****



**研 究 生 毕 业 论 文**

**（申请硕士学位）**

**论文题目 基于语义重构的文本摘要算法**

**作者姓名** **张弛**

**学科、专业名称**  **计算机科学与技术系**

**研 究 方 向 数据挖掘**

**指导教师** **王崇骏 教授、吴骏 讲师**

**2016 年 5 月 27 日**

**学号：MG1333075**

**论文答辩日期：2016 年 5 月 27日**

**指 导 教 师： （签字）**

|  |  |  |
| --- | --- | --- |
| **NJU2** | | |
| 南京大学申请硕士学位论文 | | |
| 基于语义重构的文本摘要算法 | | |
|  | | |
| 作 者： | **张弛** | |
| 专 业： | **计算机科学与技术** | |
| 研究方向： | **数据挖掘** | |
| 指导教师： | **王崇骏 教授、吴骏 讲师** | |
|  | | |
| **南京大学计算机科学与技术系**  **2016年5月** | | |
| **NJU2** | | |
|  | | |
| Text Summarization Based on Semantic Reconstruction | | |
|  | | |
| Presented By  **Zhang Chi** | |
| Supervised by  **Prof. Wang Chong-Jun**  **Wu jun** | | |
| A DISSERTATION  FOR THE APPLICATION OF MASTER DEGREE  SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE  AND TECHONOLOGY OF NANJING UNIVERSITY | | |
| **May 2016** | | |

声明

本人声明所呈交的论文是我个人在导师指导下、在南京大学及导师提供的研究环境（含标明的项目资助）下作为导师领导的项目组项目整体的组成部分而完成的研究工作及取得的研究成果。

除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

南京大学及导师所有权保留：送交论文的复印件，允许论文被查阅和借阅；公布论文的全部或部分内容；可以采用影印、缩印或其它复制手段保存该论文。

学生签名： 日期：

导师签名： 日期：

Declaration

I make a declaration here that the thesis submitted is composed of the researching work by myself and its corresponding researching results finished as a constituent part of the whole project in the project team lead by my advisor. The thesis is completed with the guidance of my advisor, and under the researching circumstances offered by Nanjing University and my advisor (including the project support indicated).

The thesis does not include other people’s researching results ever published or composed, except that are specially annotated and acknowledged somewhere in the article. Any contribution made to the research by my working partners is declared explicitly and acknowledged in the thesis.

Nanjing University and the advisor retain the copyright as follows: submitting the copies of the thesis, allowing the thesis to be consulted and borrowed; publicizing the whole or part of the thesis’ content; keeping the thesis by photocopy, microcopy or other copy methods.

Author Signature： Date：

Advisor Signature： Date：

摘要

互联网技术的快速发展产生了数据爆炸和信息过载的问题，同时现代生活节奏的加快催生了用户快速阅读的需求，使得文本自动摘要技术成为了当今科学界的研究热点。相比其他自然语言处理任务，自动摘要技术的挑战在于摘要的评价指标无法精准量化，极具主观性，而且自动摘要往往深受冗余信息的困扰。目前主流的自动摘要算法是通过预先定义某个指标，对所有句子进行打分，然后对句子排序并抽取top-k作为生成摘要。然而这些抽取排序模型一方面对句子独立打分，孤立了句子之间的联系，忽略了文章的结构信息；一方面选取的评分指标通常是词素级别或者统计特征，缺乏语义信息。针对这些缺点，我们设想一个高质量的摘要能够很好地还原原文的语义，进而提出了语义重构模型：通过寻找能够以最小损失重构原文语义的句子集作为最后的生成摘要。

本文的工作主要包括两个方面：

(1)针对词袋模型的高维稀疏、缺乏语义信息的现象，设计了两种简单有效的语义向量化方式表示文本，分别是基于神经语言模型的词嵌入加权方法和基于多层自编码网络的深度降维方法。并通过句子分类实验证明了这两种向量化方式都能得到紧凑且具有语义的文本表示。

(2)分别设计了基于二次规划的线性重构策略和更为平滑灵活的非线性重构策略，以得到能最佳还原原文的句子并作为结果摘要。另外通过冗余消减手段在改进了重构策略并提高了摘要质量。最后在DUC标准数据集上的摘要实验对比，证明了本文的语义重构模型的合理性和有效性。

**关键词**： 自动摘要 语义重构 词嵌入 语义表示

Abstract

With the rapid development of web technology, here comes the problem of data explosion and information overload. Therefore the technology of automatic text summarization becomes the hotspot in computer science. In contrast with other NLP tasks, the challenges that automatic summarization face with are that the judge issue of summary is too subjective and there always lots of redundancy lying in the result summary. Most existing models score sentence by predefining some features and select the top-k sentences as result summary. However these ranking models score each sentence independently without considering the relationships between sentences. On the other hand, these predefined features usually are lexical or statistical, which cannot capture the semantic meanings of text. To counter these shortcomings, we assume that a good summary can reconstruct the original document, and we propose the semantic reconstruction model basing on this assumption.

The proposed model selects the sentences that can best reconstruct the original document as the result summary. Our work in this paper consists of two parts:

1. Semantic representations of sentence. Given that the bag-of-words vector can not capture the semantic meanings, we use two approaches to learn compact and semantic representations for sentence:(1) weighted mean of word embeddings; (2) deep coding. The semantic representations can be used as the input of reconstruction model.

2. Reconstruction strategy is the key of semantic reconstruction and aims to find the most relevant sentences. The reconstruction strategy in this paper includes a simple linear function and flexible nonlinear function, respectively basing on quadratic programming and neural network. Besides, redundant sentences can be reduced by redundancy reduction algorithm to improve the summary quality. And the summary experiments basing on the DUC datasets validate the effectiveness of our model.

**Keywords**: automatic summarization, semantic reconstruction, word embedding, semantic representation

目录

[第一章 绪论 1](#_Toc451969756)

[1.1 自动摘要的研究背景 1](#_Toc451969757)

[1.2 自动摘要的分类 2](#_Toc451969758)

[1.3 自动摘要的挑战 3](#_Toc451969759)

[1.4 本文工作和组织结构 5](#_Toc451969760)

[第二章 文本摘要的相关研究 7](#_Toc451969761)

[2.1 句子排序抽取法 7](#_Toc451969762)

[2.1.1 基于统计信息 7](#_Toc451969763)

[2.1.2 句子聚类和图模型 9](#_Toc451969764)

[2.1.3 机器学习 10](#_Toc451969765)

[2.2 基于语言学方法 11](#_Toc451969766)

[2.2.1 词汇链 11](#_Toc451969767)

[2.2.2 LSA 12](#_Toc451969768)

[2.2.3 互参信息和修辞结构 12](#_Toc451969769)

[2.3 特殊文体或领域的摘要方法 13](#_Toc451969770)

[2.3.1 医学摘要 13](#_Toc451969771)

[2.3.2 期刊摘要 14](#_Toc451969772)

[2.3.3 邮件摘要 14](#_Toc451969773)

[2.3.4 网页摘要 15](#_Toc451969774)

[第三章 文本的语义表示 17](#_Toc451969775)

[3.1 引言 17](#_Toc451969776)

[3.2 词嵌入加权 18](#_Toc451969777)

[3.3 深度降维 21](#_Toc451969778)

[3.4 实验对比 25](#_Toc451969779)

[3.5 本章小结 27](#_Toc451969780)

[第四章 原文语义重构策略 28](#_Toc451969781)

[4.1 线性重构策略 28](#_Toc451969782)

[4.1.1 目标函数 29](#_Toc451969783)

[4.1.2 优化方法 33](#_Toc451969784)

[4.2 非线性重构策略 34](#_Toc451969785)

[4.2.1 模型结构及训练 35](#_Toc451969786)

[4.2.2 摘要提取 40](#_Toc451969787)

[4.3 冗余消减 41](#_Toc451969788)

[4.4 实验对比 43](#_Toc451969789)

[4.4.1 数据集和评测工具 43](#_Toc451969790)

[4.4.2 对比实验介绍 45](#_Toc451969791)

[4.4.3 实验结果及分析 45](#_Toc451969792)

[4.5 本章小结 49](#_Toc451969793)

[第五章 总结与展望 50](#_Toc451969794)

[5.1 工作总结 50](#_Toc451969795)

[5.2 未来展望 51](#_Toc451969796)

[参考文献 52](#_Toc451969797)

[致谢 57](#_Toc451969798)

[附录 58](#_Toc451969799)

2. 绪论

随着互联网技术的发展和产生数据的快速膨胀，信息过载(Information Overload)的问题变得日益严重；同时现代社会的生活节奏变得越来越快，让更多的用户提出了快速获取知识的需求，因此自动文摘(Automatic Summarization)技术便应运而生。文摘又可以称为文本摘要(Text Summarization)或者文档摘要(Document Summarization)。自动文摘技术是通过电脑程序从文本集合中自动提炼出简洁连贯的语言段落，并保留原文的主旨思想，以达到信息浓缩的目的。

移动通信的发展让更多用户选择从手机端获取资讯，然而由于博客、新闻等文章篇幅往往很长，手机屏幕尺寸严重制约了阅读效率。2013年雅虎以3000万美元收购了17岁德国学生Nick D’Alosio基于IOS平台开发的新闻自动摘要应用Summly，以利用其适配移动设备体验的自动摘要技术来加强公司内部产品，如雅虎财经和雅虎体育。从这个备受关注的事件可以看出，自动摘要技术是个蕴含商机的前沿研究热点。

关键词提取(Keyword Extraction)是一种和文本摘要思路类似的NLP任务，目的是从文档中找到一组和原文话题最为相关的单词或短语。关键词提取任务可以通过文本挖掘(Text Mining)和信息检索(Information Retrieval)技术解决，比如文档主题模型(Topic Model)[1]或者随机游走模型[2]。文本摘要和关键词提取任务形式上都类似特征选择(Feature Selection)问题，前者找特征句，后者找特征词(短语)。然而文本摘要任务的难度更高，因为相比词语，句子的文本粒度更大、信息更丰富，不能用单个特征表示，因此不能用简单的特征选择算法如卡方检验直接求得；此外，文本摘要还要考虑摘要冗余问题，句子之间的冗余信息会使摘要质量很差。但是，关键词提取往往可以作为文本摘要的早期步骤，进而用关键词的包含情况来衡量句子重要性。

* 1. 自动摘要的研究背景

随着PC和手机的普及，以及互联网技术和移动通信技术的迅猛发展，人类社会进入了大信息时代。据中国互联网发展统计报告称，截至2010年我国网民数量近4.2亿，手机用户达2.77亿，网络普及率不断提高。同时，以Facebook、Twitter为代表的社交网络(Social Networking Services, SNS)风靡全球，让越来越多的网民依赖于从网络媒体获取或发布信息，使得互联网已经成为最大的信息集散地。随之出现的现象就是信息过载，或称为信息爆炸，据worldwidewebsize[[1]](#footnote-1)统计世界最大的搜索引擎公司Google所存储索引的网页数量超过了450亿，每天处理近100PB的数据；根据数字宇宙研究报告称，未来8年里人类将产生超过40ZB(=40万亿GB)的数据量。海量的数据再加上远快于人类消化的更新速度，一方面需要谷歌等搜索引擎公司利用庞大的计算机集群建立索引，以提供个性化检索服务；另一方面需要高效的自动摘要技术来对大规模文本进行关键信息提取，以提供快速浏览功能。

自1958年Luhn[3]在IBM Journal上发表的第一篇自动文摘领域的论文已近六十年，期间自动摘要技术得到了广泛研究和长足发展。然而自动摘要技术在处理速度和摘要质量上依然有诸多不足，包括：(1)处理速度太慢，赶不上数据产生和信息传播的速度。(2)摘要质量在组织性、可读性等方面相比人工摘要相去甚远。因此，工业界对成熟高效的自动摘要技术有急迫的需求。此外在一些特定领域，自动摘要技术的研究严重不足。比如医学摘要和基因摘要(Gene Summarization)，对摘要质量的专业性和准确性要求极高。因此，文本摘要技术依然是NLP领域的热点分支。

* 1. 自动摘要的分类

所有文本摘要的思路都是对文档进行压缩以提炼出能反映原文主旨的简短段落。但从不同角度不同需求出发，文本摘要可以有多种分类方式。

总体上，摘要的类型可以分为抽取式文摘(Extractive Summary)和生成式文摘(Abstractive Summary)。抽取式摘要从原文最具有代表性的句子或段落而并不改变原句，直接组成文摘；相反的，生成式摘要通过对原文进行语义理解并分析主旨，然后通过语言模型来重新组织语言来形成文摘。简而言之，两者区别是前者直接选择现存的句子；后者通过NLP技术生成新句。高质量的生成式摘要能够贴近人类给出的摘要，更具有可读性。然而受限于实现难度，绝大部分的摘要研究包括本文提出的模型，都是基于抽取式摘要。

从处理的文本数量的角度，自动摘要系统又可以分为单文本摘要(Single-document Summary)和多文本摘要(Multi-document Summary)。多文本摘要每次从若干篇相关的文档中生成一份摘要，这些文档通常围绕同一个主题各有侧重又相互联系，构成一个主题文档群(Topic Group Documents)。多文档摘要的难度要远远大于单文本摘要，因为多文档文摘通常存在非常严重的冗余性问题，主题文档群里的所有文档都围绕同一主题，因此容易提取到大量重复信息；此外每个文档各自的侧重点也很难把握。

从用户需求角度，摘要可以分为普通摘要(Generic Summary)和基于查询摘要(Query-based Summary)。普通摘要就是不考虑任何用户需求，即生成的摘要能让任何人看懂原文的大概含义。而查询相关摘要是面向用户的，针对用户提出的特定需求去生成和其查询相关的摘要。输入的用户查询可以是句子、词组，甚至话题，因此查询相关摘要任务相比之下更容易一些，生成摘要的质量也通常更高。一个理想的自动摘要系统应该既能处理普通摘要，也能面向特定用户生成查询相关摘要，而且当给出查询信息越丰富时，生成的摘要质量更高，如TextRank[4]以及最大边界相关模型(Maximal Marginal Relevance, MMR)[5]。

摘要从内容可以分为指示性摘要(Indicative Summary)和信息型摘要(Informative Summary)。指示性摘要能够给出诸如长度、写作风格等文章信息，提供快速浏览功以便用户大概了解文章结构相关，进而深入原文进行选择性阅读。而信息性摘要就是原文提炼出来的短文，可以直接代替原文而不影响阅读体验。

此外，还有一些特殊类别的摘要，如更新式摘要(Update Summary)会根据用户的历史阅读记录来对摘要进行更新；而头条式摘要(Headline Summary)只给出一个句子作为摘要；辅助式摘要(Aided Summary)依赖人工对自动摘要进行后处理；基因摘要是对基因相关信息如基因产物、突变表型和基因集合等生成摘要。

* 1. 自动摘要的挑战

自动文摘技术已经发展了近六十年，期间得到了广泛研究和长足进步。然而机器生成的摘要依然差强人意，无法完全代替人工摘要以满足人们快速、准确获取信息的需求。相比于分词(Word Segmentation)、命名实体识别(Named Entity Recognition)等早已在工业界成熟应用的NLP任务，自动文本摘要依然步履蹒跚，面临不少困难和挑战。

自动文本摘要技术目前遇到的挑战主要存在以下几个方面：

(1)摘要本身没有一个准确量化的定义。摘要算法的设计思路往往是通过预先定义一个评价指标，然后提出相应的优化方法来得到指标得分最高的句子集合作为最后生成的摘要。同样的流程在社团发现(Community Detection)中也有体现，社团在概念上是在网络中的联系紧密的节点簇，并没有客观量化的定义，可以通过侧面定义模块度、中心度等指标衡量节点簇的结构强度以寻找社团。但是社团结构可以通过在空间中的物理分布情况来评价，而相比之下摘要的评价是个主观复杂的过程，难度更高。

(2)文本摘要的评价指标过于复杂。无论是人工摘要还是机器摘要，凸显性、覆盖度、切题性、冗余度、组织性、流畅性等都是要考虑的方面。其中较为重要的几个评价指标如下：

凸显性(Salience)：又称切题性，高质量的摘要必须契合文章的主旨，每个句子都和原文高度相关，不重要的信息显然要丢弃；

覆盖度(Coverage)：在凸显原文主题的同时，要涵盖尽可能多的话题点，不能挂一漏万、舍本逐末；

冗余度(Redundancy)：冗余性是自动摘要最容易出现的问题，即句子之间存在大量的重复信息(Overlapping Information)。冗余性问题在多文本自动摘要中尤其突出，因为主题文档群都是围绕着同一主题，部分内容高度相似；

组织性(Fluency)：指生成摘要的句子顺序能够读起来逻辑顺畅、条理分明。对于自动摘要来说可以通过人工后处理实现;

可读性(Readability)：一般是人工摘要以及生成式自动摘要的目标。生成式摘要往往依赖一个优秀的语言模型，让生成的新句语法正确、结构自然。

(3) 文本摘要的评价过于主观。摘要评价的主观性体现在，对文章的主题理解本身就是个主观的过程，因此人工摘要以及自动摘要的评价也极具主观性。

所谓“一千个读者心中就有一千个哈姆雷特”，对同一篇文章的主题，不同的人基于不同的立场有不同的理解。这种现象让文本摘要的质量好坏成为非常主观的评判。因此大多数NLP任务尤其是自动摘要任务所用的标准数据集都使用相对客观的热点新闻材料，因为时事新闻通常具有主题突出、立场鲜明的特点。尽管如此，当涉及到多文本摘要时，文档主题群里的多个文档都围绕同一主题引申出各有侧重的话题。对这些话题的取舍，是先主后次还是面面俱到，即使对语言专家来说也是难以定度。

文章主题的主观性，直接导致了摘要评价的主观；再加上摘要的组织性、可读性等指标无法客观量化地评判摘要质量，导致即使是人工生成的摘要，其质量评价也是仁者见仁智者见智。而机器摘要的评价往往又依赖人工摘要，比如最权威的自动摘要评价指标Rouge就是将结果摘要和专家给出的参考摘要进行相似度对比。相反的，对其他诸如词性标注(Part-of-speech Tagging)、分词等简单的NLP任务来说，所得词性或者分词结果是确定的、客观的，直接对比模型输出和预期结果就可以对模型表现进行评估。因此，缺乏客观、精准的评价指标和预期结果，也是自动摘要面临的重要挑战。

(4)自动摘要模型的不足。目前主流的自动摘要模型是基于某种特征给输入文档的句子打分，然后对所有句子按分值排序并选取top-k作为结果摘要。这种排序模型存在两个不足：一方面所选取的特征往往是词素级或统计性的简单特征，不能把握句子的语义信息，即无法分辨语义相似但表述不同的句子；另一方面给句子独立打分的过程并没有考虑句子之间的联系，忽略了原文的结构信息，且容易导致冗余。

本文为了解决上述不足，一方面通过模拟人脑思考的过程将文本语义向量化，提取句子的深度表示，把握句子的语义信息；另一方面，通过基于原文重构来整体选取句子，避免对原文结构信息的忽略。

* 1. 本文工作和组织结构

为了解决主流自动摘要模型中忽略句子语义信息和句子之间联系，以及词素特征或统计特征缺乏语义性的缺点，本文提出了基于语义重构的自动摘要算法。算法分为两个步骤，一方面是通过简单有效的文本向量化模型将句子和原文表示成紧凑的语义向量；另一方面，基于原文重构，找到能最有效还原文本语义主题的句子集，作为生成摘要。

本文后续组织结构如下：

第二章介绍当今文本自动摘要领域的相关研究和基本方法。

第三章设计词嵌入加权和深度降维两种简单有效的语义表示方式对文本进行语义向量化，生成重构模型的输入表示。

第四章分别介绍了线性重构策略和非线性重构策略，各自通过原文重构来得到高质量的摘要。此外通过冗余消减算法，提高摘要表现。最后辅以实验论证。

第五章是总结与展望,包括对论文系统的主要工作进行总结并针对本文工作的可完善之处进行展望和探讨。

1. 文本摘要的相关研究

文本摘要发展了近六十年，涌现出各种经典的算法和模型。最主流的自动摘要模型是句子排序抽取模型，其衡量句子重要性的方法，由最初的词频统计、对数似然比等统计信息，发展到图模型方法，再到各种有监督的机器学习方法。随着语言模型、知识图谱等技术的发展，基于语言学的各种方法也开始流行，包括词汇链和潜在语义分析(Latent Semantic Analysis, LSA)等。

以上方法都是面向普通摘要和基于查询的摘要，对于特殊文体或领域的文章摘有其他需求，如医学摘要必须考虑严谨性、期刊摘要可以利用科学论文的固有结构等。

* 1. 句子排序抽取法

句子的排序抽取是最主流的自动摘要方法，其重点是如何衡量句子的重要性，其中包括非监督的统计信息法和图模型法，以及监督的机器学习方法。

* + 1. 基于统计信息

Luhn在自动摘要领域的开河之作[3]，将词素频率作为衡量重要性的标志。Luhn假定词频能够反映词概率(Word Probability)，即原文的词在生成摘要中出现的概率。词概率也是最简单的衡量词重要性的标志，计算公式为该词出现频率除以输入文档中的所有词个数*。*给定输入文档的词概率分布，摘要的似然概率可以通过多项式分布计算出：

|  |  |
| --- | --- |
|  | (2-1) |

其中M是摘要的字数，是词在摘要中出现的次数。Nenkova[6]分析了30个DUC主题文档群，对比了四个语言学专家给出的参考摘要和机器生成的自动摘要，发现人工摘要的似然概率要略高一些，证明了词频在衡量重要性上的价值。

但正如Luhn自己指出，很多高频词和其重要性并不相称。因为所有文档中的词语总体出现情况服从齐普夫分布(Zipfian Distribution)[7]：单词出现的频次与它在词频表中的排名的常数次幂成反比。而在信息检索领域非常流行的tf\*idf权重，正好能够解决这个问题[8][9]。

|  |  |
| --- | --- |
|  | (2-2) |

*D*表示所有的文档，表示词*w*出现过的文档数。被称为逆文档频率(Inverse Document Frequency, IDF)，表明了该词在背景文档里的先验概率，其值越高说明该词越与某特定话题高度相关，而idf值很低的词通常为停用词(Stop Words)。

对数似然比(Log-likelihood Ratio)[10]是比tf\*idf更有效的词重要性衡量指标，可以很好的找到能够高度概括原文的词，这些词在文献中普遍称为主题标志(Topic Signature)[11]。和tf\*idf类似，topic signatures也和话题高度相关，即在其他话题的文档中极少出现。但对数似然比是通过设置阈值来判断输入文档中的词是否是主题标志。词*w*在所有背景文档*D*以及和输入文档*I*中分别出现的概率比较分为两种情况：

|  |  |
| --- | --- |
|  | (2-3) |
|  |

输入文档和背景文档都可以看作词序列，每个词的出现都符合伯努利实验，因此文档的似然值可以通过二项分布公式计算：

|  |  |
| --- | --- |
|  | (2-4) |

其中*N*为文档长度，*k*为*w*出现次数。对数似然比定义为：

|  |  |
| --- | --- |
|  | (2-5) |

其中和分别从输入文档和所有背景文档计算得到。-2的数据分布在统计学被称为分布，可以用来区分词是否为主题标志，主题标志就是似然概率比随机预期值更大的词。可以通过查询分别表来获得某个值所对应的概率，比如对数似然比为10.83的概率为0.001。句子的重要性就可以通过包含的主题标志数量衡量，而不需要对词进行加权。Lin和Hovy[11]设计的SUMMARIST[12]系统就是利用主题标志来进行自动文本摘要。而之后[13][14][15]使用的摘要算法也分别利用主题标志来进行多文档摘要。主题标志相比词频和tf\*idf表现更好，因为它是根据实际分布给出判断词语是否具有主题代表性的界限[16]。

* + 1. 句子聚类和图模型

在多文本新闻摘要中，由于主题文档群里的多篇新闻都是围绕着同一主题的，因此重复出现的相似信息可以认为是主题相关的重要信息。不少学者[17][18][19]就从句子聚类着手，来寻找主题句。一种简单可行的思路是先对所有句子进行聚类，然后从每个类簇里挑选一个代表性句子组成摘要，同时尽可能的消除冗余信息。句子聚类模型的一个显著缺点是每个句子只能严格划分到一个类簇中，但实际上有些句子会阐述多个观点。因此，表现方式更加灵活的图模型开始流行起来。

图模型可以兼在词和句子的层面剖析其重复性或重要性。句子之间的相似度往往通过重叠词数计算，因此高频词可以联接多个句子，而且相似的句子可以同时提高彼此的权重。因此，基于图的方法同时拥有词频统计模型和句子聚类模型的优点。此外，图模型可以直接计算句子的重要性值，而不是对词重要性的加权来给句子打分。典型的图模型中[2][4]，点代表句子，而边代表句子之间的联系，边权值为相似度。点的值又称为中心度(centrality)，中心度的概念从社会网络启发而来，将所有句子组成的图看成一个网络，句子的中心度就是其和原文中心思想的相关性。中心度可以通过普适的图算法解决，比如PageRank：当所有的边权值都归一化成概率分布后，从每个点出发的边权值之和为1，整个图进而变成了马尔科夫链(Markov Chain)，而边概率值构成了转移矩阵。随机过程（Stochastic Process）算法可以计算出任意*t*时刻的点矩阵值，直到收敛到平稳分布(Stationary Distribution)。最终的收敛值越高的点，其对应的句子重要性越高，越可能选择为摘要句子。

DUC04官方曾经对比话题标志算法[14]以及图模型算法的摘要结果，发现两种模型的表现都很出色，其中话题标志算法的表现略胜一筹。而图模型在通用性方面有无可比拟的优势，无论是单文本还是多文本摘要都表现良好，并且图模型无需语言处理，无论中文摘要还是英文摘要均可适用[20]。在多文档摘要中，可以通过在同文档句子以及不同文档的句子之间使用不同的权重指标给联接边赋值，来实现区分单个文档的主旨以及多个文档反复出现的共同话题[21]。

* + 1. 机器学习

越来越多的句子重要性的指示特征被提出后，机器学习方法能够有效的结合这些不同特征。Kupiec[22]利用了朴素贝叶斯分类器对摘要进行预测。他使用了五个特征，包括：Paice[23]提出的固定短语特征、Luhn[3]提出的位置相关的段落特征、Edmundson提出的词频特征以及Kupiec本人新提出的大小写特征和句子长度特征，并假设这些特征彼此之间独立，样本文档为摘自工程信息（Engineering Information）的188篇科技文章，每篇文章的类标为由语言专家给出的参考摘要。经过实验证明，多特征结合的方法显著提高了摘要效果，但其中词频这一特征反而降低了摘要质量。

之后，Conroy和O’Leary[25]利用隐马尔科夫模型(Hidden Markov Model, HMM)来提取文本摘要。相比贝叶斯分类器，该模型不需要特征之间彼此独立这一前提假设。他们的模型使用了三个特征：句子在文中的位置、每个句子中的词个数和输入文档的词概率。此外，他们还探索了马尔科夫依赖性，即某句子出现在摘要中的概率依赖于它的前句是否在摘要中。在后续的HMM模型中，句子的主题标志个数也被作为特征。

在普通摘要中，基于机器学习的摘要算法的表现并没有比图模型或者词频的非监督模型有太大的提高。但是在一些与结构或文体相关的摘要中，机器学习算法的表现远远胜出，因为在这些任务中，分类器可以用来区分特定的信息种类，比如科普论文中的文学背景的句子摘要，以及会议记录中是否达成一致协议。

这些机器学习算法中的一个固有问题是训练文本必须带有摘要类标，而人工对文本进行摘要标注非常耗时耗力[26]。因此一些生成式摘要的研究[27][28][29][30][31]致力于将摘要和原文自动对应，即能够分辨出输入文档中的摘要句子和非摘要句子。由于不同摘要者的摘要方式不一样，因此很难直接从单篇摘要去辨别摘要度信息(Summary-worthy Information)。为了解决这个问题，Chali[32]提出了通过相似度计算从原文和参考摘要找出高度相似的句子对，并非完全一一对应。

总之，在抽取式摘要算法中，监督学习并没有显著提高普通摘要的表现。仅利用单个特征例如话题词或者图模型的中心度就能得到较好结果。然而至今并没有很好的模型能结合这些特征而不使用监督模型进行自动摘要。

* 1. 基于语言学方法

之前的方法主要是基于统计信息来对句子排序抽取。然而通过对文本的语义理解和词句结构的分析，更能发掘原文的主旨性和话题性。有些模型依赖于人工建立的语义资料库如WordNet，如词汇链方法；而有些模型从大量未注释文本推断语义信息，如LSA。

* + 1. 词汇链

词汇链[33][34][35]试图呈现原文或段落的主旨。词汇链的原理是话题通常由若干相关的词语联系起来而表达的，如 “汽车”，“轮子”，“公路”等暗示同一个话题的词序列，尽管这些词可能无法顺畅的组织成句子。词汇链的生成极度依赖于WordNet[36],一个由专家撰写的、将海量词语分门别类地梳理过后的词汇网络，网络中联接的词都是具有话题或者类别相关性。此外，词汇链模型需要一定程度的语言学处理，比如词性标注以及话题相关的分词处理。

Barzilay和Elhadad的摘要系统[33]首先将输入文件进行分词，通过WordNet建立词汇链，然后给词汇链打分，然后从每条高分链中选择一个句子。他们的创新点在于该模型创建词汇链的方式更合理，能够很好的处理一词多义现象，例如" bank"一词可有“银行”和“河边”等多种意思。Barzilay的算法里将文本中所有可能的词汇链都建立之后才对多义词进行消歧处理，消歧方式为该词的实际词意选择在链中有最多联接的语境。之后的研究进一步改进了建立词汇链的时间复杂度和消歧准确率[37][35]。

词汇链比词频等特征更能反映句子的重要性，因为很多不同的词可以指向同一话题。词汇链的特征包括长度，即链中的词个数，以及均匀度(homogeneity)，即链中不同的词的个数除以长度。通过词汇链选择句子的具体方式为每个链选出代表词，然后从文本中选择第一个包含该词的句子即选为摘要句。

由WordNet推导的词汇链后来演变成概念集合(Concept Set)，可以更好的避免一词多义的问题。在多文本摘要系统DEMS[38]，用WordNet的同义词、上位词以及下位词关系推导概念集合。不同于词汇链只是将语义相关的词找出并相联接，DEMS系统将有五种以上释义的词都剔除。概念集合建立之后，对每个集合的所有词频相加作为该概念的频度。比如集合C={bank, river, fish, bait}，即使每个单词的频次可能很低，但是整个概念可能经常出现。

然而词汇链以及概念集合都强烈依赖WordNet，导致生成摘要的表现严重受制于WordNet的覆盖及标注效果。因此不需要人工参考资料的LSA是一种很好的选择。

* + 1. LSA

D.Scott提出的LSA[39]是一种通过观察词的共现情况来学习文本语义表示的非监督模型。Gong和Liu[40]提出使用LSA来进行单文本和多文本的普通新闻摘要，算法思路是先将输入文档表示成一个词-句矩阵*A*，行表示文档的词，列表示句。*A*的每个项表示词*i*在句子*j*中的tf\*idf值。然后对矩阵*A*进行奇异值分解(Singular Value Decomposition, SVD)分解成三个矩阵的内积：*。*其中的每一行分别表示一个话题，而每一列表示一个句子。Hachey[41]也提出了一种基于SVD的算法，该模型更接近LSA的原始思想，根据大量背景文档的词共现信息来建立初始矩阵*A*，而不是仅仅要摘要的输入文档。经过实验对比发现，矩阵分解可以大幅改进摘要质量，然而词共现特征并没有比tf\*idf提高很多。

* + 1. 互参信息和修辞结构

反映同一语义实体的不同语段除了词汇链、LSA，也可以通过利用输入文档的互参信息(Coreference Information)实现。起初互参信息在文本摘要中专门用来衡量句子重要性[42][43]，但并没有得到显著提高生成摘要的效果。Steinberger在[44]中使用首语重复法(Anaphora)生成互参信息并将其输入到基于LSA的摘要模型[40]中，大幅提高了摘要效果。实验中所有指向同一个实体的所有词句，都被替代为首次提到该实体的词句；然后将生成的文档输入到传统的LSA模型中产生摘要，结果发现生成的摘要质量相比传统LSA模型反而变差了。而另一个实验中每个句子的实体都被当成衡量句子重要性的特征，但不改变指向实体的词句，最终生成的摘要质量有明显提升。

另外一些研究通过分析输入文档的论述机构来产生单文本摘要，其中包括修辞结构理论(Rhetorical Structure Theory, RST)[45]。RST需要将文档表示成一棵树，RST中最小的文本分析单元是Elementary Discourse Units(EDUs)，通常是子句从句。相邻的EDU通过修辞关系连接起来，更大的单元递归参与形成关系，直至形成一个覆盖全文的多层树结构。Marcu[46][47][48]证明RST能够在单文档新闻摘要得到很好的结果。

* 1. 特殊文体或领域的摘要方法

前面各种摘要算法都是针对一般摘要而设计的，却也往往能生成较高质量的面向用户需求的摘要，比如基于查询摘要。然而，这些方法通常不适用于特殊文体或者某些特别领域的摘要。当输入文件有某种特定结构或者其他独特的特性，摘要算法可以利用这些特性使得摘要质量进一步提高。比如期刊文章经常有一个总结章节直接概括出文章的关键信息，这个总结段显然能给摘要提供很多关键信息。又比如在医学或者法律等特殊领域，往往对摘要结果有专业性和准确性要求，而且除了输入文档外通常有丰富的相关资源来帮助自动摘要过程。本小节主要介绍一些文体结构和领域不同于新闻的文本摘要研究。

* + 1. 医学摘要

医学领域的摘要任务是不适用普通摘要算法的典型例子。在这个领域，摘要算法通常由精确定义的需求决定，比如帮助医生对治疗手段的决策，或是去检索和一些特别病症相关的最新研究。医学文章往往有固定可知的结构，而且更重要的是，医学界会有大规模的医学资料提供数以百万的概念名称及相关语义信息。

当病人等非专业人士去搜索医学信息时通常没法给出精准的需求，却可能需要一些专业信息。Centrifuser[49][50]是个能帮助用户搜索信息的摘要系统，实现了多文本摘要和基于查询摘要。它能够从多个文档中选择和查询问题相关的话题段落，这些段落能够给出指向原文档的导航链接，因此Centrifuser更像是指示性摘要，给用户提供了快速浏览功能。

医学期刊文章摘要对于医药知识的搜索极有参考意义。著名的期刊摘要TAS[51]是病症数字图书馆PERSIVAL[52]的一部分。TAS也是个基于查询的多文档摘要系统，能将根据查询搜索出来的论文集合进行摘要生成摘要。TAS的特殊之处在于它并不是抽取式摘要，而是通过从自居单元中抽取信息填充到预定义的模板，再进行排序重组生成摘要句子。此外，TAS还根据病人记录来过滤论文，给医生提供适合该病人的相关论文摘要结果。

* + 1. 期刊摘要

期刊论文往往有较为固定的结构形式，有些摘要算法从结构入手来寻找关键信息，比如Teufel和Moens[53]根据句子的修辞状态(Rhetorical Status)来提取摘要。

此外，存在引用的论文通常会包含被引语段的概括信息，因此有研究者根据论文的引用链接来辅助抽取摘要。Nanba和Okumura[54]创建的摘要系统能够为若干篇相关科学论文自动生成简介，并且将论文之间的关系可视化。通过规则匹配，该系统能够识别被引用的区域，并将每个被引区域分成三类：(1)引用的模型或者方法(2)和相关工作的对比或讨论(3)其他。而Mei和Zhai[55]提出了影响力摘要(Impact Summarization)，通过寻找被大量引用的语段区域，并利用语言模型从这些区域中对句子进行排序，进而抽取摘要。另外引用摘要(Citation Summarization)是另一种基于引用信息的单文本摘要模型。Qazvinian 和Radev[56]提出从输入文档引用的其他文档中提取出引用部分的重要语段作为摘要，而这些语段很可能重复表述、存在冗余信息，因此可能通过2.1.2节中的句子聚类或图模型来抽取和提炼关键句。

* + 1. 邮件摘要

邮件摘要分为单封邮件摘要和多封邮件摘要，包括对邮箱摘要和邮件线摘要。对每封邮件提取一个话题可以帮助用户了解收件的优先级，知道哪些邮件必须立刻回复以及迅速查找相关历史邮件。而邮箱摘要可以提供一个浏览界面帮助用户迅速定位感兴趣的邮件。

单封邮件通常通过选择能反映邮件主题的名词短语来完成摘要。Gister[57][58]是个邮件自动摘要系统，能够结合语言过滤以及机器学习方法来选择名词短语来组成摘要。而微软的研究者[59]通过支持向量机(Support Vector Machine, SVM)对邮件句子进行分类，判断是否为任务指示。先前的全文摘要算法也可以适用于邮件摘要，但是需要一些预处理。Lam[60]在对邮件摘要之前，先去掉了开头致辞、引用文本、结尾谦词以及邮件签名，然后使用一个IBM的自动摘要系统[61]完成摘要。

另一种常见的邮件摘要任务是对一个邮件线(Email Thread)进行摘要，邮件线即主邮件(根邮件)被反复回复后形成的邮件系列。Nenkova和Bagga[62]提出的摘要系统能从邮件线中生成指示型摘要，他们只对邮件线的前两封邮件各抽取一个句子：在根邮件抽取包含最多出现在邮件主题的名词的最短句子；在后续回复邮件中，提取和主邮件有最多重复词语的句子。而Rambow[63]从邮件的对话本质入手，使用了两种基于机器学习的抽取式摘要算法：第一种和普通摘要类似，使用词频等特征；第二种算法依赖邮件结构特有特征，包括回复邮件数、句子主题相似度等。而他们使用的基于规则的分类器RIPPER，并得到非常好的摘要效果。

此外，对存在大量邮件、邮件线的邮箱和邮件归档的摘要是个困难的任务。Newman和Biltzer[64]提出的多文本摘要模型能够快速浏览邮箱。而Carenini提出利用图模型来进行群邮件摘要，并展示邮件个体之间的联系。图模型中反复出现的词被定义为线索词(Clue Words)，统计线索词的频率可以用来给句子的重要性评分。实验证明基于线索词的模型效果非常理想。

* + 1. 网页摘要

网页信息的内容越来越丰富，为了满足人们快速从网页获取知识，网页摘要的需求变得愈加紧迫。大部分网页摘要研究都依赖于一个开放目录DMOZ[[2]](#footnote-2)来获取大规模网页摘要资源，DMOZ中网页内容和对应摘要被组织成多层结构，越是热门话题的网页节点越在上层，每个网页都有对应的人工摘要。

早期的网页摘要都基于网页本身去建模[65][66]。Berger和Mittal[65]使用基于统计机器翻译的模型，来给DMOZ的网页和摘要内容进行对齐，并使用了两个模型分别生成摘要词及其顺序，最后生成的摘要中会用到输入网页中未出现过的词。Buyukkokten等[66]用了基于Luhn思想的简单摘要模型，能够生成可变长度的摘要并显示在小型手持设备上。而Delort[67]通过计算余弦值当作句子之间的重复率，来获取句子的网页相关性。Sun[68]利用微软的搜索引擎来获取用户对网页的请求以及数据的点击流，并以三元组形式表示（用户,查询，网页），并结合Luhn的模型和LSA算法提高了摘要质量。Choi等[69]利用非监督算法提取网站摘要，来增强赞助广告的表现。

在线论坛和博客博文的自动摘要也有不少相关研究。Zhou和Hovy[70]利用类似语音、邮件摘要模型来对论坛进行摘要，他们注意到论坛通常包含关于多个交错的话题的异步交互，通过识别回复对象和讨论话题，最终给每个话题提供子摘要。Hu[71]通过对博客博文的评论内容进行词频统计，来衡量句子重要性，最终生成博客摘要。

1. 文本的语义表示
   1. 引言

在对文本进行NLP任务时，除了诸如分词、去停用词、词根化(Stemming)等一系列数据清洗的必要步骤外，还需要将文本转为可计算的特征表示，作为模型的输入进行学习。文本的语义向量化一直是NLP界最热门的研究方向之一，尤其当深度学习技术的发展和普及后。

文本向量的语义性在数学上主要体现在，语义越相似的文本之间的语义向量的距离也越近，包括欧式距离和夹角余弦：

|  |  |
| --- | --- |
| 欧式距离: | (3-1) |
| 夹角余弦： | (3-2) |

文本向量化模型分为两种，一种是忽略文本中的单词顺序，只保留句子的统计信息如词频、多样性。最经典的就是词袋模型(Bag of Words,BOW)。词袋向量可以作为文本的特征直接输入到分类器，进行文本分类和情感分析等任务。

词袋模型的建立过程很简单，下面是两个简单的文档样本句：

(1) Jack likes to play music. Rose likes music too.

(2) Jack also likes to play football.

根据这两个样本建立一个词典包含所有出现的单词，如果是未清洗过的原始文档，则以词根建立。

{ " Jack ", "like ", " to ", " play ", " music ", " Rose ", " too ", " also ", " football " }

根据每个词语在词典的索引和在文档中的出现频次可以对以上两个句子建立词袋向量：

1. [1,2,1,1,2,1,1,0,0] (2) [1,1,1,1,0,0,0,0,1]

每个维度的权值可以取例子中的词频，也可以是tf\*idf, 甚至可以是二元值，即1表示出现0表示未出现。而词袋模型的缺点也很明显，向量表示的特征维度高、非常稀疏，而且缺失语义信息。

另一种向量化模型和单词输入顺序有关，可以区分"Mary loves Jack"和"Jack loves Mary"这两个句子。词语的顺序很难直接量化输入，但循环神经网络(Recurrent Neural Network, RNN)能够通过时间序列的变化，实现变长词串到语义向量的映射(sequence-vector)。

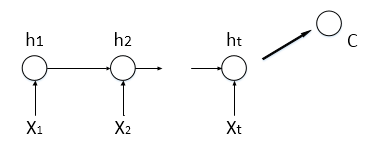


图3.1 RNN结构示图

如图，为输入文档的词串，RNN将文档看成一个随时间变化的词序列，每输入一个新词，隐含层就进行一次更新:

|  |  |
| --- | --- |
|  | (3-3) |

这样，隐含层充分利用了上文的历史信息，并始终保持最新状态，直到最后输出文档的语义向量*c*。由于带有词序信息，RNN训练出来的文本向量相比词袋模型更具有语义，在实验上直接体现在输入到同个分类器中，带词序信息的模型在情感分类任务的表现更好[72]。此外，实现了变长串到向量(sequence-vector)映射的模型，可以适用更复杂的NLP任务，比如配合语言模型可以很容易扩展成sequence-vector-sequence模型，以完成直接从源语言到目标语言的机器翻译任务[73]。然而RNN优化困难、结构复杂，且容易丢失较久之前的历史信息。

本文中尝试使用了两种简单的方法学习文本的紧凑的语义表示，分别是词嵌入加权和深度降维，分别在3.2节和3.3节阐述。

* 1. 词嵌入加权

词嵌入(Word Embedding)表示词语在连续空间到特征向量的态射，相当于单词的语义向量表示。相比于文本的语义向量化问题，单词的语义向量化问题简单很多。对文本中所有词的语义表示进行加权平均得到的向量，可直接作为整个文本的语义表示，用数学公式表示为：

|  |  |
| --- | --- |
|  | (3-4) |

其中是该文档中出现过的所有词，表示该词的语义向量(词嵌入)，为词嵌入的加权系数函数。公式3-4中有两个关键点，一个是词的语义表示形式，一个是加权函数。

对于词的语义表达形式来说，最简单的是one-hot表示，即将每个词表示为一个长度为词汇表大小的bit向量，其中只有一个bit为1，其他bit为0。虽然one-hot能够唯一区分不同的词，但是这种简单的映射完全没有考虑词和词之间的关系，语义信息更无从谈起。因此one-hot形式作为词的直观表达，常常在一些复杂模型中当作原始输入进行处理。

词嵌入最常用的获得方式就是神经网络。Bengio[74]最早通过训练神经语言模型来计算词序列的概率，顺便得到了一份词嵌入作为副产品。如今研究者们通过神经网络来处理包括词性标注、机器翻译的各种自然语言处理任务时，都能顺便得到词向量矩阵，而且得到的语义效果通常比简单的语言模型更好，但要求训练集文本必须有词性或者译文等类标。因此本文使用最简单的基于n-gram语言模型的前向神经网络来训练词嵌入向量矩阵。

基于n-gram的统计语言模型的思想可以表示为:

(3-5)

其中表示第*t*个词，表示()词序列。前半部分为语言模型中词序列的联合概率表达形式。后半部分的为n-gram的基本假设，即每个词只和上文*n*-1个词有关。

基于n-gram的神经网络语言模型结构如图：

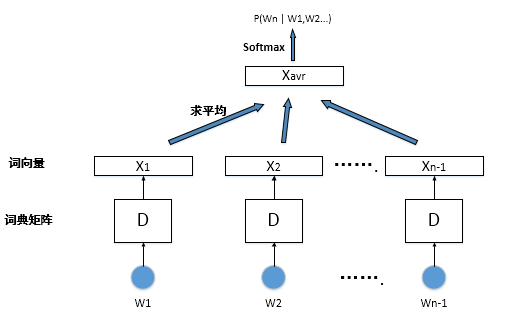


图3.2 基于n-gram的前向网络语言模型

图中表示前*n*-1个词，用one-hot形式输入。矩阵D是词向量矩阵，每一列都是一个单词的词向量。就是对应的语义向量，可以通过单词对应的one-hot向量点乘词典矩阵D得到。是前*n*-1个单词的语义向量均值，输入到一个分类器中去预测下一个单词。最后的输出值是每个单词的预测概率，可以通过softmax激活函数实现：

|  |  |
| --- | --- |
|  | (3-6) |

其中是输出层单词激活前的值，*y*的计算方法为：

|  |  |
| --- | --- |
|  | (3-7) |

其中*U*,*b*分别为softmax激活函数的参数。对于输入文档{},模型的目标函数是最大化预测概率的似然函数：

|  |  |
| --- | --- |
|  | (3-8) |

其中*T*是文档长度。用随机梯度上升(Stochastic Gradient Ascent, SGA)可以求出网络中的各个参数：

|  |  |
| --- | --- |
|  | (3-9) |

其中(*U, b, D*)。词向量矩阵D是本文所需要用到的参数。

词向量的权重函数*weight()*要能够反映词在所在文段中的重要性。单词的重要性主要体现在两个方面：统计信息和位置信息。统计信息可以用tf\*idf值衡量，位置信息可以通过该词是否在标题出现过来判断：

|  |  |
| --- | --- |
|  | (3-10) |

其中是位置信息函数，如果在标题出现值为1，反之为0；是统计信息函数，为的tf\*idf值。是个增益参数，用来放大在标题中出现过的词的权重，本文取值0.25。

* 1. 深度降维

通过模拟人脑大脑皮层的感知过程，深度模型可以生成人工智能级别的摘要。神经生物学表明，大脑皮层之所以有多种认知能力，是因为其复杂的层状物理结构[75]。当神经语言系统中的多个脑区，包括布罗卡氏区(Broca’s Area)和韦尼克区(Wernicke’s Area)，进行即使最简单的词素-语义处理(Lexical-semantic Processing)，也会有数十层的皮层参与其中[76]。受此启发，本文使用了一个无监督的多层神经网络模型来对文本进行降维，学习出文本的深度特征作为语义表示。

Hinton[77]在科学杂志首先提出通过使用深度自编码网络来对样本进行维度消减来提取深度特征，并成功运用在手写识别和文件检索中。类似于PCA的非线性泛化，深度自编码器通过一个自适应的多层编码网络(Encode Network)将原始输入压缩成低维紧凑的特征码(Feature Coding)，再通过一个结构相似的多层解码网络(Decode Network)将特征码恢复数据。随机初始化多层神经网络中的边权值很可能导致梯度扩散和局部极值问题，用限制玻尔兹曼机(Restricted Boltzmann Machine, RBM)进行预训练来初始化网络可以避免这些问题。

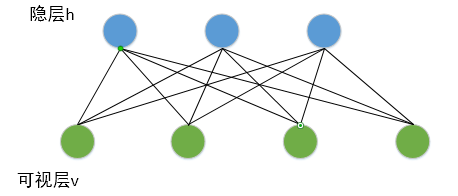


图3.3 RBM双层结构视图

RBM网络在结构上有两层，下面*v*层为可视层(Visible Layer)，一般是输入层；上面*h*层为隐含层(Hidden Layer)，即特征表示层。*v*层和*h*层相互连接构成二部图，而同一层的节点间互不连接，因此在隐藏节点和可视层的节点分别是条件独立的：

|  |  |
| --- | --- |
|  | (3-11) |

因此在输入*v*的时候可以通过得到*h*，得到*h*后再通过可以得到*v’*。RBM预训练的目的是让*v’*接近 *v*，从而使得隐藏层*h*可以作为*v*层输入数据的另一种特征表达。

本文中我们最终建立并训练了一个2000-800-300-100的多层神经网络，能够从一个句子或一篇文档中学习出100维的特征码。为了训练这个网络，我们从DUC语料库中的纽约时报(New York Times)和美联社(Associated Press)中摘选了10000个句子，进行去停词、词根化等数据清洗步骤后取词频最高的2000个词作为词典，将每个句子表示为2000维的词袋向量作为网络的输入。底层的输入层和顶层的输出层都是2000个节点，对应句子或文档的2000维语义向量。

整个建模过程包括三个阶段: RBM预训练, RBM拼接和整体调参。

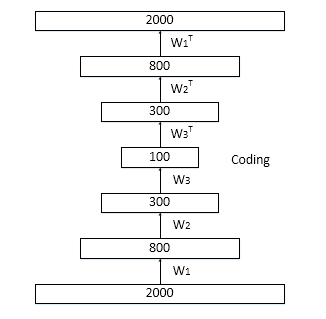


图3.4 深度网络结构,100个节点的隐层为coding层

在预训练阶段，需要训练三个RBM,分别是2000-800, 800-300, 300-100。RBM是个生成模型，预训练过程是无监督的。RBM的能量函数定义为：

|  |  |
| --- | --- |
|  | (3-12) |

其中*a*、*b*分别是可视层*v*层和隐藏层*h*层的偏置参数，*w*是联接网络矩阵。RBM是一种能量模型(Energy-based Model, EBM)，能量模型的概率分布通过能量函数定义，RBM的联合概率分布为：

|  |  |
| --- | --- |
|  | (3-13) |

其中*Z*是一个归一化因子。当输入层*v*确定时*h*的概率分布为：

|  |  |
| --- | --- |
|  | (3-14) |

预训练的三个RBM中，除了底端的RBM (2000-800)有连续数值输入，其他的两个RBM的隐含层和可视层的所有节点均为01二值，即。每个节点被激活的概率为：

|  |  |
| --- | --- |
|  | (3-15) |

其中为sigmoid激活函数。为了更新参数*W, a, b*，RBM采用的Gibbs采样(Gibbs Sampling)。Gibbs采样可以根据一个复杂概率分布生成数据进行采样，因此可以根据*v*层采样*h*层，再从*h*层采用*v*层，反复迭代：

|  |  |
| --- | --- |
|  | (3-16) |

其中和分别是可视层和隐含层在第*i*次迭代采样的结果。利用这些样本可以用梯度上升方法进行参数更新，具体算法可以用Hinton的对比散度(Contrastive Divergence)方法。

训练完一个RBM的参数之后，可以将这个RBM的隐藏层激活值作为下一个RBM的可视层输入进行训练。经过逐层预训练后，每个RBM的两层节点都成为特征表示，进而将3个RBM联接起来，并复制一份进行翻转，可以拼接成一个图3.4中所示的多层网络。其中底部及顶部的RBM(2000-800)的样本输入值为0~1的连续值，并非01二元值，因此在顶层输出的时候用logistic函数可以得到0~1的激活值，用交叉熵作为整个网络在整体调参阶段的目标函数：

|  |  |
| --- | --- |
|  | (3-17) |

其中代表输入层节点*i*的强度，代表输出层节点*i*的重构值强度。最小化交叉熵的优化过程可以用梯度下降求解。用交叉熵作为损失函数的原因以及梯度下降法的迭代过程会在4.3.1节里详细阐述。为了加快训练过程，可以将样本集划分成minibatch进行训练。

不同于图像处理，图像输入是一个像素向量，每个像素的取值固定为0-255的整数。因此只需要将像素向量除以255就能保证每个输入值在[0,1]之间，这时候输出层用一个logistic激活函数就可以拟合要重构的输入值。但是本文模型的目标是要对文本进行降维处理，输入文档样本通常是以tf值为权重的词袋向量，然而tf值会随着文本长度变长而增加，取值范围不确定，因此词袋向量必须除以文档长度，变为一个词概率向量作为原始输入。这样会导致两个问题：(1)以词概率向量作为输入时，输入层的所有节点之和恒为1，但是输出的logistic激活函数没有这个约束，并不保证能理想的重构输入；(2)词概率的值往往很小，尤其当文本很长的时候趋近于0，这将使得节点的激活值非常小，进而导致训练出来的网络边值往往会很大，使得重构效果非常差。

为了解决这个问题，Hinton的trick是让底层的RBM(2000-800)以词袋向量输入，避免输入值太小导致无法激活，而顶端的RBM(800-2000)改用softmax激活函数，以保证输出层的激活值之和为1，这时候只需要将目标函数改为激活值向量和输入文档的词概率向量的交叉熵，就可以解决上文提到的两个问题。而Salakhutdinov[78]使用了一个更为复杂的约束泊松模型(Constrained Possion Model)对输入向量建模。

上述的两种模型都比较复杂，且都用了softmax作为激活函数使得计算量比较大。本文给出的解决方法是对输入进行归一化处理，将词概率向量进行离差标准化(Min-max Rescaling)：

|  |  |
| --- | --- |
|  | (3-18) |

Min-max归一化后的输入向量其实就是之前的词概率向量的线性投影，但每个维度的值都被放大了几倍到几十倍，并依然保持0-1的取值范围。同时，归一化后的向量并没有各维度值之和为1的约束，因此只需用简单的logistic作为激活函数就能有效实现数据重构。

* 1. 实验对比

为了证明两种文本向量化方式的有效性以及比较两者优劣，本文做了一个文句分类的对比试验。从DUC2007语料库中我们抽取了6篇新闻，分别是"An Interview with BURMA’s AungSan Suu Kyi"， "Gingrich Wraps up His Official Duties"，"Spain Facing its Own Fears on Separatism"，"Search for Suspected Abortion Clinic Bomber Still Wide Open"，"E-Commerce: Coming Soon to A Coffee Shop Near You"和"Napster Strikes Dissonant Chords"。每篇新闻均大约有一百个句子。

分别对所有句子进行两种方式向量化后，使用常见的高维数据可视化工具t-SNE (T-distributed Stochastic Neighbor Embedding)[79]来观测他们的分类表现。t-SNE是一种流形学习方法，通过保持数据点的相邻关系把数据从高维空间降低到二维平面，特点是数据分布的边缘呈圆形，适合实验结果作图对比。t-SNE的降维原理就是让高维数据在映射到的低维空间中依旧保持和之前类似的分布。在原高维空间中两个点和的相似性定义为：

|  |  |
| --- | --- |
|  | (3-19) |

其中*n*为点的个数，表示*j*为*i*的邻居的概率：

|  |  |
| --- | --- |
|  | (3-20) |

其中是以点为中心的高斯分布的方差。而在映射后的低维空间，点和的相似性计算公式为：

|  |  |
| --- | --- |
|  | (3-21) |

通过优化算法可以使映射后的低维空间的相似度矩阵和原高维空间的相似度矩阵相近，此时得到的映射点就是降维后的数据。

对两个语义向量化模型学习后的所有句子样本进行t-SNE的降维，左侧是词嵌入加权模型结果，右侧是深度降维模型结果。其中同一个颜色的点代表来自同一篇新闻，因此聚类效果越好代表句子分类效果越好，即语义向量化的质量更高。

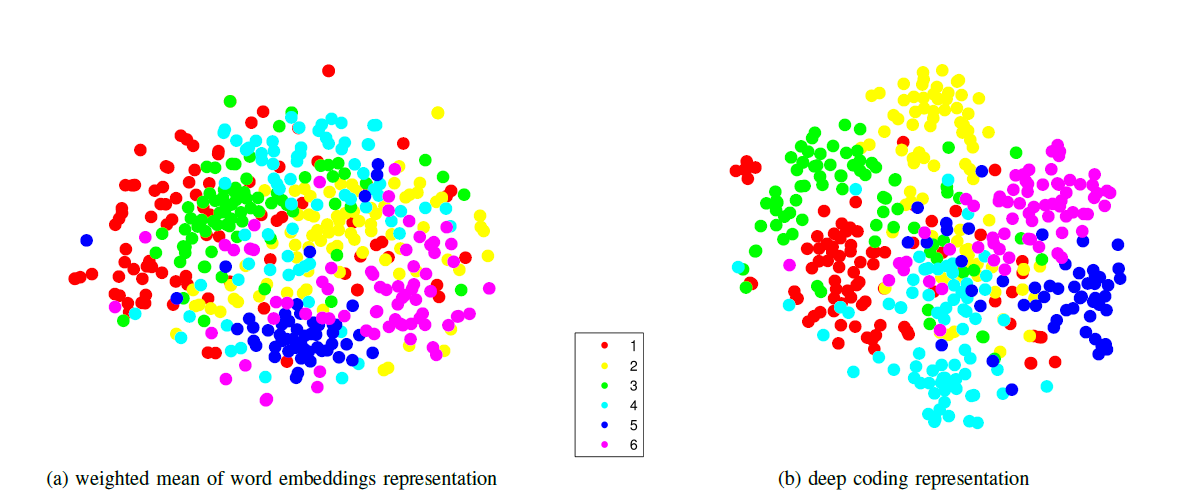


图3.5 语义向量化实验对比，左为词嵌入加权，右为深度降维

实验结果图像表明两种向量化方式在文句分类方面的效果都很好，说明两种模型都能学习出文本的低维语义表示。而对两种方法的表现进行比较，可以发现深度降维得到的向量分类特征更加明显，如图3.5所示每篇文章的句子集都构成了一个扇形，视觉上可以直观的通过斜率加以区分。

* 1. 本章小结

在实现自动摘要算法的过程中，文本的语义表示是不可或缺的重要步骤。一方面，将原文整体以及所有句子的语义向量化后，才能为后续的重构算法提供原始输入；另一方面，将文本用语义特征而不是简单的统计特征去向量化表示，可以使得文本向量包含更多的语义信息，从而让生成的摘要句子更契合主题。

相比于词袋模型的高特征和语义缺失，我们希望能够生成紧凑和富有语义的文本向量。当前最热门、效果最好的语义向量化是RNN，但是结构比较复杂。在本文中用了两种相对简单的模型生成文本的语义表示。一种是词嵌入加权，将复杂的句子向量化(text-vector)问题转化为相对容易的词向量化(word-vector)问题，词嵌入可以通过一个基于n-gram的前向神经网络训练。另一种方法通过模拟大脑皮层的深度结构，训练一个多层的神经网络学习得到文本的深度特征作为语义向量，其中利用输入归一化简化了模型。

经过句子分类的实验论证，两种方法训练出来的语义向量表现良好,都可以学习出文本的紧凑的语义表示。

1. 原文语义重构策略

经过第三章的两种语义向量化方法后，可以得到每个句子甚至整个文本的语义表示，进而再输入到设计好的重构模型进行摘要提取。语义向量的意义主要在于提高了生成摘要的凸显性。而本章的重点就是设计一个有效的重构策略，让生成的摘要能够以最小的重构损失去还原原文档的语义。

重构模型的基本思想是得到能够最佳还原文本的句子集作为生成的摘要，其中有效的重构策略是获得高质量摘要的关键。重构策略包括重构函数的设计，重构源和重构对象的选取。具体来说，重构函数可以是线性的，也可以是非线性的。重构源可以直接对所有可能的候选句子集进行重构，找到最优句子集直接作为摘要；也可以是以单个句子逐一重构，得到每个句子的重构贡献度，以此排序取top-k。重构对象可以对所有句子逐一重构，累积重构误差；也可以是直接对整个文档进行一次重构。

He在[80]中重构策略是利用线性重构函数对原文中的所有句子作为重构目标，逐句进行重构并累计重构误差作为目标函数，最后找到总重构误差最小的候选句集作为摘要。Liu在[81]采取的重构策略是构建一个由若干RBM栈式堆叠而成的深度模型进行逐句重构。重构源和对象都是每个句子的词袋向量，经过深度网络的非线性重构后能在隐含层节点提炼出若干个原文概括性最高的关键词，再对每个句子以关键词为权重进行打分选取top-k。

在4.2节和4.3节中，分别设计了基于线性函数和非线性函数的重构策略，对原文语义进行重构。此外，4.4节又从冗余消减方面，对重构策略进行一定程度上的优化，以提高生成摘要的质量。最后在4.5节进行实验对比论证。

* 1. 线性重构策略

He的重构策略是对原文中的所有句子，逐句进行重构并累计误差，最后找到总重构误差最小的候选集。该重构过程的目标函数为：

|  |  |
| --- | --- |
|  | (4-1) |
|  |
|  |

其中*V*是所有句子集合，*X*是候选句集，也是*V*的子集。是每个句子的线性组合系数。是正则化系数，防止参数过大导致的过拟合现象。

该目标函数的优化过程是NP-hard，无法在多项式时间内解决，经过二次规划后可以等价转化为如下函数：

|  |  |
| --- | --- |
|  | (4-2) |
|  |

该目标函数可以通过贪心算法获得近似解，最后得到的最优*X*就是生成的摘要。

和He的原句重构策略不同，本文的重构对象是原文的语义向量，而不是重构所有句子叠加误差，避免了复杂的泛函分析，降低了计算复杂度；另一方面He的重构模型用的是词袋模型作为输入，而本文以语义表示作为输入，以重构原文语义为目的，使得生成的摘要更具语义性和凸显性。

* + 1. 目标函数

用第三章中的语义表示模型，将清洗过的文档中的所有句子以及整个文档分别用语义向量表示，令句子集为矩阵*S* = {,,…}，其中每列的表示第*i*个句子的语义向量；又令*d* 为整个文档的语义向量，特征维度同。在本文中，重构源是所有句子，经过线性组合后得到重构结果：

|  |  |
| --- | --- |
|  | (4-3) |

其中是句子对应的重构系数。重构目标是原文的语义向量，并用欧式距离平方作为重构损失：

|  |  |
| --- | --- |
|  | (4-4) |

表示线性组合系数向量，表示向量*a*到*b*的欧式距离。重构损失表示为欧式距离的平方并乘以因子1/2的目的是为了求导方便计算。该目标函数在形式上同回归问题中的最小二乘估计（Least Square Estimation, LSE），求解过程是一个连续可微的凸优化问题，可以直接用梯度下降法求最优解。中系数为0的项对应的句子被认为文档不相关句子，因为这些句子并未参与原文重构。因此最终生成的摘要只会从中非零项对应的句子集中挑选。

损失函数除了用欧式距离计算的重构误差外，生成的摘要总是有个长度限制。比如做后续对比实验的摘要测试系统DUC中，对每篇文档的摘要结果有严格的长度限制要求。因此我们希望组合稀疏*w*中含有尽量少的非零项，必须在目标函数上附加一个稀疏惩罚项：

|  |  |
| --- | --- |
|  | (4-5) |

其中是L0范数形式，用来统计向量中非零维度值的个数；是一个正则化参数，用来平衡稀疏性和重构损失。

然而L0惩罚项的优化过程是个NP-hard问题，无法在多项式时间内解决。常用的方法是用L1正则项或者L2正则项替代L0正则项，将目标函数变成凸优化问题。

L1正则项的表达形式为：

|  |  |
| --- | --- |
|  | (4-6) |

L2正则项的表达形式为：

|  |  |
| --- | --- |
|  | (4-7) |

L0替代为L2范数后目标函数变为岭回归(Ridge Regression)问题，也是个连续可微的凸优化问题。对比之前的最小二乘，L2范数能够在给出最优解和求得很小的系数之间做平衡。尽管L2正则项能够实现参数收缩(Variable Shrinkage)的目的，但是它并没有很直接地将系数强制为0，而是仅仅逼近为0。[82]已经证明L1范数能比L2范数在处理稀疏性方面更优秀，尤其当矩阵*X*中存在不相关的特征的情况。

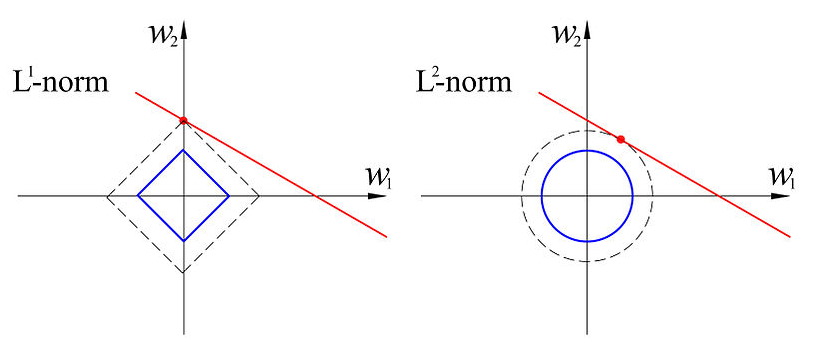


图4-1 L1范式相比L2范式能得到更稀疏的最优解

图中模拟的是二维空间的优化过程，虚线部分表示随变化而伸缩的范数约束域(norm-ball)，红线为损失函数等高线，红线和norm-ball相交的点就是最优解的系数。从图上也可以直观地看出L2能比L1得到更平滑更小的系数，但是L1的norm-ball更容易在特征轴上相交，即能得到更稀疏的系数。基于以上考虑，我们用L1正则化作为稀疏惩罚，并得到对应的目标函数：

|  |  |
| --- | --- |
|  | (4-8) |

其中是一个用来平衡稀疏性和重构损失的参数。带L1范式正则项的二次回归问题曾经分别独立的表达为Least Absolute Selection and Shrinkage Operator(LASSO)回归问题[83]以及Basis Pursuit Denoising(BPDN)[84]问题。公式4-8称为Lasso的拉格朗日形式或者非约束形式(Unconstrained Formulation)。非约束形式Lasso是个凸优化问题，但对时的并不可微，无法用类似最小二乘估计或者岭回归问题的方法去求全局最优解。因此更多地被写成约束形Lasso (Constrained Lasso)去求解：

|  |  |
| --- | --- |
|  | (4-9) |
|  |

约束形Lasso的目标函数是个凸函数，而且约束条件定义了一个凸集，因此是个凸优化问题，而且任何满足约束的局部最小值都是全局最小值。在一般的Lasso回归模型中，参数的取值需要通过交叉验证(Cross-Validation)方法去确定，因为值太大会让回归系数过于稀疏导致拟合不足；而太小会导致过拟合(overfitting)现象。但是在本文中并不需要学习出一个能精准预测类标的回归模型，而是通过给定的*S*、*d*优化目标函数以得到最优的。因此，通过手动或设定程序自动调整的值，可以获得理想的进而生成满足篇幅要求的摘要。

此外当生成摘要的原文档比较短的时候，即文档的句子数*n*远小于语义向量的维度*m*时，用L1范数正则项可能会遇到退化现象，即当存在一组相互高度相关的句子，Lasso会倾向于只从中选择一个句子而忽略其他几句。原Lasso表达式附加L2范式惩罚的弹性网络(Elastic Net)[85]可以克服这个缺点：

|  |  |
| --- | --- |
|  | (4-10) |
|  |

其中是一个平衡参数，当 = 0 和 时目标函数分别变成Lasso和岭回归，因此可以看出弹性网络是Lasso和岭回归的一种折中。往往是一个接近1的值，使得弹性网络既有Lasso的稀疏功能，又能避免由于句子间高度相关而导致的选择退化。弹性网络适用于文档句子数*n*远小于语义向量维数的情况*m*，而本文实验所用的标准测试集DUC每篇文档都有300个句子以上，远大于语义向量维度100，因此对比实验所用模型更适用Lasso算法。

公式4-10通过手动调整的值，可以获得理想的进而生成满足要求的摘要。摘要的篇幅限制包括句数限制和字数限制。当限制句数时：

|  |  |
| --- | --- |
|  | (4-11) |
|  |

其中*t*的值为摘要结果的限制句数。当限制生成摘要字数时：

|  |  |
| --- | --- |
|  | (4-12) |
|  |

其中表示第*i*个句子的长度，表示摘要结果的字数限制，表示求0次方值，此处定义。

* + 1. 优化方法

Lasso是一种典型的凸优化以及二次规划问题，可以求得全局最优解。学术界对Lasso问题已经给出多种解法。Efron的最小角度回归(Least Angle Regression, LARS)[86]算法不仅发现了Lasso和boosting之间紧密的数学联系，还通过结合前向逐段回归(Forward Stagewise Regression)将求解过程优化至相当于最小二乘法的复杂度。而Friedman[87]的坐标下降法(Coordinate Descent)是最快的Lasso解法。

用LARS算法求解约束性Lasso的步骤：

* 初始化为0，此时残差*δ = d*。找到与*d*相关性最大的自变量
* 若和*δ*正相关则增加对应系数的值，反之减小，同时更新残差 。直到另有与*δ*的相关性与同样大
* 在(,)的联合最小二乘方向同时增加(,)，直到另有与当前残差*δ*也有同样大的相关性
* 迭代直到所有*s*都被激活

数据处理也是重要的一个步骤，包括数据中心化和归一化。在进行优化之前，文本矩阵*S*的各列都要标准化为0-均值、1-方差，而文档向量*d*将中心化为0-均值。通过中心化为0-均值，可以消除截距。通过对*X*进行1-方差，可以保证*S*中的所有列都处于相同的数值尺度进而让所有的系数也处于同一范围，这样避免了一个高范数的列倾向得到很小的系数。

可以证明稀疏参数*λ*值变大，得到的最优系数*ω*的非零项越多，而且将包含较小的参数得到最优解的所有非零项，因此得到的摘要字数更长：

|  |  |
| --- | --- |
|  | (4-13) |

其中是在给定*λ*下用LARS或者坐标下降算法求得的最优解。生成摘要的长度关于变量*λ*单调增，因此可以用二分迭代法确定*λ*的取值。

|  |
| --- |
| 算法4.1：线性重构生成摘要 |
| Input：文档的语义向量*d*，文档中所有句子的语义表示集合矩阵*S*，生成摘要的字数限制*l*  output：生成摘要句子的索引集合summary  Description：从文档中通过语义重构的思想找到线性重构损失最小的句子集作为生成的摘要 |
| 1. 将*S*的所有列及向量*d*中心化为0-均值，另将*S*的所有列标准化为1-方差 2. 初始化*λ*的大小极值 ， 3. 如果, 输出 *summary* = {}。是的取值精度，本文取0.0001。否则继续步骤(4) 4. *λ*取中值，并用LARS算法或者CD算法求出对应的最优解 5. 计算生成摘要的字数长度，如果 < *l*，表明组合系数过于稀疏，即*λ*值偏小，因此更新极小值新来缩小*λ*取值范围；反之*λ*值偏大，更新。返回步骤(3) |

* 1. 非线性重构策略

在4.2节的重构策略中，我们采用了一个线性的重构函数，通过从候选句子集合进行线性组合，对原文档的语义向量进行重构。而在本节的重构策略中，采用的是非线性重构函数。相比于线性模型，非线性模型通常更加灵活、更加平滑，对数据的拟合更加精确；同时，从人脑决策的角度分析，人工生成摘要的过程是个大脑皮层中多层神经参与活动认知的过程，而非线性的模型通过学习更高层、更抽象的表达，能够模拟这一过程。最简单常用的非线性模型就是采用非线性激活函数的神经网络。

Liu在[81]采用了一个思路类似重构的深度模型。该模型通过由多个RBM堆叠而成的多层神经网络，对输入的文本进行概念提取(Concepts Extraction)，因而隐含层也被称为概念层。隐含层节点经过计算可以获得输入文档的关键词，进而作为句子的重要性评价进行排序抽取。

* + 1. 模型结构及训练

本文使用了一个100-80-100的两层神经网络去模拟了文本的重构过程，结构包括了输入层*x*、话题提取网络H1、隐含层*c*、原文重构网络H2和输出层*y*。

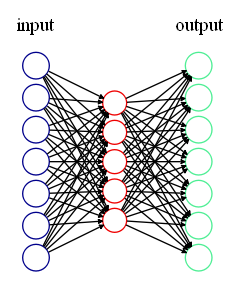


图4-2 前向重构网络结构示图

输入层*x*：输入文档(句子)的语义向量，包含100个神经元。

H1：对输入文档(句子)进行概念提取，由于经过了非线性函数的激活，可以得到相比输入层更为抽象的特征表示。

隐含层*h*：相当于输入层压缩后的抽象表示，每个神经元可以理解成一个话题，每个输入文本都映射成80维的话题向量。

H2：构成一个重构网络，对隐含层的话题向量进行原文重构。激活函数是sigmoid函数，以达到非线性重构的目的。

输出层*y*：重构得到的结果。

在该重构神经网络的训练阶段，每次输入的样本是文档的句子*x*。该网络尝试学习出一个的重构函数，因此对于每个句子的重构过程来说，重构误差可以用输入和输出欧式距离来表示：

|  |  |
| --- | --- |
|  | (4-14) |

用均方误差(Mean Square Error, MSE)衡量该网络原始状态下的损失函数为：

|  |  |
| --- | --- |
|  | (4-15) |
|  |

其中为权重衰减项(Weight Decay)，用来防止网络边权值过大造成的过拟合现象，*λ*是权重衰减参数，用来平衡重构误差和权重衰减。表示网络H*l*中节点*i*和节点*j*的联接边权值。

隐含层*c*模拟的是输入文本的话题分布，而每个输入的句子通常都只是涵盖很少的话题，因此除了重构误差外，我们为这个网络另外加上隐藏神经元的稀疏限制，这样能够让隐含层学习到每个句子的侧重话题，而不是仅仅实现降维压缩。

稀疏限制的思路是抑制隐含层神经节点的激活值。在使用sigmoid函数作为激活函数的时候，隐层神经元的激活值接近1时认为该节点处于激活状态，而激活值接近于0的时候认为该节点处于抑制状态。因此稀疏性限制就是让隐藏层神经元尽量处于抑制状态，即激活值接近0。这种稀疏性限制从模拟人脑的角度分析，是因为神经元被激活是需要能量的，每次大脑皮层对输入文本的词素-语义分析时，只有小部分神经元工作，其他都处于被抑制状态。

用表示隐藏神经元*j*在输入样本*x*下的激活值，在训练集*X*=｛*x*1,*x*2..*x*n｝上隐藏神经元的平均激活度表示为：

|  |  |
| --- | --- |
|  | (4-16) |

为了实现隐藏神经元的稀疏限制，在我们的优化目标中加入一个额外惩罚因子。但是该惩罚不是简单的将隐藏神经元的每次激活值直接累加，因为理想中的激活值是个接近于0的一个很小的值，却又不完全等于0。因为所有点激活值为0训练出来的网络会导致连接边极大，导致重构效果极差。因此，我们定义一个稀疏性参数*ρ*。*ρ*通常是接近0的较小值，如0.05。通过鼓励隐藏神经元的激活值都尽量接近*ρ*，可以抑制隐藏神经元的激活程度。

因此稀疏限制的关键就是惩罚显著偏离*ρ*的情况。量化区别两个分布的指标有互信息、卡方检验、KL散度等，此处平均激活度和*ρ*都是离散序列，KL散度是个最合适的选择：

|  |  |
| --- | --- |
|  | (4-17) |

因此整个网络的稀疏惩罚用KL散度表示为：

|  |  |
| --- | --- |
|  | (4-18) |

其中*h*为隐层节点个数，在本文模型中*h* = 80。整个神经网络的损失函数包括重构误差以及稀疏惩罚：

|  |  |
| --- | --- |
|  | (4-19) |
|  |

而总损失函数对各层网络参数*W*和*b*的偏导数分别是：

|  |  |
| --- | --- |
|  | (4-20) |
|  |

其中是单个样本的重构损失的偏导值，需要用反向传播(Back Propagation, BP)算法计算得到。

为了方便计算，定义单个样本在网络中的流程如下：

\

图4-3 重构网络每个样本计算流程

其中*z*表示网络加权值，*a*表示激活值，*φ*()表示激活函数，在本文中是sigmoid函数。隐含层为第一次激活值，输出层*y*为第二次激活值。整个网络的计算过程简写为。

后向反馈算法的具体流程描述如下：

对每个输入样本，进行前向传播并计算网络中所有的激活值。然后对网络中第*l*层的每个节点*i*计算其传播残差，表示该节点占输出层残差的比重。其中输出层的残差表示的是激活值和预期值*x*的差距，可以用欧式距离计算：

|  |  |
| --- | --- |
|  | (4-21) |
|  |

其中*m*为输入层、输出层的节点个数，本文中为100。如果不考虑稀疏限制，隐藏层的残差为：

|  |  |
| --- | --- |
|  | (4-22) |
|  |

考虑了稀疏限制后，隐藏层的残差改为：

|  |  |
| --- | --- |
|  | (4-23) |

得到输出层和激活层的残差公式后，就可以计算单个样本重构损失分别对边权值*W*和偏置项*b*的导数：

|  |  |
| --- | --- |
|  | (4-24) |
|  |

其中*l*为0或1，分别计算是隐藏层和激活层关于*W*，*b*的偏导。

网络的训练过程，首先初始化网络参数(*W,b*)，将每个参数和赋值为接近0的极小值，可以通过正态分布随机生成()。随机初始化而不是全0初始化的目的是实现对称失效(Symmetry Breaking)，防止所有的边权重学习到相同值。初始化后，参数可以通过梯度下降法进行优化，由于损失函数非凸，可能得不到全局最优解，但是对于层数较浅的神经网络来说局部最优解也足够理想，也能实现文本重构功能。梯度下降算法每次迭代时，将文本的所有句子逐一输入进行训练，迭代到很小。每次迭代的具体过程：

1. 初始化梯度,
2. 对于*i* = 1：*n*

用BP算法计算偏导数和，并更新计算 和

1. 更新参数*W*,*b*

以及

其中*lr*是学习率。

梯度下降迭代训练完后，模型网络已经能够模拟原文重构的过程，并在隐藏层实现了话题分布的提取。

本文的摘要模型和3.3节中用于语义向量化的深度模型在结构上有几点区别：

1. 网络结构上：向量化模型用的多层神经网络，摘要模型是个浅层神经网络
2. 参数初始化：向量化模型需要用RBM预训练，防止梯度弥散；而摘要模型层数较浅，可以随机初始化网络参数
3. 隐含层：向量化模型的隐含层即编码层节点个数较少，目的是实现深度降维；而摘要模型的隐层节点数相对较多，但有稀疏限制，目的是实现话题提取
4. 损失函数：向量化模型的损失函数为输入输出的交叉熵，而摘要模型的损失函数为输入输出的重构误差加上稀疏惩罚
   * 1. 摘要提取

在线性摘要模型中，可以一次对多个句子进行组合重构，因此可以直接通过优化重构误差来直接获得最优句子集，作为生成的摘要。而神经网络每次只能处理一个句子，因此最理想的摘要策略是用该网络模型给每个句子以原文重构效果为标准，逐一打分并排序取top-k作为摘要。因此非线性重构策略的关键是定义一个量化单个句子重构原文效果的指标。

重构原文的效果体现在两个方面，一个是重构损失，一个是话题契合度。和训练时候一样，句子*s*的重构损失用欧式距离计算：

|  |  |
| --- | --- |
|  | (4-25) |

其中*d*是整个文档的语义向量，是句子*s*的语义向量在网络模型中的输出。话题契合度是指句子的话题分布和文档主旨的相似度，用相对熵计算：

|  |  |
| --- | --- |
|  | (4-26) |

其中表示输入句子*s*对应的第一层激活值，即隐藏层向量。*h*表示文档*d*对应的隐藏层向量。然后对文档所有句子求总损失：

|  |  |
| --- | --- |
|  | (4-27) |

对句子的排序值排序并取最小的*k*个句子作为生成的摘要。整个算法的流程：

|  |
| --- |
| 算法4.2：非线性重构生成摘要 |
| Input：文档的语义向量*d*，文档中所有句子的语义表示集合矩阵*S*，生成摘要的字数限制*l*  Output：摘要句子的集合*Summary*  Description：从文档中通过语义重构的思想找到非线性重构最好的句子集作为生成的摘要 |
| 1. 将*S*的所有列和*d*中心化为0-均值，另将*S*的所有列标准化为1-方差 2. 建立稀疏自编码神经网络，参数(*W*, *d*)随机初始化为 3. 以句子集{,,…}为样本进行BP训练，并用梯度下降法优化得到(*W, d*) 4. 文档*d*输入网络，得到隐藏层向量*h*为文档主题分布 5. for *i*=1:*n*      1. 将*score*数组中最小的*k*个句子输出为*summary* |

* 1. 冗余消减

能以最小重构损失还原文档的句子，被认为是与文档主题联系最紧密的句子，因此重构策略抽取出来的句子满足了凸显性这一指标。此外，本节通过加入冗余消减功能，来完善重构策略。

冗余就是句子之间的重叠信息。生成式摘要可以通过句子的融合重组,提炼出互不冗余的新句。然而对于抽取式摘要模型来说，冗余消减的实现非常困难。一方面，冗余度难以计算，以最经典的基于差异最大化(maximal diversity)的摘要算法MMR[5]为例：

|  |  |
| --- | --- |
|  | (4-28) |

其中*sim()*是文本相似度函数。*C*表示所有句子的集合，*S*是已经选中的句子集合，*是C*和*S*的差集即未选中句子集合。MMR算法的思想就是初始化*S*，然后每次挑选在剩余句子找到和集合*S*最不相关(最大边际相关)的句子并添加到*S*，反复迭代至达到长度限制，得到的集合*S*即为生成摘要。可以看出，冗余度往往要么在已选中的摘要句子集之间计算，或者通过待定的候选句和已选中的句子之间计算，前者通常是用于对摘要模型的改进，而后者依赖贪心模型逐句选择并迭代更新候选句集。

另一方面过度的冗余消减可能导致信息损失(Information Distortion)。比如下图出现的情况，三个句子,,,每个点表示一个话题，每个句子都涵盖了部分话题。如果选择句1和句2组成结果摘要，能够涵盖了更广的话题，却存在信息冗余；如果单选句3，则会导致信息损失。摘要算法必须在冗余消减和信息损失之间做出衡量，力求以最小的信息损失消减最多的冗余信息。

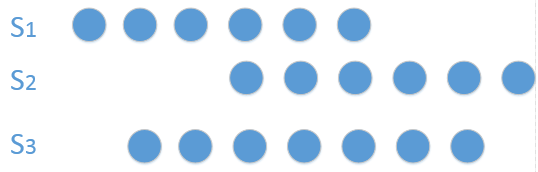


图4-4 对三个句子进行冗余消减,每个点表示一个话题

尤其在多个文本摘要任务中，主题衍生的分话题更广，各个话题高度相似，而重复率最高的信息往往是最重要的信息。消除所有的冗余信息很难通过算法直接实现。因此，本小节只是在一定程度上消除冗余信息并进行探讨，以完善重构策略。

在线性重构策略中，我们认为组合系数非零的句子参与了重构，而零系数对应的句子未参与重构。但在参与重构的句子中，负系数对应的句子可以理解为是被消减的部分，而正系数对应的句子才是真正提供话题参与重构的部分。因此，在公式4-12的基础上加上系数非负的限制，可以强制让所有的句子参与正重构进而减少冗余句子的产生：

|  |  |
| --- | --- |
|  | (4-29) |
|  |

系数非负限制的Lasso又叫Positive Lasso，可以通过修正的LARS算法求解。求得的最优化系数中的正项所对应的句子将被选为摘要句。

而在非线性重构策略中，摘要是通过选择重构度top-k的句子集合生成，在排序提取top-k的过程中，每个候选句和之前选中的句子都进行相似度对比，可以通过欧式距离或者夹角余弦计算。如果相似度大于预先设定的阈值，则认为该句子为冗余句，不被加入候选句集。

* 1. 实验对比
     1. 数据集和评测工具

本文采用的文本摘要数据集为DUC(Document Understanding Conference)数据集[[3]](#footnote-3)。DUC会议是美国国家标准与科学研究院(National Institute of Standards and Technology, NIST)自2001年起主持组织的，每年举办一次，自2008年起更名为文本分析会议(Text Analysis Conference, TAC)。DUC/TAC先后启动了单文档、多文档、面向查询、动态摘要以及更新式摘要等多个不同类型的摘要评测系统。每年都有超过20支队伍参加摘要系统比赛，而且会议组织者也一直在持之以恒的改进摘要评估的方法。DUC数据集已经成为自动文摘领域最普遍也是最权威的标准评测数据集。在本文中，采用了DUC会议在2006年和2007年的数据集，都是关于复杂话题的多文档摘要任务，分别包含50个和45个主题文档群。每个主题文档群各包含有25篇摘自纽约时报(New York Times)或美联社(Associated Press)等媒体的短新闻，同文档群里的短新闻都是围绕同一主题而报道，例如DUC2007的第6个文档群的25个新闻都是由“缅甸军政府延长昂山素季软禁期”这一事件延伸出来的，具体包括“素山昂季与诺贝尔奖和平奖”、“反对派示威者遭大规模逮捕”、“英美法等国家及人权组织纷纷谴责缅甸军政府”等相关新闻。每篇短新闻约有十几个句子共300多个单词。最终要从每个主题文档群生成一个250字以内的摘要，即3%左右的压缩率。此外每个主题文档群会提供四份由语言学专家给出的人工摘要作为评测参考。

自动摘要算法的评测指标一般是通过与人工文摘对比，在给定粒度文本单元的共现情况来计算。DUC/TAC对参赛队伍提交的摘要一般有4种不同的评测方式，分别是Basic Element(BE)、Pyramid、Rouge以及专家评分。其中BE系统的文本单元粒度是Basic Element，即名词动词等中心词以及中心词与修饰词之间的关系，通过对比自动摘要和专家摘要的Basic Element来给出评分。Pyramid评测系统的文本粒度是摘要内容单元(Summarization Content Unit, SCU)，即同一语义的词集，评测过程比较复杂费时。而Rouge是使用最方便最广泛的摘要评测系统，也是本文中所使用的评价方法。

ROUGE （Recall-Oriented Understudy for Gisting Evaluation）是由Lin[88]等人提出的自动摘要评测方法，原理是基于统计自动摘要和参考摘要之间的重复单元，如n-gram、次序列和词对。其中以n-gram为文本单元粒度的计算公式为：

|  |  |
| --- | --- |
|  | (4-30) |

其中*Ref*表示专家摘要，表示专家摘要中出现的n-gram个数，表示在专家摘要和测试摘要中共同出现的n-gram次数。除了以n-gram为文本粒度的Rouge-N之外，ROUGE系统还提供了其他一些常用评价指标：

1. Rouge-L: 以最长公共子序列(Longest Common Sequence, LCS)为评测单元粒度
2. Rouge-W：以加权LCS为评测单元粒度
3. Rouge-S：以skip-bigram为评测粒度，skip-bigram表示不相邻的两个词，例如"tomorrow is another day"中的"tomorrow"和"day"可以组成一对skip-bigram
4. Rouge-SU4：评测粒度结合skip-bigram和2-gram并且间隔距离不超过4

可以看出Rouge-N是个基于召回率(recall)的评测指标，但Rouge系统可以给出所有指标的precision、recall和F-measure分数。以Rouge-L为例：

|  |  |
| --- | --- |
|  | (4-31) |
|  |
|  |

其中*X*和*Y*分别是专家摘要和测评摘要，*m*和*n*分别为*X*和*Y*的长度。LCS(*X,Y*)表示*X*和*Y*的最长公共子序列长度。在DUC中，*β*是个很大的值，接近于无穷。因此F-score在数值上很接近于召回率recall。这样设置的原因是高质量的自动摘要的关键不仅要准更在于全。因为根据公式4-30,一个很短的摘要，比如只选择文档首句，往往能得到一个高得离谱的准确率值。因此在自动摘要研究中，使用Rouge测评系统的研究者们都会使用recall值或者f-score值做评分对比。

本文中采用DUC06和DUC07数据集，以Rouge-N、Rouge-L和Rouge-SU4作为评测指标,给出各个算法的F-measure值进行对比。

* + 1. 对比实验介绍

为了证明本文提出的语义重构模型的有效性，我们另外找了几个代表性的摘要算法进行实验对比。包括排序抽取模型SPSR(Self-Present Sentence Relevance)[89]，两个选择模型DSDR[80]和TopicDSDR[90]以及两个由NIST提供的Baseline：

* Baseline 1：NIST称之为简单参照(Simple Baseline)，对每个主题文档群Baseline1只是将每篇新闻的首句加入到摘要中，直到生成的摘要凑满250字。由于首句在文档尤其是新闻中的重要性，Baseline1的效果远好于随机选取句子作为摘要。
* Baseline 2: NIST在DUC2007评测系统中给出的泛式参考(Generic Baseline)。该基准是由一个名为CLASSY04的自动摘要系统生成，该系统是DUC04的优胜者，CLASSY04基于隐马尔科夫模型，用了5个状态表示隐含的摘要句或非摘要句子，用标志词元(Signature Tokens)作为观察序列的特征。
* SPSR：通过最小化文档结构的隐模糊相关性(Latent Implicit Relevance)来给句子打分，并选择top-k作为摘要
* DSDR：通过重构原文选择重构损失最小的句子集作为生成摘要，通过二次规划和泛函分析进行复杂运算
* TopicDSDR：DSDR的改进型，将DSDR中每个句子的向量表示由词袋模型改为由隐狄利克雷 (Latent Dirichlet Allocation，LDA)[1]主题模型生成的话题向量
  + 1. 实验结果及分析

各算法在DUC06和DUC07数据集的跑分结果分别在表中呈现，其中LR表示线性重构模型，NR表示非线性重构模型，-w表示使用了词嵌入加权的向量化方式，-d表示使用了深度降维的向量化方式。Rouge指标包括Rouge-1,Rouge2,Rouge3,Rouge-L以及Rouge-SU4的平均F-measure值。

表4.1(a) 各算法在DUC06上的平均F值，LR和NR为线性和非线性重构模型

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Method | Rouge-1 | Rouge-2 | Rouge-3 | Rouge-L | Rouge-SU4 |
| Baseline1 | 0.32082 | 0.05267 | 0.01372 | 0.29726 | 0.10408 |
| SPSR | 0.35383 | 0.05922 | 0.01555 | 0.31354 | - |
| DSDR | 0.33168 | 0.06047 | 0.01482 | 0.29850 | - |
| TopicDSDR | 0.37365 | 0.07073 | - | 0.34172 | **0.13190** |
| LR-w | 0.38365 | **0.07142** | **0.02001** | 0.34113 | 0.12944 |
| LR-d | 0.38463 | 0.06826 | 0.01494 | **0.34785** | 0.12831 |
| NR-w | 0.38368 | 0.06264 | 0.01435 | 0.33841 | 0.12384 |
| NR-d | **0.38477** | 0.06381 | 0.01442 | 0.34134 | 0.12621 |

表4.1(b) 各算法在DUC07上的平均F值，LR和NR为线性和非线性重构模型

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Method | Rouge-1 | Rouge-2 | Rouge-3 | Rouge-L | Rouge-SU4 |
| Baseline1 | 0.33475 | 0.06490 | 0.01856 | 0.31074 | 0.11278 |
| Baseline2 | 0.40059 | 0.09272 | 0.03058 | 0.36317 | 0.14467 |
| SPSR | 0.37070 | 0.06716 | 0.01844 | 0.32704 | - |
| DSDR | 0.39573 | 0.07439 | 0.01965 | 0.35335 | - |
| TopicDSDR | 0.39849 | 0.08200 | - | 0.36164 | 0.14562 |
| LR-w | **0.41431** | 0.09558 | 0.03106 | 0.37355 | 0.14861 |
| LR-d | 0.41124 | **0.09588** | **0.03152** | **0.38636** | **0.15061** |
| NR-w | 0.39982 | 0.09264 | 0.02935 | 0.33841 | 0.14384 |
| NR-d | 0.41008 | 0.09381 | 0.02942 | 0.34134 | 0.14621 |

其中粗黑字体的数字表示所有算法中表现最好的跑分。从实验结果可以看出本文的线性重构模型LR和非线性重构模型NR的表现都比较理想，除了DUC06数据集的Rouge-SU4指标在TopicDSDR算法的表现更好外，其他评测指标的平均F值都是最高的。除了LR和NR模型外，综合表现最好的是TopicDSDR和Baseline2，其中TopicDSDR是基于DSDR的重构算法，但是该算法用LDA模型生成的话题向量作为文本表示，提高了句子表示的语义性，因此较DSDR模型的表现提高不少；而Baseline2的CLASSY04算法是一个基于隐马尔科夫的监督模型，并得益于通过多次提交来获得表现反馈，大幅提高了摘要质量。而Baseline1和SPSR模型表现较差，因为Baseline1只是简单挑选每篇新闻首句并凑成摘要，虽然效果远好于随机抽取，但肯定差于任何一个精心设计的算法；而SPSR是个排序模型，给句子打分的时候孤立了句子之间的联系，而且该模型的思想是最小化隐文档相关性，但是由于文本表示的词袋向量很难掌握句子之间的相似度和相关性，导致算法效果并不理想。

为了具体观察每一篇新闻的表现，在DUC06数据集随机抽取了23个主题文档群，并对比线性重构模型LR和Classy04在每个主题上的ROUGE-1跑分。斜线上的点表示LR的表现更好，斜线下的点表示Classy04的表现更好。可以看出使用了词嵌入加权方式和深度降维方式向量化的LR模型的表现都比Classy04明显出色。

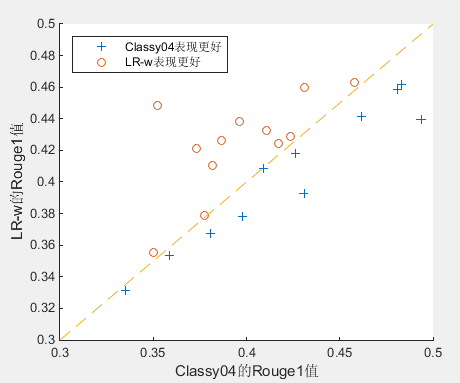


图4.5(a) LR-w模型和Classy04算法在DUC07的Rouge1值对比

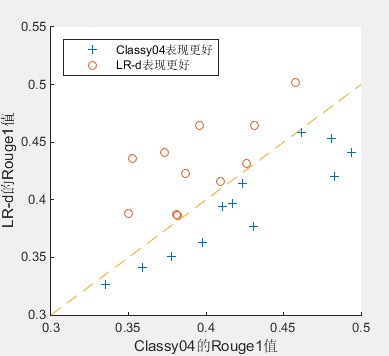


图4.5(b) LR-d模型和Classy04算法在DUC07的Rouge1值对比

通过对表4.1和表4.2中跑分数据横向对比本文提出的两种重构策略，可以发现NR模型除了DUC07中的Rouge1值比LR略高外，其他指标都不如LR。为了分析非线性重构模型表现不济的原因，在公式4-27中调整重构误差及话题契合度的平衡系数的值。

图4.6(a) NR模型在DUC06的Rouge1均值随的变化情况

图4.6(b) NR模型在DUC07的Rouge1均值随的变化情况

如图所示值越大，即重构误差项所占比重越大，Rouge值均趋于下降。可见损失函数公式4-27中只有话题契合度起作用，而考虑重构误差反而降低了摘要质量。因此能找到一个衡量重构误差并且能提高摘要质量的指标意义重大。

* 1. 本章小结

在第三章学习到文本的语义表示后，本章的目的就是建立重构模型，让生成的模型能够最小损失的重构原文语义。

重构模型的关键是高效的重构策略。本章中用两种重构策略，一种是线性的组合重构，另一种是非线性的重构。线性重构法的目的是找到线性重构原文的句子集，加上稀疏限制并辅以冗余消减；非线性重构相比线性函数更平滑也更灵活，实现方法是训练一个重构神经网络，计算出每个句子的重构效果值，进行排序抽取。此外，通过冗余消减能够在一定程度上优化摘要的结构性，完善了重构策略。

经过对比实验论证，两种重构策略的模型表现都非常出色，能够提取较高质量的自动摘要。

1. 总结与展望
   1. 工作总结

为了满足快节奏时代人们的快速阅读需求，自动摘要技术成为研究热点。为了解决主流模型中孤立句子联系和统计特征缺乏语义的缺点，本文提出的基于语义重构模型，即假设高质量的摘要能够以最小损失重构原文。

语义重构模型分为两步，分别是语义表示和重构策略。语义表示就是将文本、句子语义向量化，让句子之间的语义相似度能够通过语义向量的距离来衡量。本文使用了两种语义向量化方式：词嵌入加权和深度降维。词嵌入加权模型将文本向量化问题转为相对简单的词语向量化问题，词的语义向量化可以通过训练一个基于N-gram的神经网络语言模型来实现。而深度降维模型通过多层神经网络学习出输入文本的深度语义表示，深度网络使用了多个RBM预训练来避免随即初始化网络带来的缺点。此外，将输入文本的词概率向量进行离差标准化可以适应任何长度的输入文本。

重构策略是重构模型有效性的关键，本文分别尝试了线性重构策略和非线性重构策略。线性重构策略就是对所有句子进行线性组合，来还原原文的语义。当重构误差最小时，参与重构的句子集就是生成摘要。由于摘要的长度限制，重构误差要考虑重构系数的稀疏性惩罚，可以用L1正则项来表示。目标函数形式类似Lasso回归，可以通过LARS算法或者坐标下降法求最优解。而非线性重构策略是通过非线性激活的神经网络实现，该神经网络除了实现输入重构外，隐含层被加上了稀疏限制，因此能够提取输入文档的主题分布。通过对比每个句子和原文档的主题语义，能够计算每个句子的摘要贡献度，进而以此进行排序提取top-k。

此外，机器生成的自动摘要往往存在大量冗余信息。而对于抽取式摘要模型来说冗余信息一般很难去除，一方面以为冗余性计算条件的苛刻过于依赖贪心算法；另一方面冗余句子的剔除容易导致信息的损失，需要在两者之间平衡。在本文中，两种重构模型也分别在一定程度上试图消减冗余信息。在线性重构模型中，重构系数为非0对应的句子被认为参与了原文重构，但其中负系数所对应的句子在重构过程中相当于是被削减的句子，即被认为是冗余句。因此通过给目标函数加上组合系数非负的限制，可以大大减少冗余信息的产生。而在非线性重构模型中，针对生成摘要是对所有句子打分排序获得，可以设置一个相似度阈值，每次对当前候选句相似度判断，即和当前摘要集合所有句子进行相似度计算，如果大于阈值则认为是冗余句，否则加入摘要集合。通过一定程度上对冗余信息的消减，可以提高摘要质量，进一步改进了本文的语义重构模型的改进。

* 1. 未来展望

本文中的模型还存在一些不足，比如只是针对单文本摘要，没有考虑多文本摘要、基于用户摘要等；另一方面非线性模型的表现不尽如意，主要是损失函数中重构误差项反而降低了摘要质量。因此未来的工作将专注于两个方面：

(1)针对非线性重构策略中损失函数的表现不足，探索新的衡量重构误差的指标，既能够反映重构误差又能提高摘要质量。

(2)除了对单文本摘要算法的改进，另外探索多文本摘要、基于查询摘要甚至特殊领域摘要比如医学摘要。

参考文献

1. Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." the Journal of machine Learning research 3 (2003): 993-1022.
2. Erkan, Günes, and Dragomir R. Radev. "LexRank: Graph-based lexical centrality as salience in text summarization." Journal of Artificial Intelligence Research (2004): 457-479.
3. Luhn, Hans Peter. "The automatic creation of literature abstracts." IBM Journal of research and development 2.2 (1958): 159-165
4. Mihalcea, Rada, and Paul Tarau. "TextRank: Bringing order into texts." Association for Computational Linguistics, 2004.
5. Carbonell, Jaime, and Jade Goldstein. "The use of MMR, diversity-based reranking for reordering documents and producing summaries." Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1998.
6. Nenkova, Ani, Lucy Vanderwende, and Kathleen McKeown. "A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization." Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2006.
7. Baayen, R. Harald. Word frequency distributions. Vol. 18. Springer Science & Business Media, 2001.
8. Sparck Jones, Karen. "A statistical interpretation of term specificity and its application in retrieval." Journal of documentation 28.1 (1972): 11-21.
9. Salton, Gerard, and Christopher Buckley. "Term-weighting approaches in automatic text retrieval." Information processing & management 24.5 (1988): 513-523.
10. Dunning, Ted. "Accurate methods for the statistics of surprise and coincidence." Computational linguistics 19.1 (1993): 61-74.
11. Lin, Chin-Yew, and Eduard Hovy. "The automated acquisition of topic signatures for text summarization." Proceedings of the 18th conference on Computational linguistics-Volume 1. Association for Computational Linguistics, 2000.
12. Hovy, Eduard, and Chin-Yew Lin. "Automated text summarization and the SUMMARIST system." Proceedings of a workshop on held at Baltimore, Maryland: October 13-15, 1998. Association for Computational Linguistics, 1998.
13. Conroy, John M., et al. "Left-brain/right-brain multi-document summarization." *Proceedings of the Document Understanding Conference (DUC 2004)*. 2004.
14. Conroy, John M., Judith D. Schlesinger, and Dianne P. O'Leary. "Topic-focused multi-document summarization using an approximate oracle score."*Proceedings of the COLING/ACL on Main conference poster sessions*. Association for Computational Linguistics, 2006.
15. Finley, L., and H. Andrew. "Lite-GISTexter at DUC’2004." *Proceedings of the 2004 Document Understanding Conference (DUC’2004), Boston, MA*. 2004.
16. Gupta, Surabhi, Ani Nenkova, and Dan Jurafsky. "Measuring importance and query relevance in topic-focused multi-document summarization."*Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*. Association for Computational Linguistics, 2007.
17. Hatzivassiloglou, Vasileios, et al. "Simfinder: A flexible clustering tool for summarization." (2001).
18. McKeown, Kathleen, et al. "Towards multidocument summarization by reformulation: Progress and prospects." *AAAI/IAAI*. 1999.
19. Siddharthan, Advaith, Ani Nenkova, and Kathleen McKeown. "Syntactic simplification for improving content selection in multi-document summarization." *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics, 2004.
20. Mihalcea, Rada, and Paul Tarau. "A language independent algorithm for single and multiple document summarization." (2005).
21. Chali, Yllias, and Shafiq R. Joty. "Improving the performance of the random walk model for answering complex questions." *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*. Association for Computational Linguistics, 2008.
22. Kupiec, Julian, Jan Pedersen, and Francine Chen. "A trainable document summarizer." *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1995.
23. Paice, C. D. "Constructing literature abstracts by computer." *Information Processing & Management* 26(1990):171-186.
24. Edmundson, H. P. "New Methods in Automatic Extracting.." *Journal of the Acm* 16.2(1969):264-285.
25. John M. Conroy, and Dianne P. O'leary. "Text summarization via hidden Markov models." *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*2001:406-407.
26. Ulrich, Jan, et al. "A publicly available annotated corpus for supervised email summarization." *Proc of Aaai Email Workshop* (2011).
27. Barzilay, Regina, and N. Elhadad. "Sentence Alignment for Monolingual Comparable Corpora." *Conference on Empirical Methods in Natural Language Processing* 2010:25--32.
28. Iii, Hal Daumé, and D. Marcu. "A Phrase-Based HMM Approach to Document/Abstract Alignment.." *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing , EMNLP 2004, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004, 25-26 July 2004, Barcelona, Spain* 2004:119-126.
29. Jing, Hongyan. "Using Hidden Markov Modeling to Decompose Human-Written Summaries." *Computational Linguistics* 28.4(2002):527-543.
30. Marcu, Daniel. "The automatic construction of large-scale corpora for summarization research." *University of California, Berkely* 2010:137-144.
31. Zhou, Liang, and E. Hovy. "A Web-Trained Extraction Summarization System.." *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1* 2003:págs. 331-336.
32. Chali, Yllias, S. A. Hasan, and S. R. Joty. "Do automatic annotation techniques have any impact on supervised complex question answering?." *Acl-Ijcnlp 2009 Conference Short Papers* Association for Computational Linguistics, 2009:329-332.
33. Barzilay, Regina, and M. Elhadad. "Using Lexical Chains for Text Summarization." *In Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization* 2010:10--17.
34. Galley, Michel, and K. R. Mckeown. "Lexicalized Markov Grammars for Sentence Compression.." *Proceedings of NAACL-HLT* (2007):180-187.
35. Silber, H. Gregory, and K. F. Mccoy. "Efficiently Computed Lexical Chains as an Intermediate Representation for Automatic Text Summarization." *Computational Linguistics* 28.4(2002):487-496.
36. Ed. "Introduction to WordNet: An On-line Lexical Database." *Fellbaum*3.4(1998):235--244.
37. Galley, Michel, and K. Mckeown. "Improving Word Sense Disambiguation in Lexical Chaining.." *Ijcai-03, Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, Acapulco, Mexico, August* 2003:1486-1488.
38. Schiffman, Barry, A. Nenkova, and K. Mckeown. "Experiments in multidocument summarization." *International Conference on Human Language Technology Research* Morgan Kaufmann Publishers Inc., 2002.
39. Scott, Deerwester, et al. "Indexing by latent semantic analysis."*Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence* Morgan Kaufmann Publishers Inc., 1999:391-407.
40. Gong, Yihong, and X. Liu. "Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis." *SIGIR '01: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval* 2001:19-25.
41. Hachey, Ben, G. Murray, and D. Reitter. "Dimensionality Reduction Aids Term Co-Occurrence Based Multi-Document Summarization." *ACL Workshop on Task-Focused Summarization and Question Answering*2006:1-7.
42. Baldwin, B., & Morton, T. S. (1998, June). Dynamic Coreference-Based Summarization. In *EMNLP* (pp. 1-6).
43. Boguraev, B., and C. Kennedy. "Saliencebased Content Characterization of Text Documents." *Proceedings of the Acl/eacl’97 Workshop on Intelligent Scalable Text Summarization* (1997):2--9.
44. Steinberger, Josef, et al. "Two uses of anaphora resolution in summarization." *Biotechnology & Bioengineering* 28.7(1986):1086-1092.
45. Bateman, J., and J. Delin. "Rhetorical Structure Theory." *Encyclopedia of Language & Linguistics* (2006):589-597.
46. Marcu, Daniel. "From Discourse Structures to Text Summaries."*Proceedings of the Acl Workshop on Intelligent Scalable Text Summarization* (2002):82--88.
47. Marcu, Daniel. "To Build Text Summaries of High Quality, Nuclearity is Not Sufficient." *In AAAI-98 Spring Symposium on Intelligent Text Summarization* (1998).
48. Marcu, Daniel. "The Theory and Practice of Discourse Parsing and Summarization." *Computational Linguistics* 28.1(2002):81-83.
49. Elhadad, N., et al. "Customization in a unified framework for summarizing medical literature." *Artificial Intelligence in Medicine*33.2(2005):179–198.
50. Kan, Min Yen. "Combining Visual Layout and Lexical Cohesion Features for Text Segmentation." *In Proceedings of the 31stWorkshop on Graph Theoretic Concepts in Computer Science- WG 2005* (2001):187-198.
51. Elhadad, N, et al. "Facilitating physicians' access to information via tailored text summarization.." *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium* 2005(2005):226-30.
52. Mckeown, Kathleen R., et al. "PERSIVAL, a System for Personalized Search and Summarization over Multimedia Healthcare Information.."*Proceedings of the First Acm+ieee Jcdl* 37.2(2001):331--340.
53. Teufel, Simone, and M. Moens. "Summarizing scientifi articles – experiments with relevance and rhetorical status." *Computational Linguistics* 28.4(2001):págs. 409-446.
54. Nanba, Hidetsugu, and M. Okumura. "Towards Multi-paper Summarization Using Reference Information." *In Proceedings of IJCAI-99* 98.82(1999):926--931.
55. Mei, Qiaozhu, and C. X. Zhai. "Generating Impact-Based Summaries for Scientific Literature.." *ACL 2008, Proceedings of the Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, Usa* 2008:816-824.
56. Qazvinian, Vahed, and D. R. Radev. "Scientific Paper Summarization Using Citation Summary Networks." *International Conference on Computational Linguistics* Association for Computational Linguistics, 2008:689-696.
57. Muresan, Smaranda, Tzoukermann, Evelyne, and Klavans, Judith L. "Combining linguistic and machine learning techniques for email summarization." *The Workshop on Computational Natural Language Learning* Association for Computational Linguistics, 2001:1--8.
58. Tzoukermann, Evelyne, S. Muresan, and J. L. Klavans. "GIST-IT: summarizing email using linguistic knowledge and machine learning."*The Workshop on Human Language Technology & Knowledge Management* Association for Computational Linguistics, 2001.
59. Corston-Oliver, Simon, et al. "Task-focused summarization of email."*Proceedings of the Text Summarization Branches Out Acl Workshop*(2004).
60. Rohall, Steven L., et al. "Exploiting E-mail Structure to Improve Summarization." *Massachusetts Institute of Technology* (2002).
61. Boguraev, B. K., and M. S. Neff. "Discourse segmentation in aid of document summarization." *Hawaii International Conference on System Sciences* IEEE Computer Society, 2000:3004-3004.
62. Nenkova, Ani, and A. Bagga. "Facilitating email thread access by extractive summary generation.." *In Proc. of RANLP’03* (2003):287-296.
63. Rambow, Owen, et al. "Summarizing email threads." *Proceedings of HLT-NAACL 2004: Short Papers* Association for Computational Linguistics, 2004.
64. Newman, Paula S., and J. C. Blitzer. "Summarizing Archived Discussions: A Beginning." *Intelligent User Interfaces* 2003:273-276.
65. Berger, Adam L., and Vibhu O. Mittal. "OCELOT: a system for summarizing Web pages." *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2000.
66. Buyukkokten, Orkut, et al. "Efficient web browsing on handheld devices using page and form summarization.." *Acm Transactions on Information Systems* 20.1(2002):82-115.
67. Delort, J. Y., B. Bouchon-Meunier, and M. Rifqi. "Enhanced web document summarization using hyperlinks." *Fourteenth Acm Conference on Hypertext & Hypermedia* 2003:208 - 215.
68. Sun, Jian Tao, et al. "Web-page summarization using clickthrough data."*SIGIR 2005: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil, August 15-19, 2005* 2005:194-201.
69. Choi, Yejin, et al. "Using landing pages for sponsored search ad selection.." *WWW 10: International Conference on World Wide Web*2010:251-260.
70. Zhou, Liang, and Eduard H. Hovy. "On the Summarization of Dynamically Introduced Information: Online Discussions and Blogs." *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*. 2006.
71. Hu, Meishan, A. Sun, and E. P. Lim. "Comments-oriented blog summarization by sentence extraction." *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management* ACM, 2007:901-904.
72. Le, Quoc V., and T. Mikolov. "Distributed Representations of Sentences and Documents." *Eprint Arxiv* 4(2014):1188-1196.
73. Sutskever, Ilya, O. Vinyals, and Q. V. Le. "Sequence to Sequence Learning with Neural Networks." *Advances in Neural Information Processing Systems* 4(2014):3104-3112.
74. Bengio, Yoshua, et al. "A neural probabilistic language model." *Journal of Machine Learning Research* 3.6(2003):1137-1155.s
75. Tai, Sing Lee, and D. Mumford. "Hierarchical Bayesian inference in the visual." *Journal of the Optical Society of America A Optics Image Science & Vision* 20.7(2003):1434-1448.
76. Leuba, G., and R. Kraftsik. "Changes in volume, surface estimate, three-dimensional shape and total number of neurons of the human primary visual cortex from midgestation until old age.." *Anatomy & Embryology*190.4(2012):65-71.
77. Hinton, G E,, and R. R. Salakhutdinov. "Reducing the dimensionality of data with neural networks.." *Science* 313.5786(2015):504-507.
78. Salakhutdinov, Ruslan, and Geoffrey Hinton. "Semantic hashing."*International Journal of Approximate Reasoning* 50.7 (2009): 969-978.
79. Laurens, Van Der Maaten,, and G. Hinton. "Viualizing data using t-SNE."*Journal of Machine Learning Research* 9.2605(2008):2579-2605.
80. He, Z., et al. "Document summarization based on data reconstruction."*AAAI Conference on Artificial Intelligence* 2012.
81. Zhong, Sheng Hua, et al. "Query-oriented unsupervised multi-document summarization via deep learning model." *Expert Systems with Applications* 42.21(2015):8146-8155.
82. Schmidt, Mark. "Least squares optimization with l1-norm regularization."*CS542B Project Report* (2005): 14-18.
83. Tibshirani, Robert. "Regression Shrinkage and Selection via the Lasso."*J of the Royal Statistical Society* 58.1(1996):267-288.
84. Chen, Scott Shaobing, D. L. Donoho, and M. A. Saunders. "Atomic Decomposition by Basis Pursuit." *Siam Review* 43.1(2001):33-61.
85. Hui, Zou, and H. Trevor. "Regularization and variable selection via the elastic net." *Journal of the Royal Statistical Society* 67.2(2005):301-320.
86. Efron, Bradley, et al. "Least angle regression." *The Annals of statistics* 32.2 (2004): 407-499.
87. Friedman, J, T. Hastie, and R. Tibshirani. "Regularization Paths for Generalized Linear Models via Coordinate Descent.." *Journal of Statistical Software* 33.i01(2010):: 1–22.
88. Carlos Flick. "ROUGE: A Package for Automatic Evaluation of summaries." *The Workshop on Text Summarization Branches Out*2004:25--26.
89. Li, Xiaodong, et al. "Document Summarization via Self-Present Sentence Relevance Model." *DASFAA* 2013:309-323.
90. Zhang, Zhiming, H. Li, and L. Huang. *TopicDSDR: Combining Topic Decomposition and Data Reconstruction for Summarization*. *Web-Age Information Management*. Springer Berlin Heidelberg, 2013:338-350.

致谢

春去秋来、时光匆匆，不知不觉中三年的研究生生活即将划上句号。这三年里，过得紧张却又充实，是人生中重要又难忘的三年。

首先感谢我的导师王崇骏教授和吴骏老师，在三年的研究生生活中给了我莫大的指导和帮助，在我科研迷茫的时候指点我，在我就业不顺的时候安慰我，他们是我在学生时期更是今后人生的指路人。

同时要感谢我的家人父母，他们永远是我最坚强的后盾，在我经济困难以及遇到挫折的时候毫不犹豫、全心全意的帮助我，让我顺利完成学业。

感谢实验室的小伙伴们在最后的三个月里陪我在一起查资料、做实验、憋论文，他们有刘勇、王涛、陈厚兵、戴恒宇、肖雨奇、李红、王茜和韩军华等，也希望他们能在今后的工作中大展宏图。并感谢我在509-1的舍友许涵斌和赵斐，在生活上给了我很大的帮助，一起苦中作乐、其乐融融，苟富贵必不相忘。

附录

研究生期间论文发表

1. Zhang Chi, et al. "Text Summarization Based on Sentence Selection with Semantic Representation." *IEEE International Conference on Tools with Artificial Intelligence* IEEE, 2014:584-590.

研究生期间参与项目

1. 江苏银行贷后风险网络预警系统

附件二

《学位论文出版授权书》

本人完全同意《中国优秀博硕士学位论文全文数据库出版章程》（以下简称“章程” ），愿意将本人的学位论文提交“中国学术期刊（光盘版）电子杂志社”在《中国博士学位论文全文数据库》、《中国优秀硕士学位论文全文数据库》中全文发表。《中国博士学位论文全文数据库》、《中国优秀硕士学位论文全文数据库》可以以电子、网络及其他数字媒体形式公开出版，并同意编入《中国知识资源总库》，在《中国博硕士学位论文评价数据库》中使用和在互联网上传播，同意按“章程”规定享受相关权益。

作者签名： 张弛

2016 年 5 月 27 日

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 论文题名 | 基于语义重构的文本摘要算法 | | | | |
| 研究生学号 | Mg1333075 | 所在院系 | 计算机科学与技术 | 学位年度 |  |
| 论文级别 | □硕士 □硕士专业学位  □博士 □博士专业学位  (请在方框内画钩) | | | | |
| 作者Email | 870727617@qq.com | | | | |
| 导师姓名 | 王崇骏教授、吴骏讲师 | | | | |

论文涉密情况：

□ 不保密

□ 保密，保密期（ 年 月 日 至 年 月 日）

注：请将该授权书填写后装订在学位论文最后一页（南大封面）。

1. <http://www.worldwidewebsize.com/> [↑](#footnote-ref-1)
2. <http://www.dmoz.org> [↑](#footnote-ref-2)
3. <http://www.nist.gov/index.html> [↑](#footnote-ref-3)