

自动文摘综述

自动文摘是自然语言处理其中的一个分支任务，其主要目的是自动化的对长文本进行处理，得到满足需求的短文本。最基本的自动文摘任务要求得到的短文本要保留原始文本中的主要语义，而针对某些更细分的任务，还要求自动文摘方法能够做出变化和适应。本篇综述主要针对自动文摘方法近些年的发展做出梳理和总结。

任务定义：

文献1^[1]给出了自动摘要任务一个较好的定义：

“Automatic text summarization is the task of producing a **concise** and **fluent** summary while preserving **key information** content and **overall meaning**”

在这个定义当中，针对自动文摘的形式和内容各提出了两个要求，其在形式上要求得到的文摘是简要和流畅的。首先文摘内容是简要的，是对原文内容进行压缩，使得人们能够更加容易把握其主要含义，这就对自动文摘的长度进行限制。目前主流的短文本摘要任务，一般要求将一段话（100词）左右，缩减成20个词以内的一句话。太长的摘要不符合人们的阅读习惯，影响人们的阅读体验。其次是要求得到的文摘是流畅通顺的，不同于信息抽取，自动文摘所得到的结果应该是符合人们用语习惯的一段自然语言，而不是原文主要信息的简单堆叠。这就要求文摘在实现信息压缩的基础上，还要具有合理的语法结构。

而针对文摘的内容，要求是要把握住**关键信息**和**整体语义**。首先文摘的关键信息是需要忠实于原文本的，是需要准确而真实的，其中关键的命名实体、数字、日期等关键信息是不应该出现错误的，否则文摘失去了准确性，其意义也就丧失了。其次文摘需要对于原文的整体含义进行把握，有一个整体上的概括。不同于句子压缩，一段原文本内可能含有两句或者三句话，需要对这几句话做一个综合性的把握，概括其整体含义，而不能仅仅偏向于某个单句，局限于某些细节，这会使文摘失去其概括性。

综合以上可以看出，自动文摘是一项充满挑战性的NLP任务。

任务意义：

文献2^[2]给出了自动摘要任务的任务意义：

- Summaries reduce reading time;
- When researching documents, summaries make the selection process easier;
- Automatic summarization improves the effectiveness of indexing;
- Automatic summarization algorithms are less biased than human summarizers;
- Personalized summaries are useful in question-answering systems as they provide personalized information;

首先，在这个信息爆炸的时代，如何从充斥着广告、谎言和不同观点的海量文本中获取有效信息已经成为了一种奢侈的能力，自动摘要技术能够帮助人们迅速及时压缩信息，获取关键点，节省大量阅读时间。其次，在信息检索和信息索引任务当中，如果有更高水平的摘要和概括，就能更好的提升这两个任务的效果。最后，人工摘要受限于个人的教育水平，政治观点，宗教信仰等因素，往往带有鲜明的个人偏见，而一个建立在海量数据集上的自动摘要方法能够有效的中和这种偏见，让我们看见更客观的世界。

- 一点个人的碎碎念：
本身我们在这个世界上看到的東西都是具有偏见的，所谓客观，只是发声者观念上的客观。现代

信息轰炸式的传播更是让我们陷入到种种偏见的漩涡当中。了解信息背后的观点，了解观点背后的立场，了解立场后的偏见，才是一个独立人应该看待这个世界的方式。

更概括的来说，自动摘要本质上是一种信息提取和重构的过程，相比于信息抽取，我们得到的不是结构化的信息，而是更容易理解的自然语言信息。摘要获取的不是孤立的元素，而是彼此相互联系相互补充的信息集合，并通过符合人们语言习惯的形式，重构和展现出来。这对于我们每个人在生活中获取有效信息是极其有帮助的。

任务方法分类：

自动摘要的分类方法很多，本文采用从方法层面划分的方式，将自动摘要分为两类：抽取式摘要和生成式摘要

- 抽取式摘要是从原文本中选取词、短语、句子进行重新组合，得到摘要文本的方法。抽取式文摘由抽取和组合两个步骤组成，抽取着重获取原文中“重要”的信息部分，组合则是将我们得到的关键信息进行重构，从而得到通顺流畅的摘要文本。
- 生成式摘要则是基于对原文的理解和概括，生成全新摘要文本的过程。抽取式摘要不局限于原文内容，但是也对自动摘要的技术提出了更高的挑战。

在实际操作当中来看，抽取式文摘和生成式文摘比较大的区别是抽取式文摘的词表是和原文词表一致的，生成的摘要中的内容是一定在原文中出现过的，而抽取式文摘的词表则没有限制，生成的摘要中的内容有可能并没有在原文当中出现。而一般人工摘要都是生成式摘要方法，这种方法生成的摘要概括性强，流畅度高，但是相应的在摘要生成过程中，更大的词表意味着更大的噪音和更多的不确定性，对于算法的要求也更高

任务数据集：

目前自动摘要任务有三个主流公开数据集，提供了训练数据与测试数据，各类方法也主要在这些数据集当中进行比较较量。数据集详细情况见下表：

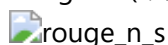
数据集名称	数据集描述
DUC数据集	DUC数据集包含DUC 2002 2003 2004评测所提供的数据，其中每份数据包含500条来自于纽约时报和美联社咨询的新闻原文本，并且每条新闻给出了4条由不同专家概括的摘要。由于其数据量较小，但是数据质量高，所以一般将其作为自动摘要任务的通用测试集。
Gigaword数据集	Gigaword数据集包含了约950万条新闻数据，在自动文摘任务当中选取每条新闻的第一句作为原文本，选取新闻的标题作为摘要内容，以此构建了约400万个原文本-摘要对。该数据集是目前自动文摘任务的主要训练集。
LCSTS数据集	LCSTS数据集是中文新浪微博数据集，由哈工大智能计算研究中心在2015年开源。以每条微博的内容做原文本，微博标题作为摘要，构成了原文本-摘要对，包含有约240条训练集，1万条验证集和1千条测试集，其中验证集和测试集都进行了人工打分标注。是质量较高的中文预料集。

任务评价指标：


自动文摘的评价一直是一个重要问题，目前主要分为两类：一类是人工评价，一类是使用ROUGE指标进行自动评价。人工评价在上个世纪是主流的评价方法，而在近年的方法中，更多使用ROUGE指标进行自动评价，但是目前最新的文章当中人工评价指标又成为了方法评价的重要补充。

ROUGE方法是在2004年由文献3^[^3]提出的一种自动评价指标，ROUGE基于摘要中n元词(n-gram)的共现信息来评价摘要，是一种面向n元词召回率的评价方法。基本思想为由多个专家分别生成人工摘要，构成标准摘要集，将系统生成的自动摘要与人工生成的标准摘要相对比，通过统计二者之间重叠的基本单元(n元语法、词序列和词对)的数目，来评价摘要的质量。通过与专家人工摘要的对比，提高评价系统的稳定性和健壮性。此方法是目前最为主流的文摘评价方法，而Rouge-1、Rouge-2和Rouge-L则是最被广泛使用的值。下面将给出相关公式：

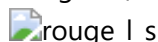
Rouge-N (单文档):



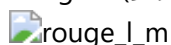
Rouge-N (多文档):



Rouge-L (单文档):



Rouge-L (多文档):



从公式可以看出，Rouge-1是对于单个词匹配进行衡量，Rouge-2是对于2-gram进行衡量匹配，Rouge-L是对于句子的最长公共子序列进行匹配。但是Rouge也有其自己的缺点，本质上是词袋模型，对于摘要的区分度并不高，有的语义相悖的句子所得到的Rouge值相差不大。同时Rouge要求和给出的标准摘要进行对比匹配，因此在标准摘要数量少的情况下，不同于标准摘要表述方式的结果将会得到很低的评价。这也是目前学界争议较大的一点。

任务方法综述：

在本部分将会对于自动摘要中有较大贡献的方法进行叙述。综述按照 传统方法 、 机器学习方法 、 深度学习方法 的顺序进行叙述，这个顺序也就是自动摘要方法逐步发展变化过程。

传统方法：

在传统方法中，由于受限于数据量，因此生成式文摘的获取十分困难，所以传统方法以抽取式文摘为主，其主要思路是从原文本中抽取关键信息并进行重构，以达到摘要的目的。

在传统方法中，首当其中的是基于规则和模板的方法。人在进行摘要工作的过程中，一般会将原文本转化为下面的三个结构：“what”，“what happened”，“who action what”。因此在自动文摘中，也构建相应的结构模板，然后从原文本当中找出对应信息，进行填空。在这种方法中，往往根据词的特征:例如 tf-idf 和词性，来进行相应词的筛选，然后填充到相应的模板当中。

在文献4^[^4]当中，作者提出了一种基于句法树的抽取式文摘方法，通过句法树，实现对于句子结构的解析，通过对树的剪枝和修改操作，得到目标文摘。这种方法也是强规则依赖，但是在实际应用中有着不错的效果。下图展示了此方法由原文本生成文摘的一个例子：



机器学习方法：

在机器学习的方法中，其最主要的思路是将自动摘要问题转化为一个删除-保留问题，即依次决定每一个词是删除还是保留，由此就将文摘问题变成了0-1序列标注的问题。而对于序列标注问题，很多机器学习方法例如逻辑回归，支持向量机，朴素贝叶斯、多层感知机，隐马尔可夫模型等方法都可以处理类似问题，相关工作已经被很多人做过，在此不再赘述。

深度学习方法：

- ABS和ABS+:

ABS和ABS+是2015年由Rush在文献5^[5]当中提出的基于注意力机制的编码-解码文摘模型。带注意力机制的编码-解码模型2014年在机器翻译当中取得了当时的最优效果，并且知道今天依然是机器翻译的主流模型。ABS和ABS+模型是首次将上述模型应用到了自动文摘的任务当中，是自动文摘任务深度方法的开篇之作。模型的输入是原文本，输出是摘要文本，通过端到端的训练，构建文摘的生成模型，该模型在Duc 2004测试集上取得了当年的最好水平，并且比传统方法和传统机器学习方法有着不小的提升。该模型的突出贡献有两点，第一是构建了自动文摘深度方法的范式，此后许多模型都是基于该模型进行修改。第二是提出了将Gigaword数据集用于自动文摘任务的方法，其构建的标题-首句对很好的解决了自动文摘任务缺乏标注数据的问题，为深度方法的应用提供了数据量的保障。下图显示了该模型的示意图：



- LCSTS:

LCSTS模型是2015年由Hu在文献6^[6]当中提出，其突出贡献是提供了中文自动摘要数据集LCSTS，该数据集使得中文自动摘要工作有了新的进展。LCSTS以每条微博的内容做原文本，微博标题作为摘要，构成了原文本-摘要对，包含有约240条训练集，1万条验证集和1千条测试集，其中验证集和测试集都进行了人工打分标注。是质量较高的中文预料集。在论文中，作者复现了上文Rush的模型，在LCSTS数据集上取得了较为不错的效果。下图为LCSTS数据集实例：



- Deletion with LSTMs 和 Extracting Sentences and Words:

Deletion with LSTMs 是2015年由Filippova在文献7^[7]当中提出的用深度方法实现抽取式文摘的模型，其本质上还是删除-保留问题的处理，使用了LSTM进行序列处理，取得了不错的结果。Extracting Sentences and Words 是2016年由Cheng在文献8^[8]当中提出的抽取式文摘模型，其划分了两个层次，分别对句子和词进行抽取，实现了对于含有多句的长文本实现自动摘要。这两个模型都使用了自己给出的数据集，下图展示了两个模型的结构图：



- RNNs and Beyond:

RNNs and Beyond是2016年由 Nallapati 在文献9^[9]当中提出的模型，该篇文章针对前面的baseline模型做出了一系列的改动和细化，最终达到了当年的最优效果。其有5个突出的改进点：

1. 引入 Large Vocabulary Trick 来解决 decoder 词表过大的问题；
2. 加入传统的 TF-IDF，POS，NER 等特征来尝试抓住句子的关键部分；
3. 引入 Generator-Pointer 来解决 OOV 和低频词的问题；
4. 引入 Hierarchical Attention 来抓住句子的重要性信息；

5. 提出新的数据集 CNN/Daily Mail;

该模型针对这些细节做了详实的实验和分析，并且开源了实验代码，使得相应的模型更加成熟，效果也比之前的baseline有了较为显著的提升，可以说是极大的提升了自动摘要技术的实用程度。

- CopyNet:

该模型是2016年由Gu在文献10^[10]当中提出的模型，CopyNet 是在 Seq2Seq + Attention 的基础上，引入了拷贝机制，使得模型对于某些关键信息的把握更加准确，即对于细节信息选择从原文中拷贝，而不是根据序列解码生成。此模型在机器翻译等其他自然语言处理任务当中也得到了较好的结果，下图为模型示意图：



- Pointer-Generator and Coverage:

该模型是2017年由See在文献11^[11]当中提出的模型。该模型的提出是为了解决自动摘要任务中常常出现的两个问题：细节错误和重复。其首先提出了Pointer-Generator用来解决细节错误的问题，对于某些文本细节，选择从原文中复制而不是解码生成，并用一个网络预测是应该解码还是复制。其次，该模型提出了Coverage机制，用coverage vector来惩罚重复词的生成。通过这两个机制实现了对这两个问题的较好解决，也取得了最好的结果。下图展示了模型结构：



任务局限及未来展望：

任务局限：

- 模型局限：

深度自动文摘的模型基本上是基于机器翻译模型进行改进和提高，然而近来想在效果上实现提升已经越来越难，越来越接近瓶颈，摘要效果距离人工摘要还有不小的距离。想要实现进一步的提升，更应该思考模型本身，毕竟机器翻译和自动摘要的关注点是不一样的，简单的使用同样的模型，很难复现机器翻译目前的效果水平。

- 评价指标局限：

目前主要以Rouge值作为文摘好坏的指标，然而经过研究Rouge本身也有很大的局限，很难把握语义信息的内容，完全不同语义的摘要，Rouge值并没有很大的区分。所以新的方法在这样一个并不敏感的指标下的效果，可能并不能被更好的衡量。目前人工评价似乎又回到了人们的视野当中，但是如何细化，如何有更丰富的人工标注语料，目前仍然是一个不小的挑战。

- 应用局限：

目前，单纯的摘要任务由于语料集的相对老旧，并没有在实际的自然语言处理技术应用当中有着很好的表现。因此很多学者希望能将其和例如信息检索、自动问答等任务进行结合，以期能够相互借鉴，相互促进。目前该任务和其他任务相结合也是一个发展方向。

任务展望：

- 未来首先希望能够将其和信息检索的任务进行结合，结合别的任务，结合别的任务的评价指标来体现自动摘要工作的价值。
- 要把工作做细，短文本摘要和长文本摘要不同，一般性的摘要和针对任务的摘要所关注的点也不应该一样，要认清他们的相同点和不同点，更好地做好这件事。

引用:

- [^1]: Torres-Moreno J. Automatic Text Summarization: Some Important Concepts[M] Automatic Text Summarization. John Wiley & Sons, Inc. 2014:23-52.
- [^2]: Gambhir M, Gupta V. Recent automatic text summarization techniques: a survey[J]. Artificial Intelligence Review, 2016, 47(1):1-66.
- [^3]: Flick C. ROUGE: A Package for Automatic Evaluation of summaries[C] The Workshop on Text Summarization Branches Out. 2004:10.
- [^4]: Cohn T, Lapata M. Sentence compression as tree transduction[M]. AI Access Foundation, 2009.
- [^5]: Rush A M, Chopra S, Weston J. A Neural Attention Model for Abstractive Sentence Summarization[J]. Computer Science, 2015.
- [^6]: Hu B, Chen Q, Zhu F. LCSTS: A Large Scale Chinese Short Text Summarization Dataset[J]. Computer Science, 2015:2667-2671.
- [^7]: Filippova K, Alfonseca E, Colmenares C A, et al. Sentence Compression by Deletion with LSTMs[C] Conference on Empirical Methods in Natural Language Processing. 2015:360-368.
- [^8]: Cheng, J & Lapata, M 2016, Neural Summarization by Extracting Sentences and Words. in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, pp. 484-494.
- [^9]: Nallapati R, Zhou B, Santos C N D, et al. Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond[J]. 2016.
- [^10]: Gu J, Lu Z, Li H, et al. Incorporating Copying Mechanism in Sequence-to-Sequence Learning[J]. 2016:1631-1640.
- [^11]: See A, Liu P J, Manning C D. Get To The Point: Summarization with Pointer-Generator Networks[J]. 2017:1073-1083.