融合句义特征的多文档自动摘要算法研究

罗森林，白建敏，韩磊，潘丽敏，孟强

（北京理工大学 信息安全与对抗技术实验室，北京 100081）

摘要：多文档自动摘要研究是自然语言处理领域的关键问题之一，为使抽取的摘要能体现多文档主题，在子主题划分的基础上，提出了一种融合句义特征的句子优化选择方法。该方法基于句义结构模型，提取句义结构中的话题、谓词等特征，并融合统计特征构造特征向量计算句子权重，最后采用综合加权选取法和最大边缘相关相结合的方法抽取摘要。选取不同主题的文本集进行实验和评价，在摘要压缩比为 15%情况下，系统摘要平均准确率达到 66.7%，平均召回率达到 65.5%。实验结果表明，句义特征的引入可以有效提升多文档摘要效果，为自动摘要技术的发展提供了一种新思路。

关键词：多文档自动摘要；句义结构模型；句义特征；自然语言处理

中图分类号：TP391；TP18

Research on Multi-document Summarization Merging the Sentential Semantic Features

LUO Sen-lin, BAI Jian-min, HAN Lei, PAN Li-min, MENG Qiang

Abstract: Multi-document summarization is one of the key issues in the field of natural language processing. In order to extract compendious sentences to reflect more accurate theme of the multi-document a new method is proposed to retrieve terse sentences. At first, we extract sentential semantic features, for example, topic and predicate, on the basis of sentential semantic model. Then we calculate sentence weight by building feature vector merging statistical features and sentential semantic features. Finally, we extract sentences by the method of Feature Weighting and Maximal Marginal Relevance. Over one set of texts are selected for experimentation and evaluation, and a good result is achieved eventually. The average precision rate of summary generated by our system reaches 66.7% and the average recall rate reaches 65.5% on condition that the compression ratio of summary is 15%. The results of experiments show that the sentential semantic features are effective on upgrading the affection of Multi-document summarization. The novel method provides a new idea in the development of automatic summarization technology.

Key words: multi-document summarization; sentential semantic model; sentential semantic feature; natural language processing

多文档自动摘要是自然语言处理领域的一个重要课题。它的目的是从主题相同或相似的文档集合中抽取重要信息生成信息丰富、语言简洁并且符合压缩比要求的摘要，从而提供一种快速浏览和获取信息的手段[1]。多文档自动摘要技术经过多年的发展出现了很多方法和技术，比较有代表性的有：美国密歇根大学的 Radev 等人[2]提出的 MEAD 多文档自动摘要系统以同一主题多数文本涉及的词作为质心，按照句子与这个质心相关性来对句子进行排序，抽取句子生成文摘；Erkan 和 Radev[3]提出了一种 LexPageRank 算法，通过特征向量中心的概念来计算句子的重要性，这种方法成功应用到了 Google PageRank 中；美国加州大学伯克利分校的 Chris Ding 等人[4]提出用非负矩阵分解（Non-negative matrix factorization，NMF）方法进行自动摘要，这种方法首先构造句子-词语矩阵，然后以最大概率抽取每个子主题中最具代表性的句子组成摘要。近年来，有些学者通过概率浅层语义分析（Probabilistic Latent Semantic Analysis，PLSA）和浅层狄利克雷分布（Latent Dirichlet Allocation，LDA）[5]来生成多文档摘要，都取得了较好的效果。

中文多文档自动摘要相比于英文而言起步

基金项目： 国家 242 项目（2005C48），北京理工大学科技创新计划重大项目培育专项（2011CX01015）

作者简介：罗森林(1968—)，男，博士，教授，E-mail:luosenlin126@126.com.

白建敏(1986—)，男，硕士，E-mail:kongjunbaijianmin@126.com.

较晚，比较有代表性的有：哈工大王晓龙教授领导的课题组[6]提出了一种面向多文档自动摘要任务的多文本框架（Multiple Document Framework，MDF），基于MDF进行信息融合生成摘要，获得了较好的结果。山东大学马军等人[7]提出了一种基于LDA的多文档自动文摘方法，该方法在ROUGE的各个评测标准上均优于SumBasic方法，与其他基于LDA模型的摘要相比也具有优势。

多文档自动摘要过程可以分解为三个任务：主题识别，主题说明，摘要提取。在摘要提取中句子权重的计算是十分重要的。目前，句子重要程度的表述大多采用统计特征，导致特征向量不能很好的表达句子的深层含义，使所选摘要句与主题产生偏差。针对特征向量的局限性本文提出了一种融合句义特征的文摘句抽取策略，通过构建句义结构模型，提取有效句义特征，进而根据综合加权选取法和最大边缘相关（Maximal Marginal Relevance，MMR）[8]方法抽取摘要。本文在多个文本集上进行实验验证了句义特征的有效性，并与两个参照系统进行对比验证了系统的优良性能。

句义结构模型是句义中的成分以及成分之间组合关系的形式化表达。通过该模型将抽象的句义形式化表达为成分之间的数理结构[9]。句义特征是能够表述句子语义的特征，句义结构模型中包含的句子结构信息和语义信息都可以作为特征表述句子的语义。如

图1所示为句义结构模型的基本形式[10]。

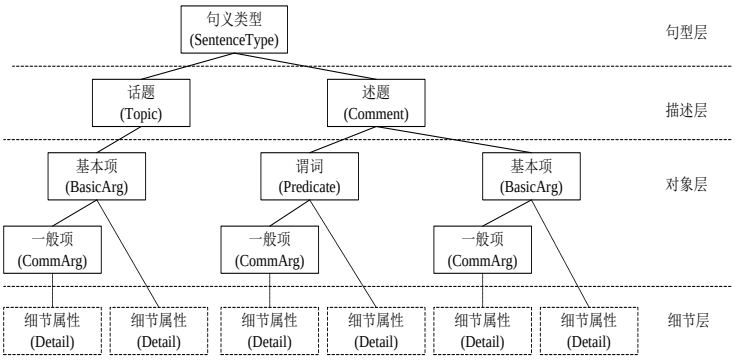

图1 句义结构模型的基本形式

Fig. 1 Basic form of sentential semantic model

1 算法原理

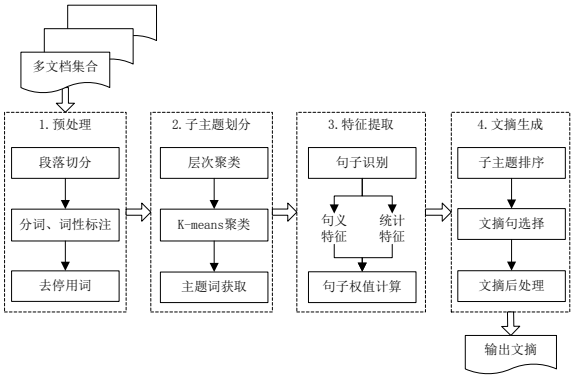

图 2 多文档自动摘要算法原理框架图

Fig. 2 Theory of multi-document summarization algorithm

系统主要包括：预处理、子主题划分、特征提取、文摘生成四个模块。各个模块的具体算法和过程如下。

1.1 预处理

生成摘要的第一步是对多文档集合进行预处理，预处理阶段进行段落切分、分词、词性标注、去停用词等操作。

预处理模块输入多文档集合，首先进行段落切分，切分后对段落进行编号。之后对文本进行分词和词性标注，采用的是中科院计算所的分词工具ICTCLAS2013。另外，为了让子主题划分中的聚类更加准确，要进行去停用词处理，本文

停用词参照《中文停用词库》。

1.2 子主题划分

多文档集合子主题是围绕着中心主题发生的现象、后果以及原因等，是对中心主题不同侧面的描述。

本文采用凝聚式的层次聚类和 K-means 聚类相结合的方法，这样即避免了层次聚类的聚类点错误选择问题又让 K-means 聚类减少人工干预。聚类后得到子主题，通过特征加权和组合词生成与过滤的方法提取关键词得到子主题的主题词[11]。

其中层次聚类的阈值设为s，当类间距离大于阈值时就停止合并，阈值的选择如式 **（1）**所示。

$$s = a\frac{\sum_{i=1}^{N}\sum_{j=i+1}^{N} sim(P_i, P_j)}{\frac{N^2 - N}{2}} \quad (1)$$

其中a为常数，经实验分析，a为 0.8 时聚类数目较为合理，N表示文档中所包含的所有段落数。段落$P_i = (W_{i1}, W_{i2}, \cdots W_{in})$，

$P_j = (W_{j1}, W_{j2}, \cdots W_{jn})$，W 为段落中的有效词，

$sim(P_i, P_j)$表示文档中两个段落之间的相似度，如式（2）所示：

$$sim(P_i, P_j) = \frac{\sum_{k=1}^{n} W_{ik} W_{jk}}{\sqrt{(\sum_{k=1}^{n} W_{ik}{}^2)(\sum_{k=1}^{n} W_{jk}{}^2)}} \quad (3)$$

1.3 特征提取

句子是人理解语言含义的基本单元，所以句子作为摘要抽取的文本单元。特征提取模块分为统计特征提取和句义特征提取两部分，最后依据特征向量计算句子权重。

1.3.1 统计特征提取

综合现有研究成果及影响句子重要程度的因素，选取表 1 所示特征为句子的统计特征：

表 1 句子统计特征

Tab. 1 Statistical features of sentences

特征类别	编号	特征代码	特征描述
统计特征	1	F_ADDRESS	句子在文本中的位置
	2	F_TFIDF	句子中有效词的平均 TFIDF 权值
	3	F_TIPS	句子中是否含有提示词
	4	F_KEYWORD	句子中主题词在所有主题词中的覆盖率

（1）句子位置特征（F_A）：美国的 P. E. Baxendale 分析了 200 个文献段落后发现，段落的论题是段首句的概率为 85%，是段末句的概率为 7%，因此应对这些特殊位置的句子赋予较高的权值，设置权值如下：

$$F_A = \begin{cases} 1.0 & \text{篇章首句} \\ 0.8 & \text{段落首句} \\ 0.6 & \text{篇章尾句} \\ 0.4 & \text{段落尾句} \\ 0.1 & \text{其他} \end{cases} \quad (4)$$

（2）平均词权特征（F_W）：句子是由词语组成的，词语的重要程度必然影响句子的重要程度。词语的重要程度与词语在文档中的出现频率和出现该词语的文档数目有关。为了去除句子长度对句子重要程度的影响，用平均词权来衡量句子的重要程度。

$$F_W = \frac{\sum_{j=1}^{n} W_{ij}}{n_i} \quad (5)$$

式中n_i为第i个句子中有效词的个数，有效词是指去除功能词后有实际含义的词语。W_{ij}为第 i 个句子中第 j 个有效词的权值，权值计算公式如下：

$$W_{ij} = \frac{tf_{ij} \log(N / n_j + 0.01)}{\sqrt{(tf_{ij})^2 [\log(N / n_j + 0.01)]^2}} \quad (6)$$

式中tf_{ij}是词语频率，$\log(N / n_j + 0.01)$是词语倒排文档频率，N是总文档数，n_j为出现该词的文档个数，分母部分是归一化因子。

（3）提示词特征（F_T）：包含提示词的句子往往高度概括了文档主题，是很好的摘要候选句。根据句子中是否含有“总之”、“综上”等提示词设置如下权值：

$$F_{\mathrm{T}}=\begin{cases}1.0 & \text{含提示词}\\ 0.2 & \text{不含提示词}\end{cases} \quad (7)$$

(4) 主题词特征(F_{K})：句子中包含主题词的个数和每个主题词的权重都可以影响句子的重要程度。主题词覆盖率是句子中所包含的主题词的加权和与子主题所有主题词的加权和的比值。

$$F_{\mathrm{K}}=\frac{\sum_{j=1}^{n} w_{ij}}{\sum_{i=1}^{c}\sum_{j=1}^{n} w_{ij}} \quad (8)$$

式中 w_{ij} 表示第 i 个句子的第 j 个主题词的权重，n 为第 i 个句子的主题词的个数，c 为子主题中句子的个数。

1.3.2 句义特征提取

针对统计特征的局限性，本文引入句义特征增强特征向量的表述能力，句义特征的提取采用课题组的研究成果[12]，通过构建句义结构模型，提取句子的主要句义成分，包括：话题基本格、述题基本格、谓词、一般格，然后通过实验筛选有效的句义特征。经分析，从句义结构模型中得到的句义特征如表 2 所示。

表 2 句子句义特征

Tab. 2 Sentential semantic features of sentences

特征类别	编号	特征代码	特征描述
句义特征	1	F_TOPIC	主题词表包含话题基本格
	2	F_PREDICATE	主题词表包含谓词
	3	F_COMMENT	主题词表包含述题基本格
	4	F_COMMARG	主题词表包含一般格

实验选取同一主题多文档集合作为语料，当摘要的压缩比为 15%时，摘要效果最好，所以实验设置压缩比为 15%。

由句义结构模型可知，基本格和谓词构成句义结构的基本框架，一般格是对基本格、谓词、某些一般格的限制和修饰，则基本格和谓词相对于一般格更重要；另外，基本格分为修饰话题的基本格和修饰述题的基本格，而话题是句子描述的对象，述题是对该对象的描述，由语义学知识可知话题基本格更重要。本文通过特征筛选实验衡量句义特征的有效性并验证理论分析的正确性。实验采用准确率、召回率、F 值对摘要进行评价，计算方法如下：

$$P=\frac{K}{N} \qquad R=\frac{K}{M} \qquad F=\frac{2\times R\times P}{R+P}$$

其中，K 为系统生成的摘要句包含在标准摘要中的数目，N 为系统生成的摘要所包含的句子数目，M 为标准摘要所包含的句子数目。

首先加入所有特征，然后按照特征编号从高到低的顺序依次去除句义特征，实验结果如图 3 句义特征筛选实验

Fig. 3 Experiments of sentential semantic feature selection 所示。

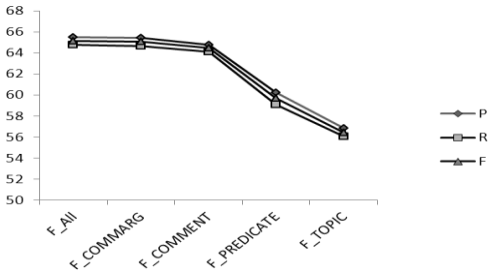

图 3 句义特征筛选实验

Fig. 3 Experiments of sentential semantic feature selection

由图 3 句义特征筛选实验

Fig. 3 Experiments of sentential semantic feature selection 可知，在不断去除句义特征后，摘要的效果越来越差，在去除 F_COMMARG 特征和 F_COMMENT 特征后，系统性能基本不变；在去除 F_PREDICATE 特征后，摘要准确率下降了 5.3 个百分点，召回率下降 5.6 个百分点；去除 F_TOPIC 特征后，摘要准确率下降了近 9 个百分点，召回率下降 8.8 个百分点。由此可得，去除 F_COMMARG 特征和 F_COMMENT 特征在现有的数据源下并没有影响，最终保留 F_TOPIC 特征和 F_PREDICATE 特征。

1.3.3 句子权值计算

由于不同特征的重要程度是不一样的，所以特征提取后要根据每个特征的重要程度获取特征向量的权向量，本文通过层次分析法获取权向量[13]。首先建立层次分析模型，然后通过对语言学的分析与实验，构造成对比较矩阵 A，当矩阵 A 为一致性矩阵时，矩阵的主特征向量就是特征的权向量，设为 $U=\{u_1\ \ u_2\ \ u_3\ \ u_4\ \ u_5\ \ u_6\}$，

各个特征的相对重要性由权向量U的各分量所确定。

本文假定各个特征相互独立，将句子的统计特征和句义特征构成特征向量F={ F_{A} F_{W} F_{T} F_{K} F_{TC} F_{PE} }, 各个特征的加权系数构成权向量U，句子权值如式 （9）所示。

$$W_i = \boldsymbol{F}_i \times \boldsymbol{U}^T \qquad (9)$$

其中 W_i 为第 i 个句子的权值，F_i 为第 i 个句子的特征向量。

1.4 文摘生成

文摘生成模块首先根据句子的权值以及子主题内的句子数目等因素，对子主题进行排序，确定摘要抽取的顺序，之后采取一定策略抽取文摘句，最后进行后处理生成可读性较高的摘要。具体步骤如下：

① 文摘句抽取前对句子进行过滤，将祈使句、问句等不适合作为文摘句的句子去掉；
② 根据有效子主题的权值高低依次选取子主题内权值最高的句子；
③ 检查候选文摘句与已选文摘句话题和谓词是否一致，如果一致，候选文摘句换为该子主题中的下一个候选句子，如果不相同则转步骤④；
④ 检查是否满足文摘压缩比要求，如果没有达到压缩比要求转步骤②，如果满足压缩比要求转步骤⑤；
⑤ 停止选取句子，输出初始文摘进行后处理。

得到初始文摘后首先进行句子排序，然后进行指代消解和平滑润色等后处理，进一步提高文摘的可读性。

2 实验及结果分析

2.1 实验数据源

实验数据来自北京理工大学信息系统及安全对抗实验中心多文档摘要语料库。该语料库主要是 2009 年热点新闻事件的新闻报道，包括 90 个主题，每个主题包含 20-50 篇不等数量的新闻语料，每篇新闻语料包含 20-80 个句子，同时每个主题包含压缩比为 5%，10%，15%的三篇标准摘要。标准摘要均由该实验中心的博士、硕士按照 BFS 语料库构建规范完成，其生成过程如下：首先每 3 人对同一主题文本集提取不同压缩比的人工摘要，然后由自然语言处理小组的 10 名博士、硕士对 3 份人工摘要进行评价并计算平均得分，将平均得分最高的摘要作为标准摘要放入语料库，如果得分相同则都放入语料库。

本文从语料库中随机选取 6 个话题进行实验，具体统计信息如表 3 所示。

实验语料	文档数目	句子数目	摘要得分	主题说明
主题	31	894	4	北京奥运会
主题	31	737	4.6	高校毕业生就业
主题	32	794	4.7	国庆六十周年大阅兵
主题	36	766	3.8	全国众志成城抗冻灾
主题	33	773	4.5	三聚氰胺事件
主题	36	714	4.6	四川 5.12 地震

表 3 实验语料统计信息

Tab. 3 Statistics information of experimental corpus

2.2 评价方法

中文多文档自动摘要的评价目前没有统一标准，大致分为内部评价和外部评价：内部评价是通过直接分析摘要的质量来进行评价，外部评价利用摘要在某一任务中的完成效果来进行评价。为了消除单一主题进行实验造成的评价指标的偶然性结果，引入评价指标算数平均值，即平均准确率$\overline{P}$、平均召回率$\overline{R}$、平均 F 值$\overline{F}$，计算公式如下：

$$\overline{P} = \frac{\sum_{i=1}^{n} P_i}{n} \qquad \overline{R} = \frac{\sum_{i=1}^{n} R_i}{n} \qquad \overline{F} = \frac{\sum_{i=1}^{n} F_i}{n}$$

其中，n 为主题数目，本文 n 为 6，P_i 为第 i 个文本集的摘要准确率，R_i 为第 i 个文本集的摘要召回率，F_i 为第 i 个文本集的摘要 F 值。

2.3 结果及分析

为了验证本文提出的多文档自动摘要系统的有效性，建立了两个对照系统与本文方法进行对比实验。

第 1 个对照系统是基于事件抽取的网络新闻多文档自动摘要系统（MSBEE）[14]，该系统引入事件抽取技术，通过主旨事件抽取及后续处理生成摘要。本文系统与 MSBEE 系统对比结果如

第 2 个对照系统是基于统计特征的多文档

自动摘要系统（MSBSF）[15]，该系统通过聚类进行子主题划分，然后对子主题内句子进行加权求和，根据句子的权值大小进行文摘句抽取。本文系统与 MSBSF 系统对比结果如

表 5 所示。

摘要压缩比在 15%情况下，三个系统的实验数据如

图 4 所示。

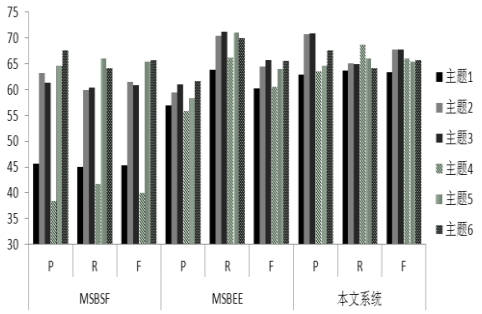

图 4 压缩比 15%时三个系统实验数据

Fig. 4 Experimental data of the systems on condition that the compression ratio is 15%

由

图 4 可见，同一个系统在不同主题下的评价结果会有一定的差异，经分析发现，标准摘要得分与系统的性能正相关。同时，不同的系统对语料的兼容性不同，本文系统兼顾不同语料的能力较优，可见，该方法适用广泛，具有一定的鲁棒性。另外，本文系统生成摘要的平均准确率最高，而且很好地平衡了准确率和召回率，使本文系统生成摘要的整体效果最优。

表 4 本文系统与 MSBEE 系统对比结果（%）

Tab. 4 The comparison with the MSBEE system(%)

压缩比	MSBEE			本文系统		
	$\overline{P}$	$\overline{R}$	$\overline{F}$	$\overline{P}$	$\overline{R}$	$\overline{F}$
5%	-	-	-	63.7	60.7	62.0
10%	-	-	-	65.4	62.3	63.8
15%	**58.9**	**69.2**	**63.6**	**66.7**	**65.5**	**66.0**

由

表 4 可见，在压缩比 15%情况下，本文系统与 MSBEE 系统相比，$\overline{P}$ 提高 7.8%，$\overline{R}$ 降低 3.7%，$\overline{F}$ 提高 2.4%，说明本文系统综合性能比 MSBEE 系统优良，尤其是在平衡准确率和召回率方面表现更好，不会出现准确率和召回率相差很大的情况。

表 5 本文系统与 MSBSF 系统对比结果（%）

Tab. 5 The comparison with the MSBSF system(%)

压缩比	MSBSF			本文系统		
	$\overline{P}$	$\overline{R}$	$\overline{F}$	$\overline{P}$	$\overline{R}$	$\overline{F}$
5%	48.2	47.1	47.6	63.7	60.7	62.0
10%	50.4	50.2	50.1	65.4	62.3	63.8
15%	**58.6**	**57.0**	**57.7**	**66.7**	**65.5**	**66.0**

由

表 5 可见，在不同压缩比情况下，本文系统与 MSBSF 系统相比平均准确率、平均召回率、平均 F 值都有较大幅度提高，其中 $\overline{P}$ 平均提高 12.7%，$\overline{R}$ 平均提高 11.4%，$\overline{F}$ 平均提高 12.1%，实验表明，通过句义结构模型提取句义特征，增加特征向量的维度和句义分析的深度，使提取的摘要更能体现主题含义，从而使本文系统优于

MSBSF。

由本文系统和 MSBSF 系统在不同压缩比下的效果可知，在一定范围内压缩比越大系统的性能越好，原因在于人工抽取标准摘要的随机性比较大，而压缩比提高、数据量变大在一定程度上克服了这种随机性，使得最终得到的摘要更加合理而使评价效果有所提高。

3 结论

融合句义特征的多文档自动摘要方法在传统句子统计特征的基础上加入句义特征，增加了句子的分析深度，使特征向量更能表达句子的含义，使抽取的文摘句更能体现主题含义，实验结果表明本文提出的文摘方法比 MSBEE 系统和 MSBSF 系统的综合性能更加优良。综上，句义结构模型在多文档自动摘要中的应用是有效的，为多文档文摘提出了一种新的思路和方向。下一步研究的重点是引入更多的句义特征并分析其作用，完善句义结构模型的特征体系，提高文摘句抽取的效果，改善文摘的逻辑性和可读性，从而生成更高质量的文本摘要。

参考文献

[1] Wang D, Li T. Weighted consensus multi-document summarization[J]. Information Processing & Management, 2012, 48(3): 513-523.

[2] Radev D R, Jing H, Styś M, et al. Centroid-based summarization of multiple documents[J]. Information Processing & Management, 2004, 40(6): 919-938.

[3] Erkan G, Radev D R. Lexpagerank: Prestige in multi-document text summarization[C]. Proceedings of EMNLP. 2004, 4.

[4] Ding C, He X, Simon H D. On the equivalence of nonnegative matrix factorization and spectral clustering[C]. Proc. SIAM Data Mining Conf. 2005, (4): 606-610.

[5] Arora R, Ravindran B. Latent dirichlet allocation based multi-document summarization[C]. Proceedings of the second workshop on Analytics for noisy unstructured text data. ACM, 2008: 91-97.

[6] 徐永东, 徐志明, 王晓龙. 基于信息融合的多文档自动文摘技术[J]. 计算机学报, 2007, (11): 2048-2054.
XU Yong-dong, Xu Zhi-ming, Wang Xiao-long. Multi-Document Automatic Summarization Technique Based on Information Fusion [J]. Chinese Journal of Computers, 2007, (11): 2048-2054.

[7] 杨潇, 马军, 杨同峰等. 主题模型LDA的多文档自动文摘[J]. 智能系统学报, 2010, (02): 169-176.
Yang Xiao, Ma Jun, Yang Tong-feng, et al. Automatic multi-document summarization based on the latent Dirichlet topic allocation model [J]. Caai Transactions on Intelligent Systems, 2010, (02): 169-176.

[8] 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. ACM, 1998: 335-336.

[9] 冯扬. 汉语句义模型构建及若干关键技术研究[D]. 北京：北京理工大学, 2010.
Feng Yang. Research on Chinese Sentential Semantic Mode and Some Key Problems[D]. Beijing:Beijing Institute of Technology, 2010.

[10] 罗森林, 刘盈盈, 冯扬等. BFS-CTC汉语句义结构标注语料库构建方法[J]. 北京理工大学学报：自然科学版, 2012, 32(03): 311-315.
Luo Sen-lin, Liu Ying-ying, Feng Yang, et al. Method of Building BFS-CTC a Chinese Tagged Corpus of Sentential Semantic Structure[J]. Journal of Beijing Institute of Technology:Natural Science,2012, 32(03): 311-315.

[11] 苏凯. 中文文本关键词提取与自动摘要技术研究[D]. 北京：北京理工大学, 2008.
Su Kai. Chinese Text Keyword Extraction and Automatic Summarization Technology[D].Beijing: Beijing Institute of Technology, 2008.

[12] 罗森林, 韩磊, 潘丽敏等. 汉语句义结构模型及其验证[J]. 北京理工大学学报：自然科学版,2013, 33(2): 166-171.
Luo Sen-lin, Han Lei, Pan Li-min, et al. Chinese Sentential Semantic Mode and Verification [J]. Beijing Institute of Technology :Natural Science, 2013, 33(2): 166-171.

[13] Saaty T L. Decision making with the analytic hierarchy process[J]. International Journal of Services Sciences, 2008, 1(1): 83-98.

[14] 韩永峰, 许旭阳, 李弼程等. 基于事件抽取的网络新闻多文档自动摘要[J]. 中文信息学报, 2012(01): 58-66.

Han Yong-feng, Xu Xu-yang, LI Bi-cheng, et al. Web News Multi-document Summarization Based on Event Extraction [J]. Journal of Chinese Information Processing, 2012 (01): 58-66.

[15] 熊颖. 中文多文档摘要关键技术研究[D]. 北京：北京邮电大学, 2011.
Xiong Ying. Research on Key Technologies of Chinese Multi-Document Summarization [D].Beijing: Beijing University of Posts and Telecommunications, 2011.