计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190500028

基于专利结构的中文专利摘要研究

束云峰　王中卿
苏州大学计算机科学与技术学院　江苏 苏州 215006
(1627405102@stu.suda.edu.cn)

摘　要　文本摘要任务旨在通过对原文进行压缩提炼,得出简明扼要的内容描述。针对中文专利文本,提出了一种基于PatentRank算法生成专利摘要的算法。首先,对候选句群做冗余处理,以去除候选句群中相似度较高的句子;然后,对专利中的权利要求书和说明书构建3种不同的相似度计算方法,以计算句子之间的影响权重;最后,选取权值高的句子输出,并将其作为专利的摘要。该算法在选取的数据集中取得了较好的效果。实验结果表明提出的算法相比于已有方法在ROUGE值上有显著提高。

关键词:文本摘要;专利;相似度计算;中文信息处理;PatentRank
中图法分类号　TP391

Research on Chinese Patent Summarization Based on Patented Structure

SHU Yun-feng and WANG Zhong-qing
School of Computer Science and Technology,Soochow University,Suzhou,Jiangsu 215006,China

Abstract　Text summarization aims to provide a concise description of the content by compressing and refining the original text. For the Chinese patented text,an algorithm for generating patent summarization based on the PatentRank algorithm is proposed. Firstly,the candidate sentence groups are redundantly processed to remove the sentences with high similarity in the candidate sentence groups. Then,three different similarity calculation methods are constructed for the patent claims and descriptions to calculate the weights between sentences. Finally,the sentence with high weight is selected as the summarization of the patent. The algorithm has achieved good results in the selected datasets. Experimental results demonstrate that the proposed method substantially outperforms existing approaches in terms of ROUGE measurement.

Keywords　Text summarization,Patent,Similarity calculating,Chinese information processing,PatentRank

1　引言

随着信息时代的快速发展,互联网上的数据量呈指数级增长。文本摘要技术,能让用户从大量数据中高效地获取所需信息。文本摘要技术通过对文本信息概括总结,提炼出文章的主旨。传统的人工提取文本摘要的方式需要耗费大量的时间,自动文本摘要技术顺势诞生。

本文的研究对象是计算机通信类的中文专利说明书。一份完整的中文专利说明书包含扉页、权利要求书、说明书和附图。扉页类似于图书的标题页,是专利说明书的一览表,上面记录了著录项目、附图和专利摘要。权利要求书是申请人要求保护的技术内容。说明书是描述发明创造的技术内容的文件部分。附图是说明书的一部分,用直观、形象的方式解释技术内容。大部分读者都习惯先从扉页中的专利摘要来了解这个专利的功能。摘要是专利说明书内容的概述,因此扉页中的专利摘要对专利来说至关重要。正是由于专利摘要极为重要,因此撰写专利摘要有着非常严格的要求,要求文字简短精炼,且能直观地描述出专利的技术内容,而这往往是让专利申请者头疼的地方。大部分专利申请者在撰写专利摘要时因为摘要内容不合格而需要多次修改,耗费了大量的时间。

针对上述问题,我们提出了一种针对中文文字的基于专利说明书生成文本摘要的方法。该方法以经典的TextRank算法为基础,核心是专利文本中说明书和权利要求书的句子间相似度的构建,利用不同的构建方法来计算句子的权重,去除冗余的句子,并根据句子权重大小来抽取专利文本中的句子生成摘要。实验证明,本文方法相比于已有方法在ROUGE值上有了显著的提高。

本文第2节介绍了文本摘要的相关工作;第3节提出了基于PatentRank的生成专利文本摘要的方法;第4节详细介绍了实验结果;最后总结全文并展望未来。

2　相关工作

文本摘要技术通过自动分析给定的文档,并摘取其中的要点信息,最终输出简短精炼的摘要。摘要的目的是通过对原文进行压缩提炼,为用户提供简明扼要的内容描述。根据不同的划分标准,文档摘要可以分为以下几种不同的类型。

根据处理的文档数量,摘要可以分为单文档摘要和多文档摘要。单文档摘要方法针对单个文档对其内容进行抽取,

基金项目:国家自然科学基金(61806137,61702518)
This work was supported by the National Natural Science Foundation of China (61806137,61702518).
通信作者:王中卿(wangzq@suda.edu.cn)

进而生成摘要。多文档摘要方法是指从包含多份文档的集合中生成一份能概括这些文档中心内容的摘要。本文研究可以被认为是单文档摘要。

根据摘要所采用的方法，可以分为生成式摘要[1-5]和抽取式摘要[6-8]。生成式摘要方法是根据文档的核心思想来重新组织生成摘要，可以重复使用原文档中的短语和词句。生成式方法直接从意义表达生成摘要句子，虽然难度大，但是更接近摘要的本质。抽取式摘要方法通过利用不同方法对文档结构单元进行评价，对每个结构单元赋予一定权重，然后选择最重要的结构单元组成摘要。抽取式方法实现简单，保留完整句子，可读性好。本文研究是一种单文档抽取式摘要。

在传统的抽取式摘要研究中，1958 年 Luhn[9] 提出了一种基于高频词的评分生成文本摘要的方法，由此开启了自动文本摘要技术的研究。Mihalcea 和 Tarau[10] 提出了一种基于图模型的文本处理算法——TextRank 算法。该算法依据句子之间的相似度来构建 TextRank 网络，迭代计算至收敛，由此得到句子的权重。Erkan 等[11] 提出了一种文本摘要算法——LexRank 算法。该算法引入了一种基于随机图的方法，根据句子的特征计算句子的权重，将向量空间模型表示成图模型，通过计算句子之间的相似度提取出相似度大的句子作为摘要。

随着近几年深度学习的发展，研究者们开始尝试将一些最新的研究运用于自动文本摘要技术。目前主流的模型是 seq2seq 模型。文献[12]提供了一种完全由数据驱动的产生生成式摘要的方法。该方法根据输入的句子，利用局部注意力模型来生成摘要中的每个字。受机器翻译中 seq2seq 技术的启发，该文将 Neural Language Mode 和带有上下文的 encoder 结合起来，encoder 和 decoder 在句子摘要任务中共同训练。文献[13]使用一种有条件的 RNN 来生成摘要，条件是卷积注意力模型，用来确保每一步生成的单词都可以聚焦到合适的输入上。文献[14]将自动文摘问题也当作一个 seq2seq 的问题，并且应用 Attentional Encoder-Decoder Recurrent Neural Networks 框架来解决这个问题，融合了很多 features 和 trick，提出了多组对比模型，并且在多种不同类型的数据集上做评测，结果所提模型都获得很好的效果。文献[15]提出把 seq2seq 模型应用于摘要生成时存在两个主要的问题：1)难以准确复述原文的事实细节，无法处理原文中的未登录词(OOV)；2)生成的摘要中存在重复的片段。该研究针对这两个问题，进一步提出融合了 seq2seq 模型和 pointer network 的 pointer-generator network 以及覆盖率机制(coverage mechanism)。文献[16]在基于 attention 的 seq2seq 生成模型基础上提出了 DGRN(Deep Recurrent Generative Decoder)模型，以对训练数据目标摘要中隐含的潜在结构信息进行建模和学习。受传统基于模板方法的启发，文献[17]提出使用现有的摘要作为软模板来指导 seq2seq 模型生成摘要，首先利用信息检索平台从语料中检索合适的摘要作为候选模板，然后扩展标准的 seq2seq 模型，使其具有模板排序和基于模板生成摘要的功能。

3 PatentRank 算法

3.1 模型的整体框架

为了计算专利文本的摘要，我们提出了一种新的 PatentRank 算法。PatentRank 算法是基于 TextRank 构建的，核心是构建中文文本中不同文本句子间的相似度，主要分为 3 部分：1)权利要求书中句子间相似度的构建；2)说明书中句子间相似度的构建；3)权利要求书和说明书中句子间相似度的构建。在本文图模型中，左侧的 S_1，S_2 和 S_3 为权利要求书中的句子，右侧的 S_4，S_5，S_6，S_7，S_8 为说明书中的句子，两个句子中的连线表示它们之间的相似度。图 1 描述了基于图模型的 PatentRank 框架。

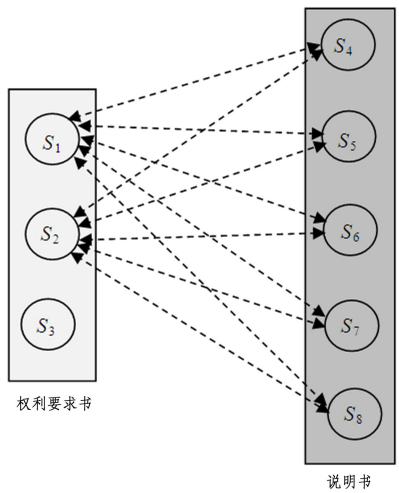

图 1 图模型的 PatentRank 框架

Fig. 1 PatentRank framework based on graph model

3.2 算法

TextRank 算法是基于图的经典排序算法，主要被应用在关键词提取、自动文本摘要领域。该算法作为一种无监督方法，不需要对语料库提前进行学习训练，实现简单并且效果不错。

摘要须具有新颖性，即候选句子包含的冗余信息尽可能少，每句话都可以独立地表示出一种独立的意思。因此要获得高质量的摘要，使生成的摘要尽可能地包含原始文档的信息，就应避免重复出现具有相同信息的句子。因此在开始进行迭代计算之前，使用余弦相似度公式计算句子之间的相似度，去除集合中相似度极高的句子，以提高生成摘要的质量。如果两个句子的相似度大于等于 $\tau=0.953$，则删去其中一句。这些相似度极高的句子并不能总结专利说明书，反而会导致迭代计算的误差。例如，权利要求书中包括前序部分和特征部分，其中特征部分大量使用“其特征是”和“根据权利要求所述的”等冗余句型。

我们可以构建一个有向有权图 $G=(V,E)$，其中 V 是点的集合，E 是边的集合。w_{ji} 是图中任意两点 V_i 和 V_j 之间边的权重，即句子 V_i 和句子 V_j 之间的相似度。$In(V_i)$ 为指向该节点的节点集合，$Out(V_j)$ 为节点 V_j 指向的节点集合。节点 V_i 的评分数值计算公式表示为：

$$ws(V_i)=(1-d)+d*\sum_{V_j\in In(V_i)}\frac{w_{ji}}{\sum_{V_k\in Out(V_j)}w_{jk}}ws(V_j) \quad (1)$$

其中，$d(0\leqslant d\leqslant 1)$ 是阻尼系数，表示从图中某一点指向其他任意点的概率。阻尼系数过大会使迭代次数增加且算法的排序不稳定，阻尼系数过小会导致迭代过程没有明显效果，因此阻尼系数一般取值为 0.85[2]。

本文提出了不同文本中句子间相似度的构建方法。

1）权利要求书间的相似度计算

采用如下公式计算权利要求书中句子间的相似度：

$$w_c = \alpha * Sim(S_i, S_j) = \alpha * \frac{\sum_{i=1}^{n}(A_i \times B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad (2)$$

其中，A_i 和 B_i 为句子 S_i 和 S_j 的向量，n 为句子向量的维度。

2）权利要求书和说明书间的相似度计算

采用如下公式计算权利要求书和说明书中句子间的相似度：

$$w_{cd} = \beta * Sim(S_i, S_j) = \beta * \frac{\sum_{i=1}^{n}(A_i \times B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad (3)$$

其中，A_i 和 B_i 为句子 S_i 和 S_j 的向量，n 为句子向量的维度。

3）说明书间的相似度计算

采用如下公式计算说明书中句子间的相似度：

$$w_d = Sim(S_i, S_j) = \frac{\sum_{i=1}^{n}(A_i \times B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad (4)$$

其中，A_i 和 B_i 为句子 S_i 和 S_j 的向量，n 为句子向量的维度。

在使用 PatentRank 算法时需要注意以下问题：1）初始值的设定，一般让所有节点初始得分为 1.0[10]；2）收敛判定，一般收敛阈值为 1×10^{-4}，即图中任意节点的误差率小于 1×10^{-4} 时达到收敛，停止迭代。研究证明，极限值取为 1×10^{-4} 时，递归计算能够更好地收敛。

当迭代结束，所有句子的权值被确定后，权值高的句子具有较高的信息量，选择这些句子作为构建专利摘要的句子。

本文的算法实现如算法 1 所示。

算法 1 PatentRank

输入：根据标点符号分句后的句子向量

输出：按句子权重大小排序后的句子向量

1. 删除冗余的句子，如果两个句子的相似度大于或等于 τ，则删去其中一句。
2. 根据 3 种不同句子相似度的构建方法构建图 G，即 G 是一个二维向量。
3. 初始化每个句子的权重为 1.0，根据式(1)更新每个句子的权重，直到每个句子权重的误差率小于 1×10^{-4} 时，更新迭代停止。
4. 将句子按权重大小排序。

4 实验

4.1 实验数据及评价标准

本文收集了 100 篇计算机通信类的中文专利说明书，并将每一份专利说明书中的专利名称、摘要、权利要求书和说明书写入文本文件中。利用自然语言处理等技术对文本信息进行处理：首先去除实验噪声，去除对提取摘要无明显影响的图片、表格和特殊符号等；然后将文本正文切分成单个句子，再使用中文结巴分词对句子进行分词，得到每个句子的词集合。

本文的评价工具是 ROUGE-1.5.5，我们把专利说明书中的摘要作为评价标准，通过本文算法生成的摘要作为评价的内容。ROUGE（Recall Oriented Understudy for Gisting Evaluation）评价法由 Lin[18] 基于机器翻译的自动评价方法研发的，用于摘要的自动评价。ROUGE 评价法的主要思想是比较系统生成的文摘与事先给定的人工文摘要间的重叠内容，通过统计两者之间重叠的基本单元的数目来评价文摘的质量。常用的评价指标有 ROUGE-1，ROUGE-2，ROUGE-W 等等。其中，ROUGE-N 中的 N 表示 n 元语法，ROUGE-N 的值是指候选的摘要与参考的摘要集合之间的 n 元语法召回率。

4.2 实验结果与分析

由于专利文本摘要任务比较新，相关工作比较少，因此本文只是和随机抽取的方法进行比较，结果如表 1 所列。

1）Random：在权利要求书和说明书中随机选取句子作为该专利的摘要，由于结果存在随机性，表 1 中的结果是重复 10 次的平均值。

2）TextRank＋权利要求书：实验文本为权利要求书的内容，去除了冗余的句子，基于经典的 TextRank 算法生成摘要。

3）TextRank＋说明书：实验文本为说明书中的内容，去除了冗余的句子，基于经典的 TextRank 算法生成摘要。

4）TextRank＋权利要求书＋说明书：实验文本为权利要求书和说明书中的句子，去除了冗余的句子，基于经典的 TextRank 算法生成摘要。

5）PatentRank＋权利要求书＋说明书：实验文本为权利要求书和说明书中的句子，去除了冗余的句子，基于 3 种不同相似度的构建，选取出权重高的句子生成摘要。

表 1 5 种方法的实验结果

Table 1 Experimental results of five methods

Method	*ROUGE*-1
Random	0.42703
TextRank＋权利说明书	0.45537
TextRank＋说明书	0.48582
TextRank＋权利要求书＋说明书	0.46148
PatentRank＋权利要求书＋说明书	**0.50402**

从实验结果来看，Random 方法的实验结果并不理想。Random 方法没有考虑到句子之间的相关性，忽略了文本的重要信息。TextRank＋权利要求书方法从权利要求书中抽取句子，考虑了句子之间的相似性。TextRank＋说明书方法从说明书抽取句子，也考虑了句子间的相似性，从表 1 中可以看到它的实验结果显然比 TextRank＋权利说明书方法好，由此可以证明说明书中具有更多的重要信息，有更多表达主旨中心的句子。同时，TextRank＋权利要求书＋说明书方法的效果比 TextRank＋说明书的方法效果差，没有达到实验预期，因为权利要求书也会含有重要的句子，对构建专利摘要也起着很重要的作用。由此我们提出了 PatentRank 算法，通过构建 3 种不同的句子相似度的方法来尝试改善实验效果。

先设定式(3)中的 α 为 1.0，尝试不同的 β 值来比较实验效果，即 *ROUGE* 值。图 2 中展示了实验结果，横坐标是 β 值，纵坐标是 *ROUGE* 值。由图 2 可知，当 β 的值为 0.4 时，*ROUGE* 值最高，为 0.50087。

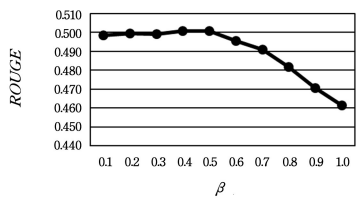

图 2 不同 β 值的实验结果

Fig. 2 Experimental results of different β

确定了 β 值后，继续尝试不同的 α 值是否有更好的实验效果。将 β 值设为 0.4，尝试不同的 α 值来比较实验结果。图 3 中展示了实验结果，横坐标是 α 值，纵坐标是 *ROUGE* 值。由图 3 可知，当 α 的值为 0.7 时，*ROUGE* 值最高，为0.504 02。

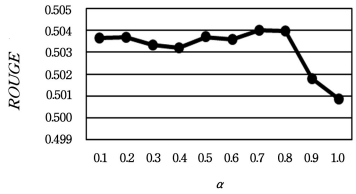

图 3 不同 α 值的实验结果

Fig. 3 Experimental results of different α

通过上述实验可知，PatentRank 算法针对说明书和权利说明书不同文本的特点，设置了 3 种不同文本句子间的相似度函数，当 α 的值为 0.7，β 的值为 0.4 时，实验效果最佳。

结束语 本文是结合权利要求书和专利说明书的中文专利摘要的研究，提出了一种新的基于 PatentRank 算法生成中文专利摘要的方法，可应用于中文计算机通信类的专利文本摘要任务。实验结果表明该方法有一定的效果，能够得到具有一定总结性的摘要。在以后的工作中，我们会收集其他领域的专利文本用于测试算法的性能，同时也会尝试 PatentRank 算法与文本整体信息结构等的结合，以提高特征句子、核心句子的权重，从而生成文本的摘要句子，使其效果进一步提高。

参 考 文 献

[1] WANG L, YAO J, TAO Y, et al. A reinforced topic-aware convolutional sequence-to-sequence model for abstractive text summarization[J]. arXiv: 1805.03616, 2018.

[2] LIN J, SUN X, MA S, et al. Global Encoding for Abstractive Summarization[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers). Melbourne, Australia, 2018: 163-169.

[3] PAULUS R, XIONG C, SOCHER R. A deep reinforced model for abstractive summarization[J]. arXiv: 1705.04304, 2017.

[4] CHEN Y C, BANSAL M. Fast abstractive summarization with reinforce-selected sentence rewriting[J]. arXiv: 1805.11080, 2018.

[5] LIU L, LU Y, YANG M, et al. Generative adversarial network for abstractive text summarization[C]// Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

[6] YASUNAGA M, ZHANG R, MEELU K, et al. Graph-based neural multi-document summarization[J]. arXiv: 1706.06681, 2017.

[7] NARAYAN S, COHEN S B, LAPATA M. Ranking sentences for extractive summarization with reinforcement learning[J]. arXiv: 1802.08636, 2018.

[8] ZHOU Q, YANG N, WEI F, et al. Neural document summarization by jointly learning to score and select sentences[J]. arXiv: 1807.02305, 2018.

[9] LUHN H P. The automatic creation of literature abstracts[J]. IBM Journal of Research and Development, 1958, 2(2): 159-165.

[10] MIHALCEA R, TARAU P. Textrank: Bringing order into text[C]// Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. 2004.

[11] ERKAN G, RADEV D R. Lexrank: Graph-based lexical centrality as salience in text summarization[J]. Journal of Artificial Intelligence Research, 2004, 22: 457-479.

[12] RUSH A M, CHOPRA S, WESTON J. A neural attention model for abstractive sentence summarization[J]. arXiv: 1509.00685, 2015.

[13] CHOPRA S, AULI M, RUSH A M. Abstractive sentence summarization with attentive recurrent neural networks[C]// Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016: 93-98.

[14] NALLAPATI R, ZHOU B, GULCEHRE C, et al. Abstractive text summarization using sequence-to-sequence rnns and beyond[J]. arXiv: 1602.06023, 2016.

[15] SEE A, LIU P J, MANNING C D. Get to the point: Summarization with pointer-generator networks[J]. arXiv: 1704.04368, 2017.

[16] LI P, LAM W, BING L, et al. Deep recurrent generative decoder for abstractive text summarization[J]. arXiv: 1708.00625, 2017.

[17] CAO Z, LI W, LI S, et al. Retrieve, rerank and rewrite: Softtemplate based neural summarization[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018: 152-161.

[18] LIN C Y, HOVY E. Automatic evaluation of summaries using n-gram co-occurrence statistics[C]// Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. 2003.

SHU Yun-feng, born in 1998, undergraduate. His main research interest is natural language processing.

WANG Zhong-qing, born in 1987, Ph.D, lecturer, is a member of China Computer Federation. His main research interest is natural language processing.