各位评委老师好，我叫孙相会，指导教师是张天成老师。我的开题题目是基于机器阅读理解的中文问答系统研究与实现。我的汇报提纲如下，我们先来看一下选题背景与意义。

左图是搜索引擎的例子，搜索引擎的原理是对互联网上数以万计的网页文档中的关键词建立索引库，根据用户输入的问题，基于关键词的检索方式，所有页面中包含问题关键词的网页作为搜索结果检索出来，然后利用排序算法，按照相关度由高到低的顺序返回网页文档。

右图是问答系统的例子，我们说搜索引擎这种基于关键词的检索方式缺乏对用户问题语义的理解，而且返回的是排序的网页文档，仍需要再次筛选答案。而问答系统利用自然语言处理技术，对问题深层次的语义分析，通过检索、推理等方法返回精确的答案，不需要再次筛选。

问答系统大致可以分为三大类：

* 基于知识库的问答系统：通过从知识库中查询与问题相关的实体，推理出实体之间的关系，返回答案。缺点在于需要预先构建好大规模的知识库。
* 基于问答对的问答系统：通过从数据库中检索出与用户问题相似的问题以及对应的答案，然后将答案与用户问题匹配，返回最相似的答案。缺点比较明显：可以回答的问题数量有限，数据库中有多少个问题就只能回答多少个答案，极度依赖于数据集，另一方面返回的答案形式单一，不具有多样性。

我的研究方向是基于机器阅读理解的问答系统，我们来看看什么是机器阅读理解。

机器阅读理解任务是给定问题以及与问题相关的一篇或者多篇文章，机器阅读这些文章后给出问题的答案。按照答案形式的不同，可以分成抽取式和生成式，生成式是指答案是从文章中归纳总结生成的一段文本，抽取式是指答案来源于文章中的某一段连续的文本。如图所示，目前我的研究计划是做抽取式任务。

基于机器阅读理解的问答系统相比于前两个：

1. 不需要预先构建大规模的知识库
2. 不依赖于数据集，同一篇文档可以回答多个问题

机器阅读理解的应用场景如下。

我们接下来看研究内容与方案：

主要研究内容包括两大部分：

* 设计并实现一个端到端形式的机器阅读理解型问答系统
* 对问答系统中的两个关键部分做深入研究，这两个关键部分是文本匹配模块和阅读理解模块

整个系统的流程是：

将用户输入的问题传输到搜索引擎接口-->从搜索引擎检索出来的网页文档中匹配出与用户问题相关的K篇文档，模型阅读这K篇文档，给出答案。

现在来看研究方案，下图分别是文本匹配模块和阅读理解模块的经典模型框架，我的研究方案就是在这两个基础框架上改进。

首先我们来看文本匹配模块的拟设计方案。我在这个模块的拟设计方案是多轮交互机制下的基于卷积结合自注意力的栅格结构。其中：

* 栅格结构Lattice我们可以看到，既考虑了字的信息，又考虑到了词的信息；
* 卷积用来提取局部特征，自注意力用来提取全局特征；
* 多轮交互机制可以增强问题和文章之间的语义交互信息。

输出0代表文章与问题不相关，输出1代表相关。

为了更好的提升文本匹配模块的性能，加快文本匹配模块的训练，拟采用知识蒸馏的机制。将BERT作为教师模型，我的栅格结构模型作为学生，将教师模型的输出概率分布作为标签指导学生模型。所以整个文本匹配模块的设计结构如图所示。

接下来我们来看阅读理解模块的设计。

阅读理解模块我是采用BERT作为基准，在BERT的基础上加上持续阅读模块和答案验证模块。持续阅读模块的作用是判断模型对文章的理解程度，只有对文章足够理解的情况下才允许模型从文章中预测答案。答案验证模块的作用是判断预测的答案与问题之间的语义关联程度，如果觉得预测的答案与问题的相关程度不高，那么重新预测答案。

最后就是利用强化学习中的奖励机制，将阅读理解模块预测的答案与真实答案之间的误差作为奖励指导文本匹配模块尽可能匹配出包含答案的文章。

如图是整个系统两大模块的训练结构。

接下来我们来看前期准备工作。

准备工作主要包括数据集的构建和基准模型的实验。

首先我们来看阅读理解数据集的构建：下表是机器阅读理解领域最常用的数据集，红色字体是中文数据集，目前我收集了五个大规模的中文阅读理解数据集，这些数据集中有的样本是不能够作为我的课题所使用的，有的数据集甚至是繁体字，我通过过滤筛选等方法将每一个数据集中可以作为我的课题所使用的样本保留下来，最后五个文件一共有152241个样本，每一个样本都是如下表所示的（文章、问题、答案）三元组，答案是文章中的某一个连续片段。

我也复现了阅读理解领域经典的模型BiDAF，下图是训练过程中记录的评估指标。Em指的是精确匹配，含义是预测的答案位置与真实的答案位置是否完全一致。F1是模糊匹配，含义是预测的答案中的单词与真实答案中的单词重叠数目。数据集中取出来20000个作为测试集，我们可以看到模型的准确率不高，甚至还有过拟合的问题。

下图是利用BERT微调的效果，由于BERT模型比较大，实验室机器配置不高，我仅仅取出来10000个样本作为训练集，3000个样本作为测试集，我们可以看到，仅仅用了很小的一部分数据集，BERT的效果已经显著优于BiDAF。

接下来就是第二个大模块，文本匹配模块数据集的构建，目前（问题，文档）这类数据集还没有，大多是（问题，答案），或者（问题，问题）这类判断问题之间相似度或者问题与答案之间相似度的数据集。因此我采取的办法是基于阅读理解的数据集构造文本匹配模块的数据集，具体构造方法是：

将原本的问题和对应的文章看作正样本

然后通过删除掉答案所在的句子或者随机选取句子替换掉答案所在的句子作为负样本

删除掉距离答案所在句子比较远的单词或者句子构造正样本，这些单词和句子对回答问题没有什么帮助。所以仍然可以作为正样本看待。

最后得到的负样本数目是430136，正样本数目是164957。

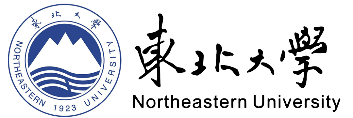
然后仍然是在BERT上微调数据集，80%作为训练，20%作为测试，可以看到在测试集和训练集上的准确率都非常高，这反映出数据集的构建方式过于简单。

下面我们来看预期研究计划：

接下来的首要计划就是文本匹配模块数据集的增强，因为根据模型的效果可以看出，数据集的构建方式过于简单，我采取两种数据增强方案：第一，通过爬取中小学语文考试阅读理解选择题来增强负样本的识别难度，如图所示，假设C是正确选项，那么问题集合选项A,B,D中的任何一个都可以与文章构造成负样本。这样构造的负样本满足答案不能从原文中抽取，而且问题和文章的相关度较高，不易识别。

此外还要增强正样本的数目，拟采取的方案是利用问题生成模型，将文章和答案作为输入，模型生成一个或多个问题，以此来构造正样本。这种方案可以增加模型的泛化性。

最后是研究计划与差异性分析：

研究计划包括：

* 实现在知识蒸馏引导下，基于卷积结合transformer结构的栅格文本匹配模块，并且在经过了数据增强的文本匹配数据集上训练。
* 实现所提出的阅读理解模型：即在BERT的基础上引入持续阅读模块和答案验证模块。
* 分别实现上述两个模块后，利用强化学习来拟合训练两个模块，最终实现问答系统。

差异性分析：

* 基于卷积结合transformer的栅格结构的提出
* 在中文问答系统中的检索匹配模块采用知识蒸馏技术
* 基于BERT结合持续阅读模块和答案验证模块的提出