·人工智能及识别技术· 文章编号:1000-3428(2018)06-0169-07 文献标志码:A 中图分类号:TP391.1

基于 Topic Signature 的动态文摘更新方法

张 祯,樊兴悦,郭禹田,吴国华

(杭州电子科技大学 网络空间安全学院,杭州 310018)

摘 要: 目前针对动态文摘的研究关注对象主要是多文档集合,其中内容随时间而更新演化,但动态文摘中存在高冗余、新颖信息丢失等问题,会影响文摘提取质量。为此,研究 Topic Signature 模型,并在其基础上提出一种新的整数规划动态文摘更新方法。根据句间相似度对每条语句的主题代表性和信息多样性进行评分,利用 Topic Signature 模型评估语句的新颖性,以提取事件中的更新演进信息。在此基础上,依据摘要生成策略,缩小解的可行域,以保证在短时间内生成高质量的文摘。实验结果表明,该方法无需进行模型训练和语言匹配,能够有效降低时间复杂度,提高文摘提取效率。

关键词: 动态文摘;Topic Signature 模型;密度峰值;整数规划模型;自然语言处理

中文引用格式: 张 祯,樊兴悦,郭禹田,等. 基于 Topic Signature 的动态文摘更新方法[J]. 计算机工程,2018,44(6):169-175.

英文引用格式: ZHANG Zhen,FAN Xingyue,GUO Yutian,et al. Dynamic summarization update method based on Topic Signature[J]. Computer Engineering,2018,44(6):169-175.

Dynamic Summarization Update Method Based on Topic Signature

ZHANG Zhen,FAN Xingyue,GUO Yutian,WU Guohua

(School of Cyberspace Security,Hangzhou Dianzi University,Hangzhou 310018,China)

【Abstract】 The dynamic summarization is to construct evolutionary content of collection. But there are some complicated problems in dynamic summarization,such as information redundancy,novelty information easily lost. To solve the above problems,this paper proposes an Integer Linear Programming(ILP) dynamic summarization update method based on Topic Signature model. According to the similarities between sentences,it calculates the representativeness score and diversity score for each sentence and introduces the Topic Signature model to determine the novelty of the sentences. Based on the summary generation strategy,the feasible region of understanding can be reduced and the high quality abstracts can be generated in a short time. Experiment result shows that the proposed method can effectively reduce the time complexity and improve the efficiency without model training and language matching.

【Key words】 dynamic summarization;Topic Signature model;density peak;Integer Programming Model(ILP);Natural Language Processing(NLP)

DOI: 10.19678/j.issn.1000-3428.0051068

0 概述

自动文摘技术[1]属于自然语言处理(Natural Language Processing,NLP)领域中的一个分支,它起源于二十世纪五十年代后期,被提出的主要原因是人工生成摘要低效耗时,亟需一种自动高效的机器算法提取摘要。自动文摘算法可以分为单文档摘要算法、多文档摘要算法和动态文档摘要算法,其中动态多文档摘要算法是近年来的研究热点。

基于聚类模型的摘要提取算法是较早用于文摘提取的模型之一,其通过挖掘句子之间的关系来抽取信息量较高的句子作为文摘,简洁有效,在工业领域已经得到了广泛的应用。针对该算法,近几年仍有学者提出改进算法。例如:文献[2]利用遗传聚类算法以及计分函数提取与主题最相关的句子作为摘要;文献[3]通过改进 K-means 聚类与图模型相结合的多文本自动文摘方法提高了摘要提取的准确度;文献[4]构建一种基于密度峰值的启发式聚类模型,其具有简洁高效、无需设定初始聚类中心的优点;文献[5]提出了语义层次聚类的多文档自动摘要算法,其结合密度聚类和层次聚类来判定句子所属主题,以提升划分的准确性,进而提高文摘的信息覆盖度。

基金项目: 国家部委基金。

作者简介: 张 祯(1978—),男,讲师、硕士,主研方向为文本挖掘、信息安全;樊兴悦,硕士研究生;郭禹田,硕士;吴国华,研究员。

收稿日期: 2018-01-03 **修回日期:** 2018-02-13 **E-mail:** zhangzhen@hdu.edu.cn

语义分析模型又被称为潜在语义索引模型。该模型对文本中词语之间的关联性进行分析,挖掘其中潜在概念和含义,在进行摘要提取时,其对文档背后的语义结构进行分析和理解,深层次地挖掘文本中的主题含义,是在文摘生成技术中研究最成熟的模型之一。语义模型的相关研究有:文献[6]提出一种提取中文观点句中评价对象和评价词主观匹配关系的方法,大幅提高了摘要提取的准确率;文献[7]提出一种基于潜在狄里克雷分配模型的多文档情感总结方法,有效地识别了关键情感句,同时提高了摘要提取的精度;文献[8]采用相邻词的共现频率进行未登录词识别,提出一种通过词汇链的构建进行中文关键词抽取和文摘生成的算法,解决了文摘内容不完整的问题;文献[9]在层次潜在狄里克雷分配模型的基础上,提出一种层次生成树摘要算法来挖掘历史文档集和更新文档集中的潜在主题结构,清晰地发现2 个文档集合中信息的差异和相同点,以便动态文摘的提取;文献[10]利用维基百科等知识密集的资源建立概念空间,并在其中对词语进行语义解释,进而解释词语文本片段的语义,得到了更好信息覆盖率的文摘。

其他较为典型的文摘提取模型有:文献[11]提出一种子主题区域划分的多文档自动文摘算法,解决了词频矩阵维数过大和过于稀疏的问题;文献[12]将文摘的提取转换为信息距离的计算问题,实现了动态文摘提取;文献[13]通过深度学习提取文档摘要,使其可从多个文档中提取适当的文摘集;文献[14]通过建立本质向量模型,提高了所提取代表信息的有效性;文献[15]利用事件间的关系构建事件网络文本表示模型,提高了摘要提取的准确率和召回率;文献[16]提出一种基于指代消解和篇章结构分析的自动摘录算法,提高了算法在受限金融领域中文本自动摘录的效率。

针对动态文档中高冗余和关键信息易丢失的现状,本文提出一种基于 Topic Signature 的整数规划模型动态文摘方法。利用 Topic Signature 模型提取元组新颖度,并通过新颖度评分算法对新颖度信息含量大小进行评估,同时增加 3 个约束条件,从而缩小解的可行域,提高算法效率。

1 基于 Topic Signature 的新颖元组提取

1.1 Topic Signature 模型

Topic Signature[17-18]基于统计学模型,主要应用于文档中主题词语的提取。该模型基于文档中描述同一主题的词语之间具有很强的关联性这一特性,从词语的实际分布来判断它们的主题代表性,效率优于以往的摘要提取算法(如 MMR baseline 和 Best Sys)。本文将 Topic Signature 模型应用于文档中元组的新颖度评分。基于假设“语义相关的词组共同出现的概率较大”,本文采用更适用于数据稀疏的情况的对数似然比的统计模型来提取主题标志词语。下文将阐述提取主题词语并进行权重估计的具体步骤。

对于同一话题的相关报道,TAC 语料库将其分为时间较早的文档集和时间较新的文档集。为评估时间较新文档集中二元组的新颖程度,本文在 Topic Signature 模型的基础上,提出了 2 条新的假设:

假设 1 $p(D_{\text{old}} \mid uts) = p = p(D_{\text{new}} \mid uts)$,*uts* 在 D_{old}和 D_{new}中出现的概率相同。

假设 2 $p(D_{\text{old}} \mid uts) = p_1 \neq p_2 = p(D_{\text{new}} \mid uts)$,*uts* 在 D_{old}和 D_{new}中出现的概率不同,*uts* 具有较强的新颖性。

文档集中的词频分布如图 1 所示,其中,D_{old}代表时间较早的文档集合,D_{new}代表时间较晚的文档集合,*uts* 代表 D_{old}中的二元组,*uts*′代表除 *uts* 之外的所有二元组,O_{11}代表二元组 *uts* 出现在 D_{old}中的频率,O_{12}代表二元组 *uts* 出现在 D_{new}中的频率,O_{21}代表 *uts*′出现在 D_{old}中的频率,O_{22}代表 *uts*′出现在 D_{new}中的频率。

	D_{old}	D_{new}
uts	O_{11}	O_{12}
uts′	O_{21}	O_{22}

图 1 词频分布

下面评估元组的新颖度。计算每一个元组的 $-2\log_a\lambda$ 值,假设元组出现的概率符合式(1)所示的二项式分布。

$$p(X=k) = \binom{n}{k} p^k (1-p)^{(n-k)} = b(k;n,p)$$
$$k = 0,1,\cdots,n \tag{1}$$

则假设 1 和假设 2 的似然度分别可表示为:

$$L(H_1) = b(O_{11};O_{11}+O_{12},p)\,b(O_{21};O_{21}+O_{22},p) \tag{2}$$
$$L(H_2) = b(O_{11};O_{11}+O_{12},p_1)\,b(O_{21};O_{21}+O_{22},p_2) \tag{3}$$

如果一个元组是用来描述事件新出现的主题或者原有主题的发展情况,那么该元组在时间较新的文档集 D_{new}中出现的概率 p_2 就会远大于在 D_{old}中出现的概率 p_1,即 $p_2 \gg p_1$,也就是说假设 2 的似然度 $L(H_2)$远大于假设 1 的似然度 $L(H_1)$,即$L(H_2) \gg L(H_1)$。因此,假设 2 和假设 1 具有较大的似然度比值,也就说明该 *term* 属于发展主题或者新出现主题中的元组,具有较高的新颖度。

基于上述的思想,对于时间较新文档集合中的元组 $term_i$(包括单元组、二元组和三元组),本文给出的新颖度计算方法,如式(4)所示。

$$Nov(term_i) = -2\log_a \frac{L(H_1)}{L(H_2)} = 2N \times I(o;term_i) \quad (4)$$

1.2　新颖元组提取

在推导出元组新颖度评分公式后,通过算法 1 详细描述了提取新颖度较高的元组作为主题词的过程。

算法 1　新颖元组的提取算法

输入　同一新闻事件不同时间的 2 个文档集 D_{old}和 D_{new}

输出　新颖度较高的单元组集合 $U = \{ < u_1, w_1 >, < u_2, w_2 >, \cdots < u_n, w_n > \}$,二元组集合 $B = \{ < b_1, w_1 >, < b_2, w_2 >, \cdots < b_n, w_n > \}$,三元组集合 $T = \{ < t_1, w_1 >, < t_2, w_2 >, \cdots < t_n, w_n > \}$

1)TAC 2008 语料库已将每一个新闻事件的相关报道,按照时间先后顺序划分为 A、B 集合,读取 A 中的文档集为 D_{old},B 中的文档集为 D_{new}。

2)利用滑动窗口算法统计 D_{old}和 D_{new}中所有二元组 b_i 出现的频率。

3)根据 $Nov(term_i)$计算元组的新颖度分值,并由高到低进行排序。

4)通过 χ^2 统计表选取在置信度为 $\alpha = 0.001$ 时的截断分值 φ,并选取新颖度较高的单元组、二元组以及三元组。

5)将新颖度较高的元组以及相应的得分整理到集合中。

6)输出新颖度较高的单元组集合 $U = \{ < u_1, w_1 >, < u_2, w_2 >, \cdots < u_n, w_n > | w_i > \varphi \}$、二元组集合 $B = \{ < b_1, w_1 >, < b_2, w_2 >, \cdots < b_n, w_n > | w_i > \varphi \}$和三元组集合 $T = \{ < t_1, w_1 >, < t_2, w_2 >, \cdots < t_n, w_n > | w_i > \varphi \}$。

1.3　句子新颖度评分

已知:当一个语句中含有新颖度较高二元组时,则该语句的新颖度也较高。本节提出句子的新颖度评分模块,具体如下:首先通过算法 1 提取新颖度较高的二元组集合 B,然后定义映射函数 f,以便将新颖度较高的二元组及其似然比值逐一映射。f 函数描述如下:

$$f(b_i) = w_i, \forall b_i \in B \quad (5)$$

因为二元组相比于单元组和三元组,其所包含的语义信息更多,所以本文将句子中所包含的二元组新颖度评分之和作为句子的新颖度得分,计算方法如下:

$$s^{nov}(i) = \sum_{j}^{|B|} f(b_j) Occ_{ij}, \forall b_j \in s_i \quad (6)$$

其中,$s^{nov}(i)$是语句 s_i 的新颖度得分,b_j 是 s_i 中的二元组,Occ_{ij}是布尔变量,当二元组 b_j 出现在句子 s_i 中的时候值为 1,否则为 0,$|B|$是新颖二元组集合中的二元组数量总和。句子新颖度评分模块是对文档集中每条语句的新颖信息含量大小进行评估。评分越高的句子,其包含的新颖信息也就越多。

2　基于密度峰值的语句信息量评估

2.1　主题代表性分值

已知:当某句和同主题下的其他语句相似性较大时,该句的信息量较大,主题代表性较强。基于此,文献[19]受基于密度峰值的启发式聚类算法(DPC)的启发,提出了一种基于句子密度的多文档摘要抽取方法(DPC-based)。该方法能够有效提取多文档集中的关键信息。本文借鉴其设计思想,首先将文本中的句子转换为向量空间中的数据点,然后利用 DPC 算法中的密度公式来计算每条语句的密度,密度越大的语句其主题代表性也就越强,具体计算公式如下:

$$s^{rep}(i) = \frac{1}{K} \sum_{j=1, j \neq i}^{K} \chi(sim_{ij} - \delta) \quad (7)$$

$$\chi(x) = \begin{cases} 1, x > 0 \\ 0, \text{others} \end{cases} \quad (8)$$

其中,sim_{ij}代表句子 s_i 和句子 s_j 的相似度,取值为 0~1,K 代表文档集中的句子总量,δ 代表一个预先设定的截断阈值,用于过滤和 s_i 相似度较低的语句。

2.2　信息差异性分值

已知:当某句和其他主题的中心句相似性较小时,该句的语义差异性较强,信息冗余度较低。基于此,可以将句子 s_i 和其他主题中心句的距离作为语义差异性分值,距离越大则差异性越强,信息冗余度也越小,具体计算公式如下:

$$s^{div}(i) = \begin{cases} 1 - \max\limits_{j; s^{rep}(j) > s^{rep}(i)} sim_{ij} \\ 1 - \min\limits_{j \neq i} sim_{ij} \end{cases} \quad (9)$$

句子 s_i 的差异性得分就是在密度(主题代表性得分)比它大的句子中寻找和它最为相近的主题句,其中最大相似度和 1 的差值即为句子 s_i 的信息差异性得分。由于 $0 \leqslant sim_{ij} \leqslant 1$,因此 $0 \leqslant s^{div}(i) \leqslant 1$。该算法和 MMR[12]冗余过滤模型非常相似,都是基于贪心的思想,每次选取和候选句子集合中差异性最大的句子作为文摘句。但不同的是 MMR 模型只和已选文摘句进行比较,而本文算法是和文本中所有句子进行比较,因此,其精确度更高。

2.3　有效长度分值

定义 1　句子有效长度就是其所包含的实词数量。

已知:若句子的主题多样性分值和信息差异性分值都较大,且有效长度较短,那么其信息量也就越高,则倾向将其选为文摘句。计算公式如下:

$$s^{len}(i) = \frac{el(s_i)}{\max\limits_{j=1}^{K} el(s_j)} \times \log_a \frac{\max\limits_{j=1}^{K} rl(s_j)}{rl(s_i)} \quad (10)$$

其中,$el(s_i)$是句子的有效长度,$rl(s_i)$是句子的真实长度。动态文摘的一个重要评价指标,是文本压缩

比,在信息量不变的条件下,文本压缩比越高,摘要的质量越好。

3 基于 Topic Signature 的整数规划文摘提取

3.1 算法描述

本文提出的基于 Topic Signature 的整数规划(下文用 ILP-TS 代替)的文摘算法,首先定义二元组的信息熵,进而定义了文摘的信息熵,然后将获取到的具有最大信息熵的摘要作为整数规划的目标函数。该模型包括 2 个重要的步骤:1)选取概念词汇;2)设置词汇的相应权重。对于二元组信息量的评估,文献[20]构建了一种回归模型。该模型需要通过一系列的特征进行训练,如二元组在句子中的位置信息、在新文档集中的频率等。本文引入的 Topic Signature 模型对时间较新文档集中的新颖词汇进行提取,能够通过似然度比值对二元组的新颖度进行量化,相对于回归模型更为简洁有效,并且不需要根据语料进行相应特征的选取。本文以获取最大二元组新颖信息量作为目标函数,将文摘的最低代表性得分、最低多样性得分以及最低有效长度得分作为规划模型的约束条件,以保证提取到的摘要具有较高的主题覆盖率和较低的信息冗余性,不仅提高了文摘质量,缩小了解的可行域,也提高了求解速度。具体计算公式如下:

$$\max\ z=\sum_{i}^{|sum|}\sum_{j}^{|B|}f(b_j)o_{ij}$$

$$\text{s.t.}\begin{cases}\sum_{i=1}^{n}r_is_i\geqslant\alpha\times REP & \text{(a)}\\ \sum_{i=1}^{n}d_is_i\geqslant\beta\times DIV & \text{(b)}\\ \sum_{i=1}^{n}l_is_i\geqslant\eta\times LEN & \text{(c)}\\ \sum_{i=1}^{n}l'_is_i\leqslant L & \text{(d)}\\ s_io_{ij}\leqslant c_j & \text{(e)}\\ \sum_{i=1}^{n}s_io_{ij}\geqslant c_j & \text{(f)}\\ s_i\in\{0,1\},\forall i & \text{(g)}\\ c_i\in\{0,1\},\forall j & \text{(h)}\end{cases}\tag{11}$$

其中,s_i 和 c_j 都是 0-1 整型变量,分别代表第 i 个句子以及第 j 个二元组是否会在提取的文摘中出现,1 代表会出现,0 代表不会出现,b_j 是新颖二元组的第 j 个元组,即 $b_j\in B$,o_{ij}是布尔变量,其值为 0 代表第 i 个句子不包含第 j 个二元组,为 1 则代表第 i 个句子包含第 j 个二元组,r_i 是第 i 个句子的主题代表性得分,d_i 是第 i 个句子的信息多样性得分,l_i 是第 i 个句子的有效长度得分,$2l'_i$是第 i 个句子的实际长度,REP、DIV、LEN 分别是时间较晚文档中所有句子的主题代表性、信息多样性以及有效长度的分值总和,不等式(a)~(d)的含义是文摘长度最大为 L,并且文摘的代表性、多样性以及有效长度得分不能小于一定的阈值,α、β、η 为相应的调节因子,不等式(e)、(f)表示当某条语句被选为文摘句时,此条语句中的所有二元组都会包含到文摘当中,同时如果某个二元组在文摘当中出现的话,说明至少有一个包含该二元组的语句被选中作为摘要句,$|sum|$是提取的动态文摘中所包含的句子数,$|B|$是集合 B 中二元组总量。

将决策变量、资源变量和约束变量用矩阵表示为:

$$\boldsymbol{S}=\begin{pmatrix}\boldsymbol{s}_1\\ \boldsymbol{s}_2\\ \vdots\\ \boldsymbol{s}_n\end{pmatrix},\boldsymbol{A}=\begin{pmatrix}\boldsymbol{r}_1 & \boldsymbol{r}_2 & \cdots & \boldsymbol{r}_n\\ \boldsymbol{d}_1 & \boldsymbol{d}_2 & \cdots & \boldsymbol{d}_n\\ \boldsymbol{l}_1 & \boldsymbol{l}_2 & \cdots & \boldsymbol{l}_n\end{pmatrix},C=\begin{pmatrix}\alpha\times REP\\ \beta\times DIV\\ \eta\times LEN\end{pmatrix}$$

其中,S 表示决策,A 表示约束,C 表示资源。将 3 个矩阵代入式(11),转换为式(12),计算公式如下:

$$\max\ z=\sum_{i}^{|sum|}\sum_{j}^{|B|}f(b_j)o_{ij}$$

$$\text{s.t.}\begin{cases}AS\geqslant C & \text{(a)}\\ \sum_{i=1}^{n}l'_i\boldsymbol{s}_i\leqslant L & \text{(b)}\\ \boldsymbol{s}_i\,o_{ij}\leqslant c_j & \text{(c)}\\ \sum_{i=1}^{n}\boldsymbol{s}_i\,o_{ij}\geqslant c_j & \text{(d)}\\ \boldsymbol{s}_i\in\{0,1\},\forall i & \text{(e)}\\ c_j\in\{0,1\},\forall j & \text{(f)}\end{cases}\tag{12}$$

转换后的形式更适用于整数规划模型框架,将其代入开源的求解器当中进行求解,若可行解不止一个的时候,选取能够使目标函数达到最大值的可行解作为最优解。例如,最优解向量为 $s_{\text{opt}}=(1,1,\cdots,0)$,那么最终选取解中值为 1 的语句 s_1 和 s_2 作为最终的文摘。

ILP-TS 动态文摘提取算法的详细步骤如下:

算法 2 ILP-TS 动态文摘提取

输入 时间较早的文本集 D_{old},时间较晚的文本集 D_{new}

输出 动态文档摘要 $Summ_{\text{update}}=\{s_1,s_2,\cdots,s_n\}$

1)通过对主题代表分值、信息差异分值和有效长度分值的计算,可得 $S=\{<s_1^{\text{rep}},s_1^{\text{div}},s_1^{\text{len}}>,<s_2^{\text{rep}},s_2^{\text{div}},s_2^{\text{len}}>,\cdots,<s_n^{\text{rep}},s_n^{\text{div}},s_n^{\text{len}}>\}$。

2)统计 D_{new}中元组出现的频率,并通过算法 1 提取新颖度较高的二元组集合 $B=\{<b_1,w_1>,<b_2,w_2>,\cdots,<b_n,w_n>\}$。

3)根据集合 S、S^{nov}以及 B 建立 0-1 整数规划模型,并通过 glpk 求解器进行求解,筛选标志位为 1 的句子作为文摘句,抽取的文摘为 $Summ_{\text{update}}=\{s_1,s_2,\cdots,s_n\}$。

3.2 时间复杂度分析

本节将对 ILP-TS 算法的时间复杂度进行分析。

从上面的算法流程可知,ILP-TS 算法主要可以分为 3 个步骤,下面对每一步进行分析。

步骤 1　对 D_{new} 文档集中每条语句进行综合评分。假设句子的总量为 K,构建句子相似值矩阵 $\boldsymbol{M}_{sim}$ 的时间复杂度为 $O(K^2)$。从主题代表性、信息多样性以及有效长度 3 个方面对句子评分所需的时间复杂度均为 $O(K)$,可得总的时间复杂度近似为 $O(K^2)$,即 $O(K^2)+O(K)+O(K)+O(K)\sim O(K^2)$。

步骤 2　提取 D_{new} 文档集中新颖度较高的二元组集合。假设语句的平均长度为 L_{avg},则 D_{new} 中二元组的数量总和不超过 $K\times(L_{avg}-1)$,所以,计算所有二元组新颖度评分的时间复杂度范围为 $O(K(L_{avg}-1))\sim O(K)$。

步骤 3　抽取 D_{new} 中的语句并生成文摘。若文摘的长度 L 较大,则选用句子综合评分摘要生成模块,该模块的时间复杂度为 $O(K)$,那么 ILP-TS 动态摘要算法的时间复杂度为 $O(K^2)$。当文摘长度 L 较短时,则可以选用基于 0-1 整数规划模型的文摘生成模块,该模块的时间复杂度和文档集合 D_{new} 中的句子总数以及生成文摘的长度有关。例如:TAC 2008 语料库中共含有 48 个新闻主题,B 文档集合中每个主题下有 10 篇报道,句子的总数量约为 800 条,如果文摘长度为 100 词,则需要提取大概 5 条句子,如果每条句子的平均长度为 20 词,因此,可行解的数量为 $C_{800}^{5}=2.7\times10^{12}$;如果文摘的长度为 250 词,则可行解的数量为 $C_{800}^{12}=4.2\times10^{36}$。通过以上的分析可知,0-1 整数规划模型的文摘生成模块所耗费的时间,会随着文摘长度的增加呈现指数增长,为了降低文摘抽取所耗费的时间,选择设置求解器来求近似最优解。

通过以上分析可知,ILP-TS 是一种非监督式文摘提取算法,它不需要语料库进行模型训练,也不需要词网知识和维基百科进行语言匹配,可以适用于任何单独以及混合语种文章摘要的提取。同时,该算法所耗费的时间较低。通过与基于潜在狄利克雷模型的文摘生成系统,以及其他效果最好的理解式文摘系统的对比实验,可得出本文提出的 ILP-TS 文摘算法具有较强的可比性,并且具有简洁有效,易于在工业领域运用等优点,从而具有较高的实际应用价值。

4　实验结果与分析

4.1　实验数据

在 TAC 2008 中,Update Summarization 任务的实验语料中的 48 个话题均来自 AQUAINT-2 新闻文档集合(AQUAINT-2 是 LDC 的一个子集,它包含了来自新华社、纽约时报等多家媒体从 2004 年10 月—2006 年 3 月的新闻报道,每个话题选取 2 个时间点,每个时间点抽取 10 篇文档来进行描述。)。

TAC 2008 语料库中的文档用类似于 XML 的一种结构化数据来进行标记处理,避免了文摘系统进行断句以及标点和段落的识别,同时也能够直接获取文档的话题。TAC 语料库中所使用到的标签含义具体如表 1 所示。

表 1　TAC 中使用的标签说明

标签	含义
< TITLE >	文章标题
< DOC >	文章 id 与类型属性
< HEADLINE >	文章标题
< DATELINE >	文章写作时间
< P >	文章句子
< NARRATIVE >	文章概要

系统评测方法是自然语言处理领域的一个重要组成部分,同样也是文摘评测客观性的保障。常见的评测准则有内部评测和外部评测 2 种:内部评测通过和人工提取的参考摘要进行重合度比对评估;外部评测将文摘系统用于完成特定的任务,根据任务完成的质量来对系统的优劣进行评估,只适用于特定的领域。

内部评测方法本文选用著名的基于 n-gram 共现的 ROUGE 评测方法对生成的摘要质量评估。该方法将机器生成的摘要和人工生成的摘要进行比对,通过比较两者之间相同词序单元的共现率来评测抽取摘要的质量,可以从查全率、查准率和 F 值 3 个方面进行,查全率、查准率的具体评判公式如下:

$$R=\frac{\sum_{S\in\{RefSum\}}\sum_{gram_n\in S}Count_{\text{match}}(gram_n)}{\sum_{S\in\{RefSum\}}\sum_{gram_n\in S}Count(gram_n)} \tag{13}$$

$$P=\frac{\sum_{S\in\{RefSum\}}\sum_{gram_n\in S}Count_{match}(gram_n)}{\sum_{S\in\{ExtractSum\}}\sum_{gram_n\in S}Count(gram_n)} \tag{14}$$

其中,R 为查全率,P 为查准率,$Count_{\text{match}}(gram_n)$ 是 $gram_n$ 在人工摘要和机器提取的摘要中出现的最大次数。查准率用于衡量机器提取的相关信息量在机器提取的所有信息中所占的比例,而查全率是用于衡量机器提取的相关信息在人工摘要信息中的比例。

F 值用于衡量摘要的可读程度以及连贯程度,它结合了查准率和查全率,计算公式如下:

$$F=\frac{2\cdot R\cdot P}{R+P} \tag{15}$$

ROUGE 的参数设置为:$-n=4$;$-r=1\ 000$;$-p=0.5$;$-l=100$。其中,$-n$ 是元组在自动摘要和一组人工摘要中共现的最大值,$-r$ 是重复取样的最小值,$-p$ 是查准率和回召率在查全率中的权重系数,$-l$ 是截取机器摘要的前 n 个字符和人工摘要进行对比评估。

ROUGE 从 2004 开始就已作为 DUC 协会的官方评价标准,本文主要选用 ROUGE-1 和 ROUGE-2 指标。虽然上述自动评估方法具有一定的客观性,并且效率较高,但仍然会受到专家的主观影响,很难把某个人工摘要作为最佳文摘,所以 TAC 提供了 A-H 6 种人工摘要作为参考,以便更为准确地反映文摘系统的提取效果。针对内部评价的上述不足点,本文同时采用更能反映用户需求、基于特定任务的外部评测方法,与 MMR baseline 和 Best Sys 两种效果较好的文摘提取算法做了效果对比,能够更有效地对文摘系统的性能进行评价。

4.2 结果分析

以下是文摘提取实验及数据分析:

本节选取 TAC 2008 语料库中的 2 个新闻话题,应用本文提出的 ILP-TS 算法生成摘要,并和人工理解式参考摘要进行对比,提取结果如下:

ID:D0801A-B

Title:Airbus announces delay in delivering new superjumbo A380

ILP-TS Summary(50 words)

Airbus announced it is two to six months behind its promised A380 delivery schedule.(摘取自 AFP _ ENG _ 20051005 文档)

Airlines are seeking millions in damages contractually agreed upon for delays.(摘取自 APW_ENG_20051113 文档)

A wing failure during testing may delay the A380's safety certification.(摘取自 APW_ENG_20060326 文档)

The U. S. is investigating whether the jet will need special safety regulations due to the turbulence it generates.(摘取自 APW_ENG_20060329 文档)

Manual Summary(50 words)

Airbus announced it is two to six months behind its promised A380 delivery schedule. Airlines are seeking millions in damages contractually agreed upon for delays. A wing failure during testing may delay the A380's safety certification. The U. S. is investigating whether the jet will need special safety regulations due to the turbulence it generates.

ID:D0802A-B

Title: Scientists fear collapsing ice shelves will hasten higher sea levels

ILP-TS Summary(50 words)

23 countries recently met in Greenland to encourage governments to act on global warming instead of arguing.(摘取自 AFP_ENG_20051207 文档)

Worldwide,countries are agreeing that the U. N.(摘取自 APW_ENG_20050823 文档)

Kyoto Protocol, which commits 39 nations to reduce greenhouse gases,is not enough.(摘取自 NYT_ENG_20050830 文档)

Infrastructure in Russia is being damaged due to the melting permafrost including airports,gas pipelines and railways.(摘取自 NYT_ENG_20051019 文档)

Manual Summary(50 words)

23 countries recently met in Greenland to encourage governments to act on global warming instead of arguing. Worldwide, countries are agreeing that the U. N. Kyoto Protocol,which commits 39 nations to reduce greenhouse gases, is not enough. Infrastructure in Russia is being damaged due to the melting permafrost including airports, gas pipelines and railways.

在话题 D0801A 的提取实验中,ILP-TS 提取的摘要和人工提取的参考摘要都涵盖了主要的 3 点内容:1)因为飞翼的问题,导致 A380 的安全认证成功时间推迟;2)空客 A380 的交付日期比事先承诺的晚 2 个月 ~6 个月的时间;3)由于交付日期的推迟,会造成空客公司数百万美元的损失。

ILP-TS 提取的摘要和人工提取的摘要在 ROUGE-1 评判准则下的重合率为 48.9%,充分说明了该方法能够提取出文档中的主要信息。

ILP-TS 与 MMR baseline 和 Best Sys 文摘提取方法做了效果对比,在 TAC2008 上的评测对比情况如表 2 所示。

表 2 TAC2008 数据集上 4 种文摘算法性能对比

方法	ROUGE-1 指标	ROUGE-2 指标
TAC 基本标准	0.299	0.065
MMR baseline 方法	0.318	0.065
Best Sys 方法	0.375	0.101
ILP-TS 方法	0.325	0.084

从以上评测结果可知,本文提出的 ILP-TS 相较于 MMR baseline 方法在 ROUGE-1 和 ROUGE-2 指标上分别高出 2.2% 和 29.2%,并且在与 TAC 2008 比赛中提取效果最好的 Best Sys 文摘系统的对比下,ILP-TS 仍然具有一定的可比性。

5 结束语

本文提出的基于 Topic Signature 的整数规划模型,充分考虑了文摘质量、解的可行域和求解速度等因素,无需对语料库进行模型训练,也不必进行词网知识和维基百科的语言匹配,可以适用于单独以及混合语种文章摘要的提取,解决了目前文摘算法时间复杂度高和只适用于特定领域的问题,简洁高效,并且具有良好的实用性。下一步拟从提高准确性角度对该方法进行改进,并将其应用于工业领域。

参考文献

[1] KANITHA K D, MUBARAK D M. An overview of extractive based automatic text summarization systems[J]. International Journal of Computer Science and Information Technology,2016,8(5):33-44.

[2] SUAREZ B S, LEON E. Genetic clustering algorithm for extractive text summarization[C]//Proceedings of IEEE Symposium on Computational Intelligence. Washington D. C., USA:IEEE Press,2016:949-956.

[3] 赵美玲,刘胜全,刘　艳,等. 基于改进 K-means 聚类与图模型相结合的多文本自动文摘研究[J]. 现代计算机,2017(17):26-30.
[4] RODRIGUEZ A,LAIO A. Clustering by fast search and find of density peaks[J]. Science,2014,344(6191):1492-1496.
[5] 胡　立. 基于语义层次聚类的多文档自动摘要研究[D]. 广州:华南理工大学,2014.
[6] 樊　娜,蔡皖东,赵　煜. 基于最大熵模型的观点句主观关系提取[J]. 计算机工程,2010,36(2):4-6.
[7] XUN J, LIU P, YANG Y, et al. Multi-document sentiment summarization based on LDA model[J]. Journal of Computational Information Systems, 2014, 10(15):6389-6399.
[8] 蒋效宇. 基于关键词抽取的自动文摘算法[J]. 计算机工程,2012,38(3):183-186.
[9] ZHU T,LI K. The similarity measure based on LDA for automatic summarization[J]. Procedia Engineering, 2012,29:2944-2949.
[10] 陈　燕,龙建勋. 基于明确语义分析的自动文摘算法[J]. 计算机工程,2011,37(3):183-185.
[11] 王　萌,徐　超,李春贵,等. 基于子主题区域划分的多文档自动文摘方法[J]. 计算机工程,2011,37(12):158-160.
[12] LI C,LIU Y,ZHAO L. Improving update summarization via supervised ILP and sentence reranking[C]//Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies. [S.l.]:Association for Computational Linguistics,2015:1317-1322.
[13] PADMAPRIYA G,DURAISWAMY K. An approach for text summarization using deep learning algorithm[J]. Journal of Computer Science,2014,10(1):1-9.
[14] CHEN K Y, LIU S H, CHEN B, et al. A novel paragraph embedding method for spoken document summarization[C]//Proceedings of Signal and Information Processing Association Summit and Conference. Washington D. C., USA:IEEE Press,2017:1-6.
[15] 廖　涛,刘宗田,王先传. 基于事件的多主题文本自动文摘方法[J]. 计算机工程,2013,39(3):236-240.
[16] 郑　诚,刘福君,李　清. 基于指代消解和篇章结构分析的自动摘录算法[J]. 计算机工程,2012,38(16):170-173.
[17] HONG B, KIM Y, SANG H L. Company name discrimination in Tweets using topic signatures extracted from news corpus[J]. Journal of Computing Science and Engineering,2016,10(4):128-136.
[18] GAUTRAIS C,CELLIER P,QUINIOU R,et al. Topic signatures in political campaign speeches[C]//Proceedings of Conference on Empirical Methods in Natural Language Processing. New York, USA: ACM Press,2017:2342-2347.
[19] ZHANG Y,XIA Y,LIU Y,et al. Clustering sentences with density peaks for multi-document summarization[C]//Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. [S. l.]: Association for Computational Linguistics,2015:1262-1267.
[20] LI C,QIAN X,LIU Y. Using supervised bigram-based ILP for extractive summarization[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. [S. l.]: Association for Computational Linguistics,2013:1004-1013.

编辑　金胡考

(上接第 168 页)

参考文献

[1] 王东波. 有标记联合结构的自动识别研究[D]. 南京:南京师范大学,2003.
[2] 王东波,陈小荷,年洪东. 基于条件随机场的有标记联合结构自动识别[J]. 中文信息学报,2008,22(6):3-7.
[3] 王东波. 基于规则的单层单标记联合结构的自动识别[J]. 文教资料,2008(9):29-31.
[4] 苗艳军,李军辉,周国栋. 统计和规则相结合的并列结构自动识别[J]. 计算机应用研究,2009,26(9):3403-3406.
[5] 吴云芳. 面向中文信息处理的现代汉语并列结构研究[M]. 北京:北京师范大学,2013:15-16.
[6] 吴云芳. 并列成分中心语语义相似性考察[J]. 当代语言学,2005(4):305-315,379.
[7] 吴云芳. 并列结构的外部句法特征[C]//2002 年全国机器翻译研讨会论文集. 北京:中国中文信息学会,中国人工智能学会,2002:110-116.
[8] 王　浩,姬东鸿,黄江平. 基于隐结构感知的并列名词短语识别研究[J]. 计算机工程,2017,43(4):217-221,227.
[9] 石　翠,周俏丽,张桂平. 面向中文专利文献的有标记并列结构的统计分析[J]. 中文信息处理,2013,27(5):43-50,59.
[10] 石　翠,王　杨,杨　彬,等. 面向中文专利文献的单层并列结构识别[J]. 现代图书情报技术,2014(10):76-83.
[11] 石　翠. 面向中文专利文献的单层并列结构识别[J]. 软件,2014,35(3):75-78,81.
[12] 黄曾阳. HNC(概念层次网络)理论[M]. 北京:清华大学出版社,1998.
[13] 苗传江. HNC(概念层次网络)理论导论[M]. 北京:清华大学出版社,2004.
[14] ZHU Xiaojian, JIN Yaohong. Hierarchical semantic-category-tree model for chinese-english machine translation[J]. China Communications,2012,9(12):80-92.
[15] 晋耀红. HNC(概念层次网络)语言理解技术及其应用[M]. 北京:北京科学出版社,2006.
[16] 朱　筠. 基本句群处理及其在汉英专利机器翻译中的应用[D]. 北京:北京师范大学,2012.

编辑　刘　冰