自然语言处理前期报告

姓名：饶文均

时间：2022.11.18

当前进度

确定选题和数据集

确定选题为Stanford Question Answering Dataset (SQuAD)。

从https://www.kaggle.com/datasets/stanfordu/stanford-question-answering-dataset下载数据集。

了解数据集构建原理

- 作者们选出了英文维基百科中头部10,000篇文章，然后从中按均匀分布随机采样出536篇。对于每篇文章，都把独立的段落提取出来，并去掉图片、表格、以及少于500个字符的段落。这样处理之后，得到了536篇文章中的23,215自然个段落。然后将这些段落随机划分为训练集

（80%）、验证集（10%）、测试集（10%）。

- 斯坦福大学通过众包的方式，让一部分人根据23215个自然段，分别进行提问并标注答案。对于每个段落，众包工人需要基于段落的内容完成最多5个问答，问题必须以文本形式写出来，答案必须在段落中被高亮标记出来。**众包工人被鼓励用他们自己的词汇来提问，而不是直接复制段落中的词汇。**
- 然后，对于验证集和测试集的每个问题，另外召集一批众包工人根据这些问题以及段落，在段落中标出最短的答案，如果一个问题是无法回答的，则不标出答案并上报该问题。在验证集和测试集中，2.6%的问题被至少一个这一波的众包工人认为是无法回答的。这样做是为了获得人类在数据集上的表现以及让评估更具鲁棒性。

观察数据集结构

数据集中的每条数据主要由包含(question、context、answer)字段的三元组构成。其中的answers字段中包含了文章编号和所截取的作为答案的段落。

	id	question	context	answers	c_id
0	5733be284776f41900661182	To whom did the Virgin Mary allegedly appear i...	Architecturally, the school has a Catholic cha...	[{'answer_start': 515, 'text': 'Saint Bernadet...	0
1	5733be284776f4190066117f	What is in front of the Notre Dame Main Building?	Architecturally, the school has a Catholic cha...	[{'answer_start': 188, 'text': 'a copper statu...	0
2	5733be284776f41900661180	The Basilica of the Sacred heart at Notre Dame...	Architecturally, the school has a Catholic cha...	[{'answer_start': 279, 'text': 'the Main Build...	0
3	5733be284776f41900661181	What is the Grotto at Notre Dame?	Architecturally, the school has a Catholic cha...	[{'answer_start': 381, 'text': 'a Marian place...	0
4	5733be284776f4190066117e	What sits on top of the Main Building at Notre...	Architecturally, the school has a Catholic cha...	[{'answer_start': 92, 'text': 'a golden statue...	0
...	...	...	...	...	...
87594	5735d259012e2f140011a09d	In what US state did Kathmandu first establish...	Kathmandu Metropolitan City (KMC), in order to...	[{'answer_start': 229, 'text': 'Oregon'}]	18890
87595	5735d259012e2f140011a09e	What was Yangon previously known as?	Kathmandu Metropolitan City (KMC), in order to...	[{'answer_start': 414, 'text': 'Rangoon'}]	18890
87596	5735d259012e2f140011a09f	With what Belorussian city does Kathmandu have...	Kathmandu Metropolitan City (KMC), in order to...	[{'answer_start': 476, 'text': 'Minsk'}]	18890
87597	5735d259012e2f140011a0a0	In what year did Kathmandu create its initial ...	Kathmandu Metropolitan City (KMC), in order to...	[{'answer_start': 199, 'text': '1975'}]	18890
87598	5735d259012e2f140011a0a1	What is KMC an initialism of?	Kathmandu Metropolitan City (KMC), in order to...	[{'answer_start': 0, 'text': 'Kathmandu Metrop...	18890

确定数据集评估方式

- F1 （即准确率和召回率的调和平均值F-Measure）
- EM （Exact Match）

EM 是完全匹配，也即是机器给出的答案必须和人给出的答案完全一样才算正确。而F1 是将答案的短语切成词，再和由众包得出的答案计算Recall，Precision 和F1 ，即是部分答对也能得分。

对数据集总结和评价

- SQuAD数据集是一个**extractive question answering**（抽取式问答，即从文章中摘取一串字符作为答案）数据集而不是abstractive question answering（生成式问答）。
- SQuAD 和之前的完形填空类阅读理解数据集如CNN/DM，CBT 等最大的区别在于**SQuAD 中的答案不在是单个实体或单词，而可能是一段短语，这使得其答案更难预测。**

个人思考总结

在选题时，我认为处理Q-A 数据集就是类似于实现“人机对话”的任务，但了解这次选择的SQuAD 数据集后我发现处理SQuAD 数据集本质上就是让机器做一个“阅读理解”的任务。

作为人类来说，我们在小学、中学和大学阶段都参加过各种英语考试，我回忆了一下我在参加各难度等级的英语考试时的答题模式：

- 小学、初中阶段，阅读理解题目难度较低，我们在做阅读理解题目时的思路是通过在原文中搜索题干的**关键词**，从而定位可以回答问题的“supportive context”，最后得到答案。
- 在高中阶段和大学四级考试中，阅读理解题难度上升，大部分题目都不能通过直接搜索关键词的方法在原文中找到相应段落，原因是命题者会将题干中的一些关键词模糊化，甚至很多时候用近义词将其替代，这样可以达到混淆我们的目的，这样主要考察我们的**词汇能力**。
- 在六级考试、雅思考试中，阅读理解难度进一步提高，命题者不再使用“近义词混淆”方法，而是直接根据原文内容重写描述，改变了原文句子的结构和逻辑，以此考察学生对**深层语法知识和文章逻辑**的掌握程度。

> 对于每个段落，众包工人需要基于段落的内容完成最多5个问答，问题必须以文本形式写出来，答案必须在段落中被高亮标记出来。**众包工人被鼓励用他们自己的词汇来提问，而不是直接复制段落中的词汇。**

根据前文提到的数据集构成信息，可以知道SQuAD 数据集中包含了很多阅读难度较高的题目（对应上文的高中及以上的题目），这对机器的学习理解能力提出了一定层次的要求。如果仅仅通过简单的上下文关键词匹配算法来处理数据集，可以预见的结果是准确率不够高。数据集的这些特点就决定了处理这个问题的最佳方法是**神经网络模型**。（这里其实存在一个问题，由于在构建数据集时研究者只是**鼓励**众包工人用自己的词汇来提问，这就造成了一种不确定性：数据集中的问题阅读难度参差不齐，且比例未知。也就是说，如果在构建数据集时更多的众包工人选择直接复制段落词汇来进行提问，也就是问题过于简单，就会造成使用“关键词匹配”算法和使用神经网络模型得到的准确率差距没有我们想象中大，甚至可能会出现简单匹配算法准确率更高的情况，但SQuAD 数据集应该不存在这种问题）。

对于这个数据集来说，其实存在些许不足。前文提到，SQuAD 数据集中的提问逻辑和高中难度以上阅读题的提问逻辑相似，但在雅思阅读题中，有一种**“NOT GIVEN”**题型，也就是“原文中未提到”。而在这个数据集构建过程中，提问者提出的问题都能在文章中找到答案，如果在构建数据集时让提问者提出一定量的具有干扰性的“NOT GIVEN”问题，可能会让数据集更完备，从而使得训练出的模型更接近人类的阅读水平。但据我所知这个问题在SQuAD 2.0中得到了改善。

下一步工作

- 学习Google BERT 模型。
- 在自己的设备上调试并跑通代码，观察输入和输出情况。