老师们好，我是李帅敏，我的中期报告题目是基于深度学习的自动文本摘要算法研究，指导老师是徐老师。

接下来，我们按照以下内容进行介绍。首先，我们介绍研究进展情况。

念PPT

自动文本摘要算法的研究在21世纪之间主要是启发式的算法，和基于人工特征的机器学习算法。21世纪之后，基于深度学习技术的文本摘要算法开始蓬勃发展，到现在仍是热点问题。

念。

我们的课题从生成摘要的不同方法展开。念。

目前，我们已经完成了前两个阶段的研究任务。

回顾基础知识，关于定义。。。。

生成式自动文本摘要的工作模式：。。。

实现方法通常是基于序列到序列的框架，编码器（encoder）先对输入文本进行编码得到输入文本的抽象表示，然后解码器（decoder）基于对文本的表示进行逐个单词的解码以生成摘要内容

模拟人类在摘要生成过程中的行为模式，更具智能性；

逐个单词生成使得生成式方法得到的摘要在内容更加连贯；

能够生成原文中未曾出现过的新的单词使得摘要的抽象程度更高。

抽取式方法由于从原文中直接抽取，具有简单、快速、和能保证事实一致性的特点。

这是我们所用到的数据集，除了DUC系列之外，其他数据集的规模都较大。

文本摘要算法的评价指标特点是主观性强，主要分为人工评价和自动评价。ROUGE 应用最为广泛，是一个基于N元共现统计的方法。

我们的基于生成式方法的自动文本摘要算法研究从编码、解码、以及编码和解码三个不同的角度进行。

首先，我们提出了一个基于局部和全局文档依赖的生成式多文档摘要模型。这一个方法提出的背景是。。。，具体的，我们。。。

**词编码层：**每个单词的向量表示由词向量和位置编码两个部分组成：。特别地，词级别和文档级别，即 。

**内部编码层**：使用Transformer层对每个文档的词序列进行编码，得到每个文档的文档表示。另外，使用Multi-Pooling将其进行长度统一。

**全局编码层**：实现某个文档与其它所有文档之间的信息交互。

**局部编码层**：构建文档语义关系图，只有语义相关的文档才会被关联。使用Graph Attention Network对语义文档关系图进行编码。

**融合层**：本层将全局文档表示和局部文档表示进行融合

实验结果如图，可以看到即使在输入长度增加的情况下，我们的模型仍然保持较好的性能，超过了一些基线模型。

接下来，从解码的角度出发，我们提出了一个基于历史和未来联合信息解码的生成式摘要模型。

该模型由编码器，反向解码器和正向解码器构成。

编码器由双向LSTM构成，对原文进行编码

反向解码器将目标摘要的逆序列作为解码目标，在解码过程中使用注意力机制关联到编码器的隐层状态。

正向解码器将目标摘要作为解码目标，在解码过程中使用两个不同的注意力机制关联到编码器和反向解码器的隐层状态。其中，编码器的隐层状态作为原文本的文本表示，反向解码器的隐层状态表示未来信息。

正反学，加深印象，增加了模型参数的多样性。

大部分已有的方法在解码过程中选择使用图4所示的束搜索方法[42]（beam search）来进行候选集合的筛选，虽然保证了最大化联合概率但是仅仅考虑已经生成的部分，而。。。

我们提出了一种推理阶段的解码方法综合考虑已生成部分的联合概率和未生成部分的ROUGE得分来进行候选序列的选择。

设计奖励预测（Reward predict）网络来预测未生成部分序列的ROUGE得分

该网络将编码器对源输入的隐层状态、正向解码器的隐层状态、针对编码器端源文本的内容向量（context vector）、以及针对反向解码器端的内容向量作为输入。

reward predict网络的目标ROUGE得分通过将某一时刻的部分生成序列的所有可能的解码序列ROUGE得分进行平均得到。

我们的实验结果如图，可以看出，我们的基于历史和未来联合信息的模型优于一些基线模型。

文本中通常包含着一些关键词或者关键短语，它们传递了一些重要的信息，在文本中扮演着重要的角色。我们认为关键词和关键短语相较于其它单词对输入文本的表示贡献了更多的信息，也就是说，关注和理解这些关键词和关键词语有助于生成更丰富的文本表示，从而提升生成式自动文本摘要的模型性能

我们的基于关键词感知编码的自动文本摘要如图所示，它由一个原文编码器，一个关键词编码器和一个关键词感知的解码器构成。

**原文编码器：**给定原文，原文的向量表示,其中表示Transformer中的自注意力机制。

**关键词编码器：**

给定的关键词, 那么关键词k\_i的向量表示为：

Transformer将一系列关键词连接起来：

由Transformer层构成，并且使用多头注意力机制完成对关键词和原文的感知。在得到原文本和关键词序列的向量表示之后，我们使用一个关键词感知的解码器来为科学文章生成摘要。具体的，关键词感知的解码器首。

实验结果如图所示，实验中我们采用了几种不同的关键词提取方式，实验证明在训练时采用标准摘要里的关键词，测试时使用原文中的关键词效果最好，且超过了一些基线模型。

念。

科学文章摘要除了要求简洁和涵盖重要内容之外还需要保持专业性，也就是说科学文章摘要是不允许出现事实性错误的，另外科学文章摘要需要传递出该研究的研究背景、研究方法、研究结果以及研究结论等科学文章要素。因此，为科学文章自动生成摘要相较于通用文本难度更大。

我们的基于机器阅读理解的抽取式文本摘要模型框架如图。

机器阅读理解（Machine Reading Comprehension, MRC）：给定给一个问题，和一个文档，片段选择类型的机器阅读理解方法旨在以给定文档为背景，选择文档内的一个片段范围作为给定问题的答案。

基于机器阅读理解的抽取式摘要方法（MRC-based Extractive Summarization）：该方法从原文本中为提前定义好的与科学文章要素相关的问题找出答案片段，并将针对不同科学文章要素的答案片段进行连接构成最终的抽取式摘要内容。

根据以上背景，我们设计了如下几个针对不同研究要素的问题。

由于摘要数据集中不存在标准的答案片段，因此我们需要自行标记。

我们首先根据摘要内容将摘要数据集中的原文-摘要对分成两类，具体的，含有明确的摘要结构的数据被定义为严格数据对（strict data pairs），其他不含有明确摘要结构的数据被定义为非严格数据对（non-strict data pairs）。然后，我们对严格数据对（strict data pairs）和非严格数据对（non-strict data pairs）分别进行标准答案片段的标记。标准答案标签的标记流程如图2-14所示。

对严格数据对的标准答案片段标记分为两个步骤：

（1）要素的聚合。首先我们将标准摘要中的句子分为背景、方法、结果和结论等要素。如图2-15所示，在严格数据对中我们很容易得到句子的要素标签。然后，我们将标记为相同要素标签的句子进行连接,分别表示要素标签为背景、方法、结果和结论的。

（2）ROUGE得分的计算。以标记关于研究背景的问题的答案段落为例，我们计算了在第一步中得到的要素标签为“背景”的序列与原文中每个句子的ROUGE得分，并且得分最高的原文序列作为与研究背景相关的问题的答案。

对于非严格的数据对，由于这类型的数据对中摘要中句子的要素标签是不明显的，而我们假设摘要必然包含研究背景、方法、结果和结论这四个部分。因此，我们通过微调BERT实现了一个句子分类模型，来预测非严格数据对中每个摘要的句子的要素标签。具体的，这个分类模型，输入格式为，具体的训练数据从严格数据对中得到，比如，,, 其中，。将非严格数据对摘要中的句子逐个输入这个训练好的分类模型，我们可以得到每个句子的要素标签。这样以来，我们可以继续使用在严格数据对中标记答案段落的方法来得到非严格数据对中的标准答案片段。

经过以上方法，对于每一个原文本，我们可以得到四个预先定义的问题以及在原文中对应的答案片段，图2-16给出了一个原文对应的问题和答案样本示例。

科学文章通常包含数以千计的单词量，因此我们使用Longformer[52]来实现基于机器阅读理解的抽取式摘要算法。Longformer是一个基于Transformer结构的模型，但是它克服了其他Transformer模型无法处理长序列的困难，特别的，它使用了一个局部窗口滑动注意力机制和一个任务驱动的全局注意力机制来处理长文本。

用Longformer对输入序列进行编码，得到的向量表示

接下来，两个线性层被分别用来预测开始和结束位置在原文本中的概率分布：

其中，和分别是两个线性层的参数，和是偏差。函数用来进行进一步的概率归一化。最后基于概率分布和可以得到开始位置和结束位置。

表2-8展示了我们所提出的基于机器阅读理解的抽取式科学文章摘要模型（MRC-Sum）和一些基线抽取式模型在三个科学文章数据集上的ROUGE-1，ROUGE-2和ROUGE-L得分。如表格结果所示，基于机器阅读理解的抽取式科学文章摘要模型性能超过了其他基线的抽取式摘要模型。

我们认为能够包含科学文章几个要素的摘要更具备全面性，基于这一想法我们提出了一个新的评价指标来评价摘要内容的全面性。具体地，我们重用了在标准答案片段标记过程中使用到的句子分类模型，来预测系统生成的摘要中每个句子的要素标签，然后我们的全面性评价指标可以通过如下公式进行计算：

我们在三个科学文章数据集上进行了得分计算，结果如表2-9。该表的结果显示，基于机器阅读理解的抽取式科学文章摘要模型能够生成内容全面的摘要内容。

上述基于机器阅读理解的抽取式科学文章摘要算法中问题每个问题的设计都是相互独立的，且对于每个文档四个问题都是相同的。为了增加问题与当前文档的联系，以及问题之间的关联程度，为此我们重新设计了问答流程来优化上一阶段的基于机器阅读理解的抽取式科学文本摘要框架。优化后的模型整体框架如图2-17所示

我们认为之前的问题设计虽然能够生成在科学文章要素方面具有全面性的摘要内容，但是简单的将彼此独立的问题对应的答案片段拼接，会导致摘要内容本身在过渡上的生硬，从而使得摘要内容整体的逻辑性不强。基于这一点考虑，我们重新设计了具有逻辑相关关系的问题，每一个问题都与上一个问题的答案紧密相连，具体的说，我们认为某一项研究的方法是在相应的背景下提出的，研究结果是在使用某种的研究方法的前提下产生的，研究结论则往往与一项研究结果有因果关系。因此关于背景要素的问题的答案应当作为一个提示内容被关联到关于方法要素的问题中去。类似的，关于方法要素的问题的答案应该作为提示内容被关联到关于结果要素的问题中去，关于结论要素的问题的答案应该作为提示内容被关联到关于结论要素的问题中去。

第二个优化存在于从答案片段的候选集到摘要内容的生成阶段，在前一个模型中，对于每个问题，都存在有一个答案片段候选集，该集合根据答案片段的概率大小进行排名，然后选择每个问题对应的答案片段候选集合中的排名最高的答案进行拼接得到最终的摘要内容。然而，在新的问答流程中，当前问题和上一个问题的答案之间有直接的关联，这有助于增强摘要内容的联系，但是同时存在一个容易错误累积的问题。也就是说，如果前一个问题的答案与正确答案的存在较大的偏颇，那么势必会造成下一个问题和下一个问题的答案偏离正确的轨道。为了解决这个问题，平衡文本内容之间相关关系以及错误累积问题，

引入了彼此独立的问题集合，并使用这些彼此独立的问题的答案片段候选集来对具有相关关系的问题的答案片段候选集进行重新排序。以“方法”要素相关的问题为例，我们可以得到两个不同的答案片段候选集合,。受到Pagerank算法的启发，我们使用中的答案片段来对中的答案片段进行重新排序。

重新排序的方法既考虑了答案片段本身的概率问题，又考虑了彼此独立的问题的答案片段候选集与具有相关关系的问题的答案片段候选集之间的语义相似度。特别地，答案片段之间的语义相似度由sentence-BERT[53]计算得到。

表2-10展示了我们的实验结果，实验证明优化后的LAMES模型较之前的模型性能有所提升，证明了我们所做的优化是有效果的。

表2-11展示了几个不同预训练模型下，替换LAMES中的具有逻辑关系的问题为相互独立问题后摘要模型的表现，实验证明具有逻辑关系的问题能够提升模型性能。

第三个研究内容，我们计划考虑综述生成任务，这一任务在摘要生成中可以被看做是一个多文档摘要任务。

我们首先分析了目前具有代表性的MS2数据集，该数据集包含407k个文档和20k个摘要。输入是与综述相关的文章，个数在1-401个不等，平均有 23个

输出部分是摘要中的结论和影响部分，平均只有两句

问题是长输入对神经网络来说有负担，且输入和输出的压缩比例过高。

我们计划。。。

优点是。。。