面向新媒体的短文本自动摘要系统研究与实现

# 第一章 绪论

## 研究背景

目前，互联网正处于高速发展时期，网络信息也出现了爆炸式增长的情况。随着各种移动终端设备的流行，人们逐渐摄弃了纯手工文本，开始频繁使用电子文本进行工作、记录和交流等，z尤其依赖于从网络上获取有效的信息资源。然而，网络环境中的信息资源数量极其庞大，杂乱无章，如何使得人们可快速、准确地从网络上获取有用的信息，是目前正在被积极研究的一个热点领域。自动文本摘要(简称自动文摘)技术作为该研巧领域的一个重要分支，可从海量文本中提取出简短的关键信息来帮助人们快速理解原来的文本内容，有效节省了人们获取有用信息的时间和精力。近年来，自动文摘系统作为一个有力的辅助工具，在各个领域发挥了作用，包括Web信息检索、新闻、邮件和会议记录等，成功地帮助了人们更加快速高效的方式获取了他们想要的信息。

自动文摘技术是自然语言处理领域重要的一个分支，初期主要广泛应用于长文本摘要任务中，例如科技论文、新闻领域等。然而随着微博、Twitter等短文本形式的社交媒体的爆炸式增长和今日头条、公众号等原创内容平台的崛起，针对短文本的自动摘要任务实现也变得更为迫切。与长文本的严谨有序不同的是，新媒体中的大多数内容都是由用户自发创造的，文本篇幅短小，内容碎片化严重，同时数据量巨大。由于用户在创造内容时几乎不受约束，语言组织形式多样化，用语随意不够规范，给识别和理解带来了很大的困难。因此，深入研究以短文本为目标数据集的自动摘要，可进一步促进自动文摘技术的发展，拓宽自动文摘技术的应用领域。

## 研究意义

摘要是对文本的高度概括总结，反映了文本的核心内容，与文本的标题标题相比，摘要的信息量更大，更具有代表性[1]。自动摘要是用户通过计算机自动地提取出文本的中心内容然后一并输出的过程[2]。它通过对信息的深度挖掘分析与浓缩，为人们日常的学习生活提供便捷，有助于用户快速高效地获取文本的核心思想，节约用户的宝贵时间和精力。由此可知，对文本自动摘要的研究是很有必要的，同时还具有很多的实际应用意义：

（1）**提高阅读速度。**我们浏览某文本时，摘要是其核心主旨的最佳体现，通过阅读摘要就可以全面理解原文，从而了解此文本是否是自己感兴趣的内容，进而决定是否继续看其中细节，节约阅读时间。

（2）**获取全面客观信息。**在通过标题和摘要快速获取过程中，我们可能受到一些为吸引关注度不客观标题的误导。同时，由于人工摘要的主观性，会遗漏一些关键性信息，而自动摘要可以避免主观偏见，帮助读者获得客观全面的信息。

（3）**节约人工编制摘要的成本，避免大量人力的耗费。**现如今无论是公众号的新闻推送，还是内容分发平台的新闻推荐。其标题和摘要部分都是人工进行编辑，不仅人力成本巨大，而且可能会因为编辑水平不足影响摘要的质量，从而影响媒体的声誉和受关注度。

# 第二章 国内外现状

## 2.1研究现状

文本自动摘要的相关技术研究从被关注一直延续到现在，尤其是网络时代的信息大爆炸对人们的生活学习影响逐渐加深，国内外越来越多的学者投身到相关领域的研究之中。历经半个多世纪，此项研究已经获得了很大的进展，并且在人工智能与语言学等领域开始有相对广泛的应用。

**2.2.1 国外研究现状**

在十九世纪50年代，海外的文本自动摘要技术已经起步，并获得了丰硕的研究成果。IBM公司的Luhn于1958年最先提出一种基于高频率词语给句子排序得到摘要句的算法，并将其发表到论文“The Automatic Creation of Literature Abstracts”内，便拉开了各式各样发展自动摘要研究的序幕。美国的Edmudson考虑句子位置和线索词等因素特征来获得句子的权重，每一个句子的重要程度均需要计算该句子的各个权重相加所获得之和，而这些句子权重的高低便成为摘要句输出的根据，相比只考虑高频词的方法，效果有所提升。LinC-Y全面考虑句子长度、句子位置、句子同文本标题的相似度等一系列特征来提高所获得摘要句的质量，充分挖掘了文本特征。Sharifi B，Huttion M和Sharifi B，INOUYE D，研究了词语加强（Phrase Reinforement，PR）和Hybrid TF-IDF在提取方法内的作用，取得了不错的应用效果。

以上都是以统计特征为研究中心的文本自动摘要技术，这些方法简单直接，但是获得的文本摘要的通顺度比较低并且具有很高的冗余。后来，为了使获得的摘要质量得到更进一步地提升改善，不同的研究者开始把知识理解和文档的篇章结构融合到文本自动摘要方法中，并且获得了良好的效果。

Morris和Hirst首先研究出了一个关于词汇链的模型并且通过模型的计算过程可以使能够构成同一主题的关系密切而且相邻的词形成一个系列，此模型对主题分割以及解析文档的结构十分有帮助。不久，Barzilay等人提出基于WordNet词汇链的多种不同模型来获得摘要句。词汇链能够将词汇间语义关系清晰地表示出来，为主题的划分和文档结构的分析提供重要的根据，从而提升文本自动摘要的质量。Gunes Erkan进行关于Lex Rank的文本自动摘要相关研究，将其内容结构用图来描述，之后获得文本摘要句的过程主要就是运用基于图的算法计算得到句子的权重，最后通过使用 DUC2004 数据集对该算法系统实施评估，证明该方法具有很好的效果。Tadashi Nomoto巧妙地运用与语料库相关的资源，通过解决使用决策树结构将句子划分为不同类别这一问题来进行摘要句的挑选，这样就能很方便地使用人工智能相关领域的知识和方法。

还有很多学者专门研究某一领域知识，结合语法、语义及语用研究成果，开发专门的应用系统，在特有的知识领域内取得了突出的研究成果。美国的GE 研究中心提出了SCSIOR系统，其在运用知识理解的基础上，利用关键词过滤以及模式匹配方法来分析待考察文本主题；然后使用与领域无关的自底向上的分析器来解析每个句子的结构特征,最后采用自顶向下的分析器得到摘要句。韩国Ulsan大学研发的基于用户请求的ROSE系统，能够根据读者的要求很好地生成读者需要的文本摘要。Munesh Chandra在提取摘要时使用k-mixture概率模型计算句子中所含名词和动词的词语权重，利用名词和动词的权重来衡量句子的语义重要性，根据一定的比例选取语义重要性排名前N的句子组合在一起作为摘要。

**2.2.2 国内研究现状**

由于中文与英文的语法构成及表现形式有着很大差异性，致使与中文相关

的信息处理研究变得相当复杂，所以国内的文本自动摘要技术起步相对较晚。

随着国内大量专家学者不断地投入人力、物力对其进行深入研究，目前中文文

本自动摘要技术获得了很大的进展。

上海交通大学的王永成教授带领实验室科研小组人员开始研究中文文本自动摘要技术，综合考虑句子位置、指示性短语、文本结构等因素成功地研发出了SJTUCAA-汉语文本自动摘要系统。之后，通过深入研究分析，他们又实

现了 OA 汉语文档自动摘要系统并且具有一定实用价值。

北京大学的李素建教授一直致力于文本内容的基础研究，认为只有从词语层到篇章层的详细解析才能找到适合文本内容处理的正确方法，并且在文本结构划分方面做不断尝试，此方法在文本自动摘要领域得出了较好的结果。

哈尔滨工业大学的王开铸教授的实验室团队首先研发了MATAS文本摘要系统，其主要是在自然语言理解相关技术上实现的，后来还继续开发了一个基于 HIT-863 I 的相关实用系统并获得了不错的研究成果。

北京邮电大学的钟义信教授实现了一种依赖涉及全信息的相关技术、对语义进行全方位划分最终重新组配成不同模块、接着运用语用知识制导的最新想法，并将其与语句理解以及摘要提取一起使用便成功地研制出了LADIES系统，在神经网络学习领域表现出了不错的效果。

# 第三章 关键技术

## 3.1主流方法

## 3.1.1 抽取式方法

抽取式摘要是从原始文本中抽取一系列可以表达原文核也思想的文本片段(关键短语、句子、段落等)，按照一定顺序排列组合而成。这些文本片段基本保持原貌，不需要修改，其表现形式通常为句子。所以，抽取式摘要任务的关键问题在于如何找出原始文档中那些适合做摘要的句子。目前的抽取式摘要方法主要以原文中的句子作为摘要单位进行抽取，因为它易于实现，同时能够完整的描述一个重要的信息点，又化段落含有更少的冗余信息，保证了生成的摘要具有相对较好的可读性。该类方法一般包括下面这几个步骤。

**文本预处理。**自然语言处理任务的标准流程中第一步都是预处理，是需要将自然语言描述的文本转化为计算机可以识别的特征的过程。在抽象式摘要任务中，首先对原文进行分词处理，得到分割后的词集，经停用词过滤之后获得文本的关键词集合，于是得到了代表该原始文本的特征元数据。

**文本表示。**对文本进行形式化处理，将其转化为计算机容易识别的格式内容。首先从文本中抽取出词或短语组成特征项，并依据其在文本中的重要性给出权重，由此便可得到文本信息的量化了的数学向量表示。文本的表示及其特征项的选取是自然语言处理任务的一个最基本问题。目前特征词的权重计算方法主要有布尔型、词频统计和TF-IDF，传统的文本表示模型则包括布尔模型、向量空间模型和概率模型，其中向量空间模型是自然语言处理中常用的主流模型。随着深度学习技术的发展，出现了以Word2Vec为代表的文本深度表示模型，可以为文本数据寻求更加深层次的特征表示。

**句子分析与抽取。**这一步骤是抽取式摘要任务的核也所在，通过对句子相似度的计算，进而排除掉文本集合中的冗余信息，再按照一定策略从句子集合中抽取出合适的句子以供生成摘要使用。句子相似度的计算最受欢迎的方法当属余弦相似度(cosine similarity)，经常与TF-IDF结合应用于分类、聚类、推荐、查询结果等许多问题当中，都有着良好的表现。而如何抽取摘要句子，是最为关键的问题。该问题通常可分为两种方式解决：一种是对文本集合中的所有句子按照特征组合的方式进行打分并予以排序，然后将分数较高的句子抽取出来作为摘要句子；另一种是先把集合中的句子经过分类处理后分到不同的子集合中，然后从各子集会中抽取合适的句子作为摘要句子。由于单文挡的内容有限，后一种方法用在单文档中并不能发挥其优势，因此多用于多文档摘要任务中。而排序之后的结果只是考虑了相关性，并没有考虑新颖性，排名靠前的几个句子表达的意思可能是相似的。因此需要做进一步处理，通常使用的是MMR算法。

**摘要句排序。**针对上一步抽取出来的摘要句子，按照一定顺序排列组合起来，使得提供给用户的最终摘要是逻辑合理和可读性良好的。对于单文档摘要，只需要按照摘要句子在原始文本中的相对位置顺序进行排序，前后连接形成摘要即可。然而多文档摘要任务中，摘要句子可能来自不同的文档，也就不便参考位置信息。此时，一般考虑依据摘要句子的内容逻辑和时间逻揖等深层文本信息来进行排序。

以上便是抽取式摘要方法通常使用的流程，研究人员在各流程环节进行创新或改进，孕育出丰硕的研究成果。已有的大多数自动摘要研究是基于抽取式方法实现的，可以将其中的一些重点方法做如下归类：

(1)**基于特征统计(feature and statistics)的方法。**特征工程在深度学习流行之前一度被视为解决特定领域问题的良药，自动摘要任务中用到的特征一般包括词频、句子长度、句子位置、句子是否包含标题词，以及句子关键词打分等。该类方法首先统计词频和位置等信息，然后计算句子权重，继而选择权重高的句子作为摘要。基于该方法的代表性系统有SUMMARIST和TextTeaser。该类方法虽然简单易用，但是其核仅仅是建立在文本表层的形式特征，缺乏深度，因此发展潜力受限。

(2)**基于词汇链(lexical chain)的方法。**词汇链是指可以描述同一个主题的相关词汇共同组成的词汇系列，最初用于文本分割问题。该类方法首先选择出候选词并根据与各词汇链的相关程度将其分配到合适的词汇链中，然后按照某种规则计算各词汇链权重，进而挑选出部分词汇链生成摘要。该方法的核屯、在于词汇语义相关性的度量。

(3)**基于图模型(graph theory)的方法。**该类方法的一般思想为：将文档的每句话作为节点、句子之间的相似度作为边的权重来构建图模型，再使用图排序算法(PageRank等)得到句子的得分，以生成摘要。代表算法TextRank、LexRank和HyperSum。

(4)**基于潜在语义分析(latent semantic analysis，LSA)的方法。**潜在语义分析使用奇异值分解(singular value decomposition，SVD)方法将高维的词-文档特征表示映射到低维空间中，解决了不同词语间的同义现象，进而发现文本背后的语义结构。该类方法通过使巧各种主题模型来挖掘词句之中的隐藏语义结构，继而从潜在主题中抽取摘要句子。经常被使用于摘要任务中的主题模型有LDA(latent Dirichlet allocation)和HMM等。该类方法相比上述方法得到的摘要质量更好，然而对文本结构的要求较高，使得其应用巧围有限。

(5)**基于聚类的方法。**该类方法多用于多文档摘要任务，通过聚类来识别出文档集合中的共同主题，并从每个簇中抽取合适的句子组成摘要。

## 3.1.2 生成式方法

生成式摘要方法是当下应用最广泛、实现最简单的自动摘要方法，能胜任对海量异构的网络文本的各种处理要求。但是由于忽略了对文本结构和语言特征的分析和理解，生成的摘要的全面性、简洁性和连贯性很难得到保证，导致可读性较差，同时也不可避免地存在兀余问题。而生成式摘要方法则需要利用自然语言理解技术对文本进行语法、语义分析，对信息进行融合，使用自然语言生成技术生成新的摘要句子，由此能够很好地弥补上述缺陷。但也正因为如此，生成式摘要方法需要对文本进行更为深入的分析，使得该任务的研巧变得困难且富有挑战性。然而随着自然语言处理技术的深入发展以及深度学习研究的日趋成熟，生成式的摘要方法受到了广泛关注，并在领域内的国际重要会议和期刊上长期占有一席之地，取得了丰富的研究成果。

生成式摘要主要通过对原始文本进行语义理解，将其表示为深层语义形式(例如深层语义图)，然后分析获得摘要的深层语义表示(例如深层语义子图)，最后由摘要的深层语义表示生成摘要文本。己有的研巧可以做如下分类：

(1)**基于结构的方法。**该类方法依靠认知模式从原始文档中识别和提取最重要的信息，然后应用语言生成器或巧应算法来生成摘要。常用的认知模式有模版、提取规则和树等其他结构。

(2)**基于语义的方法。**在此类方法中，原始文档的语义表示作为自然语言生成系统的输入以形成摘要。该方法主要通过使用自然语言处理技术，从语言学角度来处理文本进而识别出名词短语和动词短语。已有的方法包括基于多模式语义模型的方法、基于信息项的方法基于语义图结构的方法。其中基于多模式语义模型的摘要是面向多模式文档的，即文档中包含多种表现形式，比如文字和图片，因此不做讨论。而在基于信息项的方法中，摘要的内容是由原始文档的抽象表示生成而非原始文档的句子，其中抽象表示，即信息项，是文本中表达连贯信息的最小元素。基于语义图的方法旨在为原始文档创建一个富语义图(rich semantic graph，RSG)，通过使用启发式规则精简生成的富语义函，从中生成准确连贯且兀余较少的抽象式摘要。最有代表性的是基于AMR(abstract meaning representation)语义图的摘要方法，它将原始文本表示为AMR图集合，精简处理后转换为摘要图结构，从而实现由语义表达直接生成摘要。

(3)**端到端方法。**深度学习技术在各领域任务中表现突出，应用到自动摘要任务也取得了不错的成果。目前最为流行的是基于Encoder-Decoder框架的seq2seq模型，该模型不用人工提取特征，也避免了内容选择和句子重要性评估等处理步骤，只需要有足够的语料集作为输入输出进行训练即可。然而，相比传统自动摘要方法，端到端方法需要很大规模的训练语料，另外当前的主流神经网络框架不能够对长文本进行有效的语义编码，因此当前的有关研究大多只能针对句子简化(sentencesimplification)和标题生成这样的短文本摘要任务。近期工作开始尝试对长文本进行端到端摘要研巧并相应地引入注意力机制和拷贝机制， 取得了不错的效果。然而对于多文档摘要任务，主要还是依靠抽取式摘要方法来实现，从而导致生成的摘要存在着可读性差、冗余较难控制等问题。

## 3.2 数据与测评指标

## 3.2.1 数据集

当下自动摘要研究中常用的数据集包括两种，一种是用来训练深度学习模型的大规模数据集，一种是用来参与评测的小规模数据集。应用最广泛的当属DUC，其网站自2001年开始提供自动文本摘要的比赛和测评，并于2008年转移到TAC网站。这里提供的数据集都是小规模数据集，用来评测模型。Gigaword数据集非常庞大，大概有950w篇新闻文章，数据集用新闻标题作为摘要，即输出文本，用文章首句作为输入文本，属于单句摘要的数据集。CNN/Daily Mail数据集是在机器阅读理解中用到的语料，近来也被应用于摘要任务中，属于多句摘要(长文本摘要)的数据集。LCSTS是一个中文短文本摘要数据集，数据采集自新浪微博。SogouCS数据来自搜狐新闻，包含2012年6月—7月期间的新闻数据，超过1M的语料数据，包含新闻标题和正文的信息。

## 3.2.2 测评方法

自动文本摘要的评测历来是自然语言处理中比较棘手的问题，然而却是非常重要的一部分。在自动摘要漫长的成长历程中，研究者提出了许多相对可行的评测方法。自动文本摘要评测方法可分为内部评测方法和外部评测方法。其中外部评测方法是一种间接的评测方法，通过将方法生成的摘要应用到某一个特殊任务中，根据其对这项任务的贡献来评测该摘要方法的性能。内部评测方法的应用最为广泛，它通过直接分析评价生成摘要的质量来评测摘要方法的性能。在实现上需要有标准摘要作为参考，方法生成摘要与标准摘要吻合度越高，则其质量越高。多用于抽取式摘要方法的评测。

内部评测方法又可分为两类：人工评测和自动评测。人工评测传统的摘要评价方法主要由人工根据简洁性、可读性、一致性和内容含量等几个指标来评价文摘的质量。在2007年NIST组织的DUC评测4中，人工评测指标如表3-1所示。人工评测结果在很大程度上都是可信的，然而，在针对大规模文本进行评测时，人工评价需要消耗大量的人力，时间成本太高，效率太低，实现起来比较困难。

|  |  |
| --- | --- |
| 人工评测点 | 评测重点 |
| 语法性 | 以摘要的语言质量为主 |
| 无冗余程度 |
| 指代清晰程度 |
| 聚焦情况 |
| 结构和连贯性 |
| 响应 | 综合摘要的语言质量和内容质量 |

自动评测使用计算机来评价摘要效果，需要给定参考摘要作为标准答案，通过制定一些规则来给生产的摘要打分。以下是一些常用的自动评测方法:

(1)准确率(precision)和召回率(recall)

依据方法生成摘要的精度和信息覆盖率来评价摘要质量，应用准确率和召回率这两种评价指标将方法生成摘要和标准摘要进行比较。

(2)F-Measure

F-Measure是准确率(Precision)和召回率(Recall)加权调和平均，对准确率和召回率进行综合考核。该值越高，说明实验效果越好。

(3)ROUGE

R0UGE是由Lin等人在2004年提出的一种自动摘要评价方法，后被广泛应用在DUC的摘要测评任务中。该评测方法思想为通过计算方法生成摘要和标准摘要的n元词(n-gram)的共现信息来评价摘要质量。ROUGE包括一系列的评测方法，包括: ROUGE-N，ROUGE-S，ROUGE-L和ROUGE-W。其中反R0UGE-N方面最多使用的评测指标是ROUGE-1和ROUGE-2，均取得了很好的评测效果。

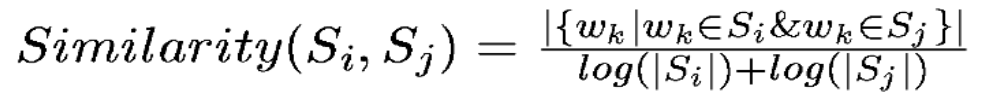
## 3.3代表模型

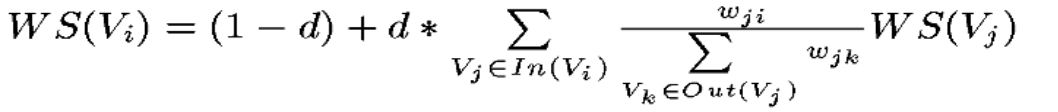
## 3.3.1 抽取式模型

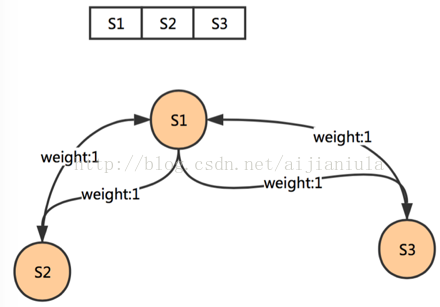
抽取式模型主要是从文章重提取重要的句子，连接成摘要。在选取关键句的方法上，使用的是rank的思想。

**3.3.1.1 Word2vec+BM25+TextRank**

借鉴了PageRank方式，将原文本拆分为句子，在每个句子中过滤掉停用词，并只保留指定词性的单词，由此可以得到句子的集合和单词的集合。将每个句子看成图中的一个节点，若两个句子之间有相似性，认为对应的两个节点之间有一个无向有权边，权值是相似度。

  
通过pagerank算法计算得到的重要性最高的若干句子可以当作摘要。由于是有权图，PageRank公式略做修改：





**3.3.1.2 TF-IDF+LDA+SVM**

该方法是使用特征工程的方式，通过句子的位置、词的词频、句子长度以及LDA提取特征，进行特征工程，再经过SVD输入SVM中进行重要度的rank，从而选出重要的句子。

## 3.3.2 生成式模型

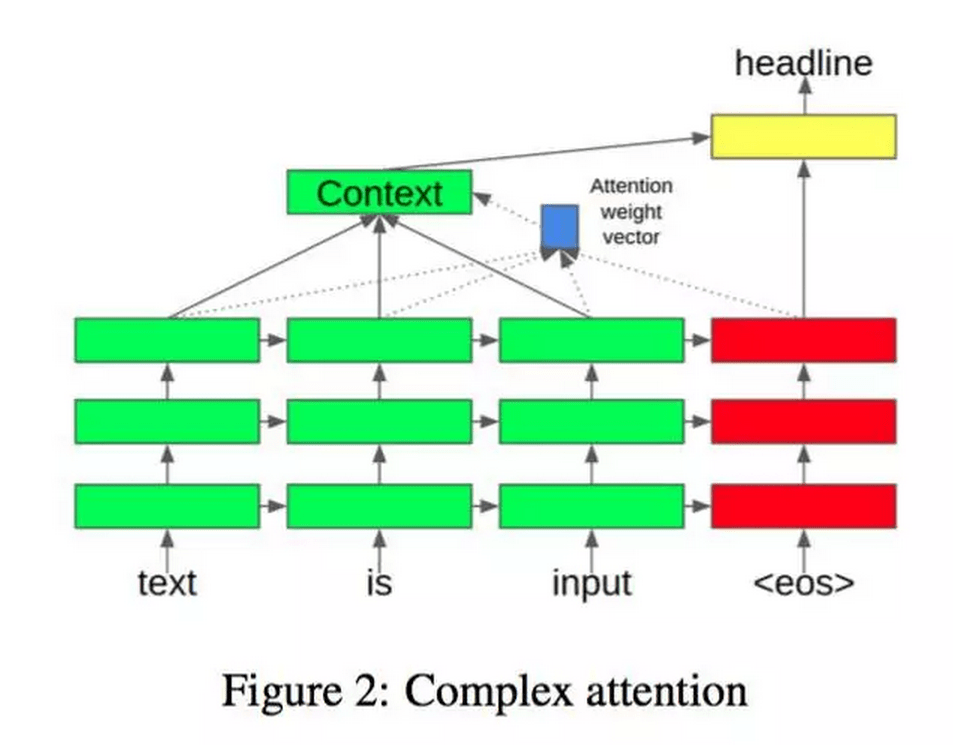
生成式模型，目前最为流行的是基于Encoder-Decoder框架的seq2seq 模型，该模型通用结构如下：

## 

## encoder部分用单层或者多层rnn/lstm/gru将输入进行编码，decoder部分是一个语言模型，用来生成摘要。这种生成式的问题都可以归结为求解一个条件概率问题p(word|context)，在context条件下，将词表中每一个词的概率值都算出来，用概率最大的那个词作为生成的词，依次生成摘要中的所有词。这里的关键在于如何表示context，每种模型最大的不同点都在于context的不同，这里的context可能只是encoder的表示，也可能是attention和encoder的表示。decoder部分通常采用beam search算法来做生成。

## 3.3.2.1 Complex Attention Model

## 模型中的attention weights是用encoder中每个词最后一层hidden layer的表示与当前decoder最新一个词最后一层hidden layer的表示做点乘，然后归一化来表示的。



## 3.3.2.2 Simple Attention Model

## 模型将encoder部分在每个词最后一层hidden layer的表示分为两块，一小块用来计算attention weights的，另一大块用来作为encoder的表示。这个模型将最后一层hidden layer细分了不同的作用。

## 3.3.2.3 Attention-Based Summarization(ABS)

## 这个模型用了三种不同的encoder，包括：Bag-of-Words Encoder、Convolutional Encoder和Attention-Based Encoder。这个模型中，让我们看到了不同的encoder，从非常简单的词袋模型到CNN，再到attention-based模型，而不是千篇一律的rnn、lstm和gru。而decoder部分用了一个非常简单的NNLM，就是Bengi于2003年提出来的前馈神经网络语言模型，这一模型是后续神经网络语言模型研究的基石，也是后续对于word embedding的研究奠定了基础。可以说，这个模型用了最简单的encoder和decoder来做 seq2seq，是一次非常不错的尝试。

## 3.3.2.4 ABS+

## Rush提出了一个纯数据驱动的模型ABS之后，又提出了一个abstractive与extractive融合的模型，在ABS模型的基础上增加了feature function，修改了score function，得到了这个效果更佳的ABS+模型。

## 3.3.2.5 Recurrent Attentive Summarizer(RAS)

## 这个模型是Rush的学生提出来的，输入中每个词最终的embedding是各词的embedding与各词位置的embedding之和，经过一层卷积处理得到aggregate vector：

## 

## 根据aggregate vector计算context（encoder 的输出）：

## 

## 其中权重由下式计算：

## 

## decoder部分用RNNLM来做生成，RNNLM是在Bengio提出的NNLM基础上提出的改进模型，也是一个主流的语言模型。

## 3.3.2.6 big-words-lvt2k-1sent模型

## 这个模型引入了large vocabulary trick(LVT)技术到文本摘要问题上。本方法中，每个mini batch中decoder的词汇表受制于encoder的词汇表，decoder词汇表中的词由一定数量的高频词构成。这个模型的思路重点解决的是由于decoder词汇表过大而造成softmax层的计算瓶颈。本模型非常适合解决文本摘要问题，因为摘要中的很多词都是来自于原文之中。

## 3.3.2.7 words-lvt2k-2sent-hieratt模型

## 

## 文本摘要中经常遇到这样的问题，一些关键词出现很少但却很重要，由于模型基于word embedding，对低频词的处理并不友好，所以本文提出了一种 decoder/pointer机制来解决这个问题。模型中decoder带有一个开关，如果开关状态是打开generator，则生成一个单词；如果是关闭，decoder则生成一个原文单词位置的指针，然后拷贝到摘要中。pointer机制在解决低频词时鲁棒性比较强，因为使用了encoder中低频词的隐藏层表示作为输入，是一个上下文相关的表示，而仅仅是一个词向量。这个pointer机制和后面有一篇中的copy机制思路非常类似。

## 3.3.2.8 feats-lvt2k-2sent-ptr模型

## 

## 数据集中的原文一般都会很长，原文中的关键词和关键句子对于形成摘要都很重要，这个模型使用两个双向RNN来捕捉这两个层次的重要性，一个是word-level，一个是sentence-level，并且该模型在两个层次上都使用attention，权重如下：

## 3.3.2.9 COPYNET

## 

encoder采用了一个双向RNN模型，输出一个隐藏层表示的矩阵 M 作为decoder的输入。decoder部分与传统的Seq2Seq不同之处在于以下三部分：

**预测：**在生成词时存在两种模式，一种是生成模式，一种是拷贝模式，生成模型是一个结合两种模式的概率模型。

**状态更新：**用t-1时刻的预测出的词来更新t时刻的状态，COPYNET不仅仅词向量，而且使用M矩阵中特定位置的hidden state。

**读取M：**COPYNET也会选择性地读取M矩阵，来获取混合了内容和位置的信息。

这个模型与第7个模型思想非常的类似，因为很好地处理了OOV的问题，所以结果都非常好。

## 3.3.2.10 MRT+NHG

## 这个模型的特别之处在于用了Minimum Risk Training训练数据，而不是传统的MLE（最大似然估计），将评价指标包含在优化目标内，更加直接地对评价指标做优化，得到了不错的结果。

# 第四章 存在问题

## 4.1评价指标

一个评价指标是否科学直接影响了这个领域的发展水平，人工评价由于带有较大主观性且无法进行量化学习，所以不再讨论，只讨论自动评价。ROUGE 指标在2004年就被 Lin 提出了，13年过去了，仍然没有一个更加合适的评价体系来代替它。而ROUGE 评价太过死板，只能评价出 output 和 target 之间的一些表面信息，并不涉及到语义层面上的东西，是否可以提出一种更加高层次的评价体系，从语义这个层面来评价摘要的效果。因此，迫切的需要一种新的体系来评价摘要效果，从而推动自动文摘领域的发展。创造新的评价指标技术上问题不大，因为计算两个文本序列之间的相似度有无数种解决方案，有监督、无监督、半监督等，但是每种方案的具体评价效果，需要大量的时间和语言专家进行测评。

## 4.2数据集

本报告的前段对当前数据集已有介绍，总体来说，英文数据集由于顶会和顶级比赛的支持，数据质量比较优秀。如Daily Mail拥有整篇的新闻和大段的媒体编辑的专业缩写标注，在中文数据集方面，LCSTS 数据集的构建给中文文本摘要的研究奠定了基础，很大程度地推动自动文摘在中文领域的发展。然而LCSTS中的数据文本（100-300字）和摘要标注（50字以内）的规模较小，和实际生产环境要求还有较大差距。虽然现在的互联网门户网站拥有大量的非结构化数据，但如何构建一个高质量的语料是一个难题。如何尽量避免用过多的人工手段来保证质量，如何用自动的方法来提升语料的质量都是难题。所以，如果能够提出一种全新的思路来构建自动文摘语料的话，将会非常有意义。

## 4.3训练速度

由于一般的研究性模型搭建均使用了tf，torch等python下的计算框架，而由于RNN系列的网络是迭代机制，无法并行。导致复杂的网络架构在训练的速度缓慢。有一些训练任务的训练时间可能需要数月之久，因此一个高效的底层语言计算框架是十分必要的。

## 4.4 重复与混乱

在抽取式模型中，由于是在全篇章中抽取高重要性句子，并顺序排列，句子间的关联性会比较差，或者出现语序混乱的问题；而在生成式模型中，由于RNN解码器的的记忆局限性，经常会产生重复生成相同词语的问题。因此两种模型都有各自的问题，研究两种模型的融合时分必要.

## 4.5 OOV(out of vocabulary)

文本摘要的本质是一个语言生成的问题，而只要是涉及到生成的问题，必然会遇到OOV问题，因为不可能将所有词都放到词表中来计算概率，可行的方法是用选择 topn 个高频词来组成词表。模型6，7和9都采用了相似的思路，从input中拷贝原文到output中，而不仅仅是生成，这里需要设置一个gate来决定这个词是copy来还是generate出来。显然，增加了copy机制的模型会在很大程度上解决了OOV的问题，就会显著地提升评价结果。

# 第五章 研究点及希望产出的结果（按优先级排序）

## 5.1抽取式和生成式相融合的模型（论文）

## 5.2高效快速的神经网络计算架构（专利）

## 5.3文本抽取开放api及演示平台（产品）

## 5.4 高质量的中文生成式数据集（开源贡献）