant a\leqslant 20\\ 0, & V_i \notin kw\end{cases} \tag{6}$$

2.3.3 句子位置信息得分

由于本文使用的数据集没有进行分段，并且经过测试得出抽取数据集中的尾句效果不佳，因此本文选择抽取新闻文本的前三句，并分别赋予不同的权重. 使用 sp 表示句子位置信息集合. 当句子位置处于第 1，2，3 句时，句子位置信息权重值 $W_{sp}(V_i)$ 分别置为 1，0.79，0.72(由后面 3.3.3 节的实验得出)；其他位置 $W_{sp}(V_i)$ 置为 0. 权重计算如(7) 式所示：

$$W_{sp}(V_i)=\begin{cases}1, & V_i \in sp,\ V_i=1\\ 0.79, & V_i \in sp,\ V_i=2\\ 0.72, & V_i \in sp,\ V_i=3\\ 0, & V_i \notin sp\end{cases} \tag{7}$$

2.3.4 线索词得分

本文从数据集中收集了包含“新华社”“大河报”“人民日报”“央广网”“中国新闻网”在内的 373 个新闻来源词，收集了包含“总之”“综上所述”“总而言之”“归根结底”等在内的 12 个具有概括性的词语. 使用 cw 表示线索词集合. 如果句子中含有线索词，线索词权重值 $W_{cw}(V_i)$ 置为 1；不包含线索词的句子 $W_{cw}(V_i)$ 置为 0. 权重计算如(8) 式所示：

$$W_{cw}(V_i)=\begin{cases}1, & V_i \in cw\\ 0, & V_i \notin cw\end{cases} \tag{8}$$

2.4 句子得分

为了使得算法的摘要与原始的文本内容更加吻合，本文对 2.3 节 4 个主要特征值分别赋予相应的权重，并设计一个加权算法计算最终的句子得分，算法公式如(9) 式所示：

$$W(V_i)=\alpha W_s(V_i)+\beta W_{kw}(V_i)+\gamma W_{sp}(V_i)+\delta W_{cw}(V_i) \tag{9}$$

其中：$\alpha,\beta,\gamma,\delta$ 分别为文本相似度、关键词、句子位置信息、线索词得分的加权系数，取值区间为[0，1]，并且满足 $\alpha+\beta+\gamma+\delta=1$.

2.5　WMMR 算法公式

WMMR算法在MMR算法基础上进行改进，用(9)式中计算得到的最终句子得分 $W(V_i)$ 替换公式(5)式中的句子与新闻文本的相似度 $sim_1(V_i, D)$. 替换后的公式如(10)式所示：

$$V_{\text{WMMR}}=\arg\max_{V_i\in D\setminus S}[\lambda\cdot W(V_i)-(1-\lambda)\cdot\max_{V_i\in S}sim_2(V_i, V_j)] \tag{10}$$

2.6　输出摘要

在计算句子得分后排序，输出排名前三的句子作为该新闻文本的摘要，同时为了确保最终摘要的可读性与连贯性，在输出摘要句时将其按照新闻文本的原始顺序输出.

3　实验验证

3.1　实验数据

本文采用 NLPCC2017 Shared Task3 评测任务的数据集(简称为 NLPCC2017)进行实验，共有 50000 篇带有参考摘要的新闻文本. 该文本特点是不分领域和类型，且篇幅较长[20].

3.2　评价标准

本文使用 *ROUGE*(Recall-Oriented Understudy for Gisting Evaluation)[21]指标对算法得到的自动摘要进行评估. 该评测方法的内部原理是将专家撰写的摘要作为参考摘要，通过统计参考摘要和自动摘要之间重叠的基本单元数，来衡量参考摘要和自动摘要之间的相似程度，从而得到生成摘要的分数[22]. 它是一种对 n 元词召回率进行抽象评价的方法. ROUGE 的评测方法中最常用的是 *ROUGE-N*(基于 N 元词). *ROUGE-N* 的公式如(11)式所示：

$$ROUGE-N=\frac{\sum_{S\in\{ReferenceSummaries\}}\sum_{gramn\in S}Count_{match}(gram_n)}{\sum_{S\in\{ReferenceSummaries\}}\sum_{gramn\in S}Count(gram_n)} \tag{11}$$

其中：$gram_n$ 代表 *n-gram* 的长度，$\{ReferenceSummaries\}$ 代表参考摘要，$Count_{match}(gram_n)$ 代表参考摘要和自动摘要之间重叠的基本单元数，$Count(gram_n)$ 代表参考摘要中基本单元数.

本文采用 *ROUGE*-1(基于 1 元词)、*ROUGE*-2(基于 2 元词)和 *ROUGE-L*(基于最长子串)的结果来评测实验效果.

3.3　实验过程

3.3.1　λ 因子选择

λ 因子是 MMR 算法中的一个重要参数，其取值区间为[0，1]，本文为了选择最佳的 λ 因子进行了多次实验. 由于 λ=0 时，忽略文本相似度对算法的影响，不适用于本文算法，因此不考虑 λ=0 的情况. 文献[23]研究发现当 λ 因子以小于 0.1 的级数增加时，*ROUGE* 值变化不明显，因此以 0.1 的间距选择 10 个点绘制折线图(图 2).

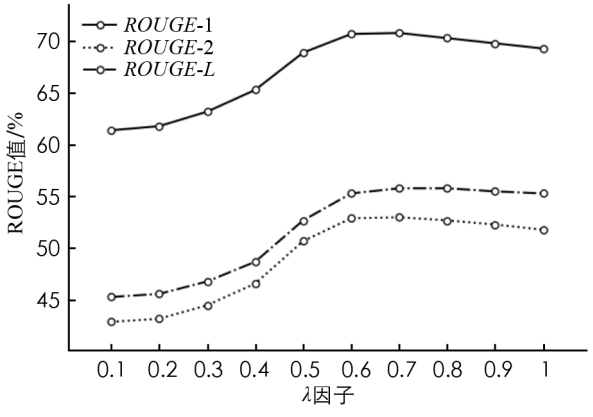

图 2　λ 因子对 ROUGE 值的影响

由图2可以看出，随着λ因子的增大，*ROUGE* 值呈先上升后下降的趋势. 当λ=0.6时，*ROUGE*-1值达到最佳结果；当λ=0.7时，*ROUGE* 值均达到最佳结果. 因此本文将WMMR算法中的λ因子设置为0.7.

3.3.2 关键词提取算法选择

本文在提取关键词时采用 *TF-IDF* 和 TextRank 两种算法，在不考虑 WordNet 对关键词提取算法的影响且文本相似度、关键词、句子位置信息、线索词的加权系数设置为0，1，0，0的前提下，将关键词数设置为20、单个关键词权重设置为0.1，测试两种算法的性能，测试结果如表1所示.

表1 关键词提取算法性能测试

算法	*ROUGE*-1	*ROUGE*-2	*ROUGE-L*
TF-IDF	75.2%	57.4%	60.1%
TextRank	74.7%	56.6%	59.1%

由表1对比得出，在提取关键词时，*TF-IDF* 算法的*ROUGE* 值均高于TextRank算法，因此本文采用 *TF-IDF* 算法提取关键词.

3.3.3 句子位置信息权重选择

本文分别抽取新闻文本的第一、二、三、四句以及尾句，测试句子位置信息对 *ROUGE* 值的影响，测试结果如表2所示.

表2 句子位置信息对 *ROUGE* 值的影响

句子位置信息	*ROUGE*-1	*ROUGE*-2	*ROUGE-L*
第一句	54.8%	38.4%	40.2%
第二句	43.3%	27.7%	31.9%
第三句	39.6%	23.8%	28.7%
第四句	37.3%	21.5%	26.9%
尾句	28.7%	15.4%	23.2%

由表2可以看出，抽取第一、二、三、四句以及尾句的 *ROUGE* 值呈递减趋势，证明在该数据集中抽取尾句的效果并不理想，因此本文选择抽取新闻文本的前三句.

为了得到最优的句子位置信息权重，在文本相似度、关键词、句子位置信息、线索词的加权系数设置为0.9,0,0.1,0的前提下，进行了5次实验：①将前三句的权重均置为1；②将第一、二、三句的权重分别置为1,0.5,0.5；③将第一句权重值置为1，第二句的权重由第二句 *ROUGE*-1 值和第一句 *ROUGE*-1 值的比值得出，即$\frac{0.433}{0.548}\approx 0.79$，同理第三句的权重值为$\frac{0.396}{0.548}\approx 0.72$；④将第一句权重值置为1，第二句的权重值为0.72(由第二句 *ROUGE*-2 值和第一句 *ROUGE*-2 值的比值得出)，同理第三句的权重值为0.62；⑤将第一句权重值置为1，第二句的权重值为0.79(由第二句 *ROUGE-L* 值和第一句 *ROUGE-L* 值的比值得出)，同理第三句的权重值为0.71. 实验结果如表3所示.

表3 句子位置信息权重对 *ROUGE* 值的影响

实验	句子位置信息权重			*ROUGE*-1	*ROUGE*-2	*ROUGE-L*
	首句	第二句	第三句			
1	1	1	1	76.8%	59.1%	62.5%
2	1	0.5	0.5	76.8%	59%	62.6%
3	1	0.79	0.72	76.8%	59.1%	62.6%
4	1	0.72	0.62	76.8%	59%	62.5%
5	1	0.79	0.71	76.8%	59.1%	62.5%

由表3对比得出，第三次实验的 *ROUGE* 值综合表现最佳，因此本文将新闻文本前三句的权重分别置为1,0.79,0.72.

3.3.4　加权系数选择

为了综合考虑文本相似度、关键词、句子位置信息、线索词对句子权重的影响，本文根据以上 4 种影响因子的加权系数 $\alpha,\beta,\gamma,\delta$(取值区间为[0，1]，并且满足 $\alpha+\beta+\gamma+\delta=1$)设置不同的加权系数组合. 并通过大量实验来选取 *ROUGE* 值最佳的组合，其中具有代表性的 11 组参数组合结果如表 4 所示.

表 4　加权系数组合结果

组别	α	β	γ	δ	*ROUGE*-1/%	*ROUGE*-2/%	*ROUGE*-*L*/%
1	1	0	0	0	76.1	58.1	61.5
2	0.9	0.1	0	0	76.4	58.4	61.9
3	0.9	0	0	0.1	76.4	58.5	62.0
4	0.9	0	0.1	0	76.8	59.1	62.6
5	0.8	0.05	0.1	0.05	77.1	59.4	62.9
6	0.7	0.1	0.1	0.1	77.4	59.7	63.2
7	0.6	0.1	0.15	0.15	77.6	60.0	63.6
8	0.5	0.15	0.2	0.15	77.8	60.3	63.9
9	0.4	0.15	0.3	0.15	77.4	60.0	63.6
10	0.3	0.2	0.3	0.2	77.1	59.8	63.4
11	0.2	0.25	0.3	0.25	76.8	59.4	62.9

由表 4 可以看出，当组别为 8，即 $\alpha=0.5$，$\beta=0.15$，$\gamma=0.2$，$\delta=0.15$ 时，*ROUGE* 值最高. 证明文本相似度对于摘要生成的效果影响最大，关键词、句子位置信息、线索词对于摘要生成的效果影响相对较小.

3.4　摘要结果分析

通过一个实际例子[24]来对比 WMMR 算法和基线算法的摘要结果. 以句号为结束符对文本进行分句，如表 5 所示. 其中，Lead3 和 Last3 算法适用性不强，不予考虑. TextRank，MMR，WMMR 3 种算法共同选取了新闻的第 1 句，故不考虑第 1 句对算法效果的影响. TextRank 算法选取第 6，8 句，关于车辆行驶方向等描述存在重复，可以看出 TextRank 算法只考虑文本的相似度，忽略文本的多样性，从而导致摘要内容存在冗余问题. MMR 算法选取第 8，10 句，其中第 10 句无实质性价值，因此 MMR 算法追求文本的多样性，可以解决冗余问题，但其抽取的文本摘要存在对原文概括能力不足、描述不准确等问题. 本文提出的 WMMR 算法选取第 2，5 句，不存在表述重复问题，并且描述出了标题内容中的“门卫大爷吓一跳”，综合考虑了诸多因素对摘要内容的影响，提高了对新闻文本摘要内容的概括能力，同时降低了摘要内容的冗余度，提升了生成摘要的质量.

表 5　文献[24]内容

项目	内容
分句结果	1、7 月 21 日凌晨 6 时左右，在泸州市区白招牌步行街，一辆奔驰越野车，或因刹车失灵，从人防办公楼侧面的门卫室对穿而过，所幸无人员受伤. 2、7 月 21 日 9 时，四川新闻网记者来到事故现场看到，肇事车已经被拉走. 3、……(省略) 4、…… 5、人防办公室保安刘师傅介绍，当时自己在办公大楼的寝室值班，突然听到“轰隆”一声巨响，就看见门卫室里面冲出一辆黑色奔驰越野车，把刘师傅着实吓了一跳. 6、刘师傅说，“那个车是从前门卷帘门冲进来的，直到把后墙撞倒了才停住. 7、…… 8、在现场记者看到，门卫室前面是一条不足 200 米的下坡，汽车从对面行驶下来，本该在前面道路转弯，可不知什么原因，该车就对直冲了下来. 9、…… 10、刘师傅说. 11、……
标题	四川一失控奔驰半夜撞穿门卫室 门卫大爷吓一跳
参考摘要	泸州一奔驰越野车凌晨将人防办公楼门卫室撞穿，致房间完全毁坏、后墙倒塌；疑因刹车失灵，司机无大碍.

3.5 实验结果分析

通过实验得出最佳的加权系数组合后，为了验证算法的有效性，将 Lead3 算法、Last3 算法、TextRank 算法、TextRank + Word2Vec 算法[25]、MMR 算法、MMR + Word2Vec 算法[26]、WMMR + Word2Vec 算法、WMMR 算法进行对比，对比结果如图 3 所示.

其中，TextRank + Word2Vec 算法、MMR + Word2Vec 算法和 WMMR + Word2Vec 算法中的 Word2Vec 模型采用 CBOW 训练方式，使用中文维基百科语料库训练，窗口大小设置为 10，词向量维度设置为 200.

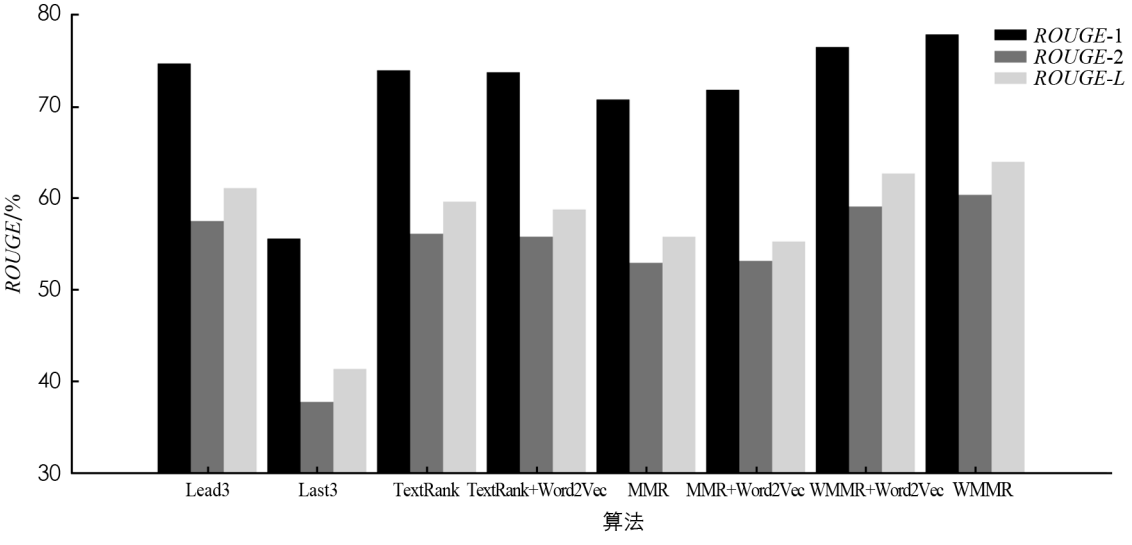

图 3 算法效果对比图

由图 3 可以看出，在面对同一数据集时，Last3 算法效果最差，*ROUGE* 值最低，证明该新闻文本的结尾部分对整篇新闻的总结性较弱；MMR 算法、MMR + Word2Vec 算法、TextRank + Word2Vec 算法、TextRank 算法的 *ROUGE* 值较 Last3 算法有小幅度提升，但稍逊于 Lead3 算法，证明 MMR 算法、TextRank 算法在抽取摘要时存在不足；Lead3 算法在传统算法中效果最好，证明该新闻文本的开头比结尾部分更能代表整篇新闻的内容，但这种仅抽取前三句的方法并未考虑诸多因素对摘要内容的影响. 将本文提出的 WMMR 算法与 Word2Vec 模型相结合，效果略逊于 WMMR 算法，证明 Word2Vec 模型不适用于本文算法.

本文提出的 WMMR 算法效果优于上述传统算法，*ROUGE* 值均达到最高，相较于 TextRank 算法提升 4 个百分点，相较于 MMR 算法提升 7 个百分点，相较于表现较好的 Lead3 算法也有 3 个百分点的提升. 这说明该算法既解决了 MMR 算法、TextRank 算法等传统算法在抽取摘要时的不足，又综合考虑了诸多因素对句子权重的影响，进一步证明本文算法的有效性.

3.6 普适性验证

在“神策杯”2018 高校算法大师赛的比赛数据(简称为神策杯 2018)、搜狗实验室整理的搜狐新闻 18 个频道 2012 年 6 月与 7 月的新闻数据(简称为 SogouCS)上验证 WMMR 算法的普适性. 为了保证实验条数的一致性，选取神策杯 2018、SogouCS 数据集的前 50000 条数据分别进行测试. 结果如图 4－5 所示.

由图 4,5 可以看出，在神策杯 2018、SogouCS 两个不同的公开数据集上，本文提出的 WMMR 算法效果均优于 MMR 与 TextRank 传统算法，证明本文算法具有普适性.

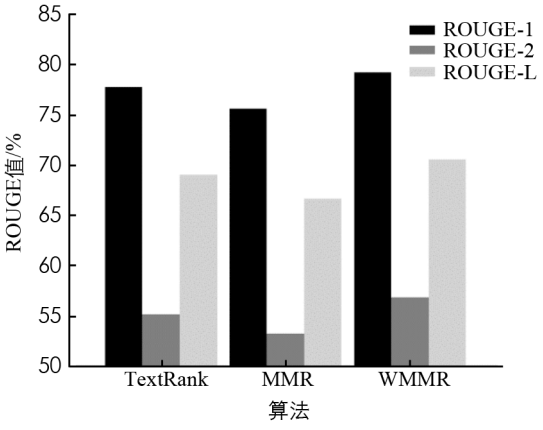

图 4　神策杯 2018 数据集实验结果

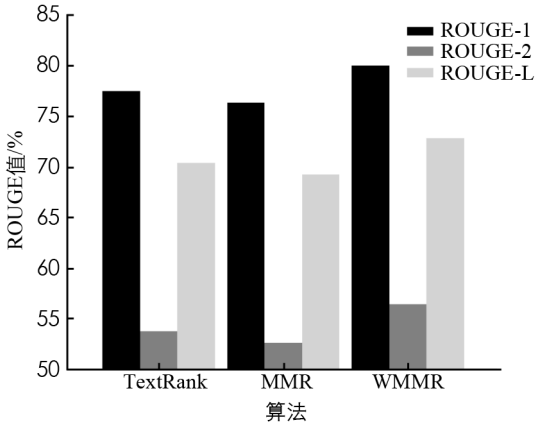

图 5　SogouCS 数据集实验结果

4　结论

本文提出了一种基于 MMR 和 WordNet 的新闻文本摘要生成算法——WMMR. 该算法综合考虑文本相似度、关键词、句子位置信息、线索词等特征对句子权重的影响，从而优化 MMR 算法中的句子得分，并在计算关键词得分时引入 WordNet 合并同义词.

在三个公开数据集上验证本文算法的有效性. 实验结果表明，本文提出的 WMMR 算法 ROUGE 值均最高，整体上明显优于其它传统算法，有效地提升了生成摘要的质量. 但本文只进行了抽取式文本摘要的方法优化，后续将尝试进行生成式文本摘要的方法优化.

参考文献：

[1] HIMA BINDU SRI S, DUTTA S R. A Survey on Automatic Text Summarization Techniques [J]. Journal of Physics: Conference Series, 2021, 2040(1): 012044.

[2] 李金鹏，张闯，陈小军，等. 自动文本摘要研究综述 [J]. 计算机研究与发展，2021，58(1)：1-21.

[3] LUHN H P. The Automatic Creation of Literature Abstracts [J]. IBM Journal of Research and Development, 1958, 2(2): 159-165.

[4] 汪旭祥，韩斌，高瑞，等. 基于改进 TextRank 的文本摘要自动提取 [J]. 计算机应用与软件，2021，38(6)：155-160.

[5] 祝超群. 基于改进 TextRank 的中文文本摘要方法研究 [D]. 武汉：武汉邮电科学研究院，2021.

[6] 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 [J]. 应用科学学报，2021，39(3)：443-455.

[7] 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 [J]. 情报科学，2021，39(2)：34-43.

[8] ELBAROUGY R, BEHERY G, EL KHATIB A. Extractive Arabic Text Summarization Using Modified PageRank Algorithm [J]. Egyptian Informatics Journal, 2020, 21(2): 73-81.

[9] ABDULATEEF S, KHAN N A, CHEN B L, et al. Multidocument Arabic Text Summarization Based on Clustering and Word2Vec to Reduce Redundancy [J]. Information, 2020, 11(2): 59.

[10] MILLER G A, BECKWITH R, FELLBAUM C, et al. Introduction to WordNet: an On-Line Lexical Database [J]. International Journal of Lexicography, 1990, 3(4): 235-244.

[11] 刘晓影，王淮，乌吉斯古愣. 基于 GAN 和中文词汇网的文本摘要技术 [J]. 计算机科学，2022：49(12)：301-304.

[12] BARUAH N, SARMA S K, BORKOTOKEY S. A Single Document Assamese Text Summarization Using a Combination of Statistical Features and Assamese WordNet [C]//Progress in Advanced Computing and Intelligent Engineering. Singapore: Springer, 2021: 125-136.

[13] XIE N T, LI S J, REN H L, et al. Abstractive Summarization Improved by WordNet-Based Extractive Sentences [C]//

CCF International Conference on Natural Language Processing and Chinese Computing. Berlin: Springer, 2018: 404-415.

[14] MIHALCEA R, TARAU P. Textrank: Bringing order into texts[C]// Proceedings of the 2004 conference on empirical methods in natural language processing. Pennsylvania: Association for Computational Linguistics, 2004: 404-411.

[15] DOM B, EIRON I, COZZI A, et al. Graph-Based Ranking Algorithms for E-Mail Expertise Analysis [C]//Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. New York: ACM, 2003: 42-48.

[16] BRIN S, PAGE L. The Anatomy of a Large-Scale Hypertextual Web Search Engine [J]. Computer Networks and ISDN Systems, 1998, 30(1-7): 107-117.

[17] SALTON G, BUCKLEY C. Term-Weighting Approaches in Automatic Text Retrieval [J]. Information Processing & Management, 1988, 24(5): 513-523.

[18] HERNÁNDEZ-CASTAÑEDA Á, GARCÍA-HERNÁNDEZ R A, LEDENEVA Y, et al. Extractive Automatic Text Summarization Based on Lexical-Semantic Keywords [J]. IEEE Access, 2020, 8: 49896-49907.

[19] CARBONELL J, GOLDSTEIN J. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries [C]//Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. New York: ACM, 1998: 335-336.

[20] 侯圣峦, 张书涵, 费超群. 文本摘要常用数据集和方法研究综述 [J]. 中文信息学报, 2019, 33(5): 1-16.

[21] LIN C Y. Rouge: A Package for Automatic Evaluation of Summaries [C]// Proceedings of the Workshop on Text Summarization Branches Out. Pennsylvania: Association for Computational Linguistics, 2004: 74-81.

[22] 张琪, 范永胜. 基于改进 T5 PEGASUS 模型的新闻文本摘要生成研究 [J/OL]. 电子科技: 1-7[2022-05-01]. DOI: 10.16180/ j. cnki. issn1007-7820. 2023. 12. 010.

[23] 曾昭霖, 严馨, 余兵兵, 等. 基于分层最大边缘相关的柬语多文档抽取式摘要方法 [J]. 河北科技大学学报, 2020, 41(6): 508-517.

[24] 杭州网. 四川一失控奔驰半夜撞穿门卫室 门卫大爷吓一跳 [EB/OL]. (2015-07-21)[2022-02-01]. https: //news. hangzhou. com. cn/shxw/content/2015-07/21/content_5854531. htm.

[25] 宁建飞, 刘降珍. 融合 Word2Vec 与 TextRank 的关键词抽取研究 [J]. 现代图书情报技术, 2016(6): 20-27.

[26] 陶兴, 张向先, 郭顺利, 等. 学术问答社区用户生成内容的 W_2V-MMR 自动摘要方法研究 [J]. 数据分析与知识发现, 2020, 4(4): 109-118.

责任编辑 张枸