题 目：《Learning from Task Descriptions》学习笔记

选 课 班：

选课序号：

姓 名：

学 号：

学 院： 信息科学技术学院

注：原文学习并翻译

**二〇二一年五月**

**论文的思想**

机器学习的任务大都是从训练样本中学习样本的分布，然后用学习到的分布在测试集上进行预测。本文提出一种框架从任务描述中进行学习，并构建了ZEST数据集，该数据集能用于从任务描述中学习的任务。

在读这篇论文之前，单纯的看标题，感觉是一篇不可思议的工作，直接从task的描述生成模型。但论文通篇读下来，发现论文做的数据集和QA任务的数据集基本差不多，只不过作者给数据集中的成分起了和QA数据集不同的名字，并且提出了针对“task descriptions to model”的新的评测指标。下面本博客从数据集制作、baseline两个方面来介绍这篇论文。

**摘要**

通常，机器学习系统通过训练数千个示例来解决新任务。 相比之下，人类可以通过阅读一些说明来解决新任务，例如可能需要一两个例子。 为了缩小这一差距，我们引入了一个用于开发NLP系统的框架，该框架可以在阅读新任务说明后综合解决该领域的先前工作，从而解决新任务。 我们使用一个新的英语数据集ZEST实例化了此框架，该数据集用于针对未看到的任务进行面向任务的评估。 将任务描述表述为问题，我们确保每个任务描述都具有足够的通用性，可以应用于许多可能的输入，从而全面评估模型解决每个任务的能力。 此外，数据集的结构可测试特定类型的系统概括。 我们发现，最新的T5模型在ZEST上得分达到12％，这对NLP研究人员提出了重大挑战。

**1简介**

如今，在受监管的NLP中，占主导地位的范例是从示例中学习，在该示例中，使用大量特定于任务的输入/输出对来训练机器学习算法。 相反，人类通过阅读描述学习执行相同的任务，然后他们能够以零击方式执行任务-实际上，这就是构建基于人群的NLP数据集的方式。 在本文中，我们认为以这种方式从任务描述中学习是通用NLP系统的必要属性，并且我们将其建议为训练和测试NLP系统的新范例。

尽管取得了这些进展，但从示例中学到的东西还是有很多严重的问题。

人们可能希望使用通用NLP系统解决几乎无限数量的任务。 通过阅读描述而不是观察示例来学习解决这些任务将解决必须为每个语言任务创建训练集的问题。 这样的系统也更适合其他领域的从业者和领域专家使用，他们可以描述他们的任务并解决它们，从而开辟新的研究途径，而在这种情况下，收集培训数据非常昂贵或不可行。

此外，我们发现，由于记住了培训分布中不感兴趣的方面，当前的监督学习技术在一定程度上取得了成功。仅从描述中教系统学习任务就可以减轻这些偏见，因为学习新任务不需要新的训练数据。

在本文中，我们综合了NLP中零射击学习的现有方法，并为思考零射击预测问题提供了正式的框架。 我们表明，以前的零射方法在应用范围和评估的严格性上都受到限制。 例如，尽管先前的工作使用零镜头预测进行文本分类，实体键入和关系提取，但我们将其推向了更为复杂的插槽填充任务。

我们在英语语言数据集ZEST（任务描述中的ZEro Shot学习）中实例化形式，其格式类似于阅读理解数据集，因为我们将后期的任务描述视为问题，并将其与文本段落配对。 我们选择这种格式，因为它提供了一种自然的方式来众包数据。

该零射数据集与典型的阅读理解数据集不同，但是，每个任务描述都与二十个不同的步骤配对，并且我们评估模型解决任务的能力，而不仅仅是为单个问题提供正确的答案 （问题，段落）对。 也就是说，给定一个问题，模型会产生一些决策函数f，正是这个函数我们对许多不同的输入进行了综合评估。 我们还特别注意完全选择轴，以评估模型针对不同类型任务描述的通用性，以特定方式更改任务描述，以系统地推动该领域朝着更有趣，更复杂的任务描述发展。

我们基于最新的最先进序列到序列体系结构评估模型，这似乎最适合在这种情况下进行零击预测的任务。 我们发现，基于T5的最佳模型（Raffel等人，2019）在此数据上仅获得12％的评分，与我们对42％的人类绩效估计相比有很大差距。 从复杂的任务描述中进行零镜头学习仍然是当前NLP系统面临的重大挑战。

**2从任务描述中学习**

本节描述了用于对未完成的任务进行零镜头泛化的框架，并将其延后至以前的工作。

**2.1从示例中学习**

考虑监督学习设置2，其中目标是学习函数****特定任务的可训练参数θ。我们将任务定义为：

•定义允许输入****，输出****的集合。

•概率分布****。

例如，在文本分类中，****是自然语言文本，而****是来自****类之一的分类标签。 在单个任务设置中，通过收集标记的示例****的数据集来学习函数****。****从****采样。 我们称其为“案例学习”。 至关重要的是，一旦构建了D，就将基础任务定义丢弃，并假设已在标记的****对中捕获该任务定义。

有很多方法可以从****进行采样以创建数据集。在诸如语言建模之类的情况下，其中****由一组规则定义的一种方法只是将规则应用于原始文本。 另一种流行的方法是使用人工注释。

在这种情况下，最常见的策略将****分解，通过某种方法从****采样（例如，从感兴趣的域中收集文本），并使用 自然语言任务描述****来描述****。 向人类注释者显示了该描述，他们使用注释为给定的****计算****。

**2.2从任务描述中学习**

从示例中学习的最大缺点是，每个新任务都需要收集一个新的数据集，以学习该任务的新函数****。

尽管任务定义包含了人类解决任务所需的所有信息，但这种方法在构建标记的数据集后也丢弃了任务定义。

此外，它使任务在测试时保持不变。我们提出的框架（我们称为“从任务描述中学习”）消除了这些限制。 首先，除了输入x之外，我们还提供了对模型的自然语言描述，而不是放弃任务定义。 其次，通过为模型提供任务描述，我们希望它能够以零射的方式推广到测试时看不见的任务。

这些修改将学习问题从拟合示例学习方法中的概率分布转移到理解任务描述的语义，以便将其应用于任务描述方法学习中的给定输入。 成功地构建以这种方式执行的模型将打开大量的NLP应用程序，从而可以通过用自然语言描述所需的输出来简化NLP系统的构建。

与从示例中学习相反，我们假设对于M个不同的任务观察到任务描述****，并且这些任务中的每个任务都有一些N个观察到的****对。

**3实例化框架**

第2节显示了用于培训和测试通用系统的框架，该系统可以执行看不见的NLP任务。 该框架中的理想系统将能够读取GLUE套件中任务的描述（Wang等人，2019），并且无需额外培训即可表现良好。 但是，这个目标远远超出了当今模型的当前功能。 为了取得进展，我们必须将问题分解为可管理的步骤。 在本节中，我们概述了我们为合理的，以NLP为重点的数据集所设想的范围，该数据集可以从任务描述中推动当前的学习状态，而又不会面临挑战而无法克服的挑战。 秒 图4描述了ZEST的数据收集过程，ZEST是根据该范围而建立的新的英语基准。

为了定义范围，我们首先考虑应用程序的类型，该模型可以成功地从任务描述中学习。 如今，构建NLP应用程序的最大瓶颈是收集标记数据。 我们的框架将消除此步骤，从而可以构建临时的NLP应用程序以轻松地从语料库中过滤，分类或提取结构化信息。

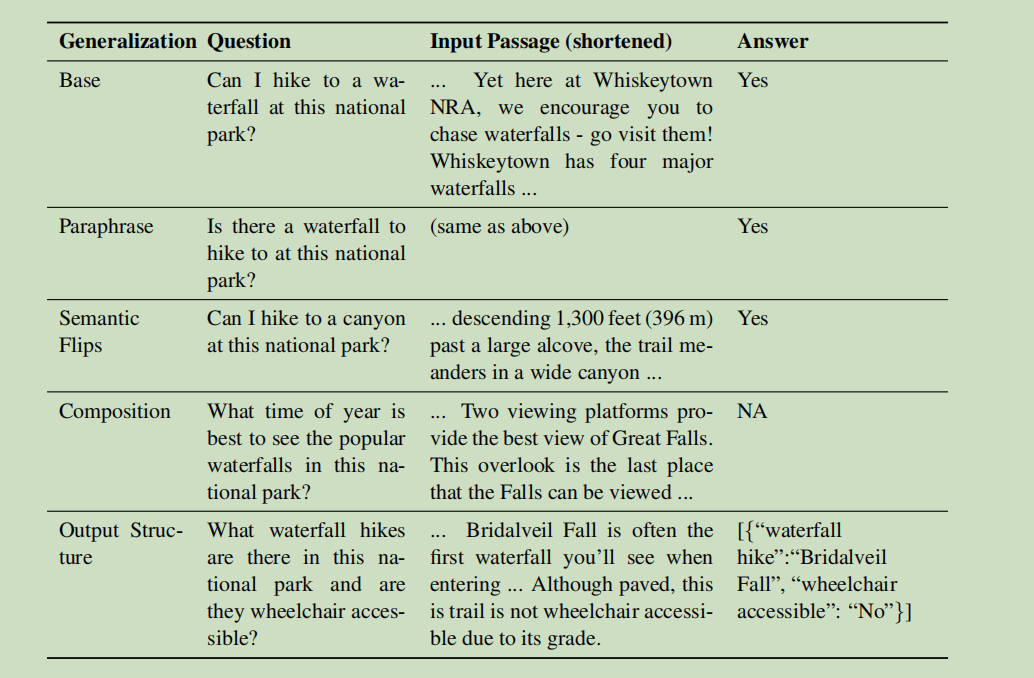
因此，重要的是在基准数据集中包含此类系统的基本任务构建块示例：分类，类型化实体提取和关系提取。 这样，它将统一零重点NLP（第2.4节）中的先前工作，该工作只关注单个任务，并且需要一个模型才能在测试时处理这些任务中的任何一个，而不是单独进行。 每个任务的模型。

更具体地说，由于每个任务τ定义了一组允许的输出****，只要在任务描述中指定了输出集，我们就可以在单个数据集中混合多个输出集****。ZEST包括最常见的输出集：离散类，输入中（可选）类型化跨度的列表以及跨度之间的关系。 表1中显示了每个示例，从任务描述中可以清楚地看到预期的输出Y。

此外，我们还包括NA输出（Ra jpurkar等人，2018），这表明在输入x的情况下无法解决任务。 例如，如果任务要求模型提取露营地名称，但输入是无关的新闻文章，则输出为NA。 在实际环境中，要合理地识别无法解决的任务非常重要，因为在这种情况下，无法合理预期每个可能的任务都可以通过每个可能的输入来解决。

### 4、数据集制作

下面给出数据集的一个表格



最左边一列是task的类别，第二列的Question，相当于输入x，第三列的Input Passage，相当于description，第4列的Answer相当于输出y。

制作数据集的时候，base类型task的description是从U.S. presidents, dog breeds, U.S. national parks三个领域收集的文章，Question是根据这些文章内容，也就是Input Passage提出的问题，这里的问题是人工提出的。下面四种类型都是根据Base的问题变化而来的。

Paraphrase：是在Base的基础上，Question 更改措辞方式，但是不更改Question表达的意思。

Semantic Flips：是在Base的基础上，对Question做尽可能小的改动，但是改动改变了Question表达的意思

Composiyion: 使用and、or等连接词，将base任务组成新的任务

output Structure：输出不是简单的分类，而是有结构的输出，相当于做了实体抽取和关系的抽取

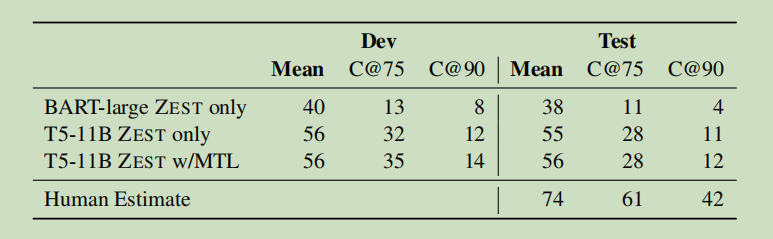
**5、Baseline**

**5.1 评价指标**

因为本文提出的数据是针对“task description”的数据集，一般使用的指标F1值不适合该数据集。因此本文提出了一个新的评价指标C@T。举个例子来说明，对于Base类的任务，单个的任务可以用accuracy来评价，那么C@90是0.7就代表模型在70%的Base类任务上，准确率可以达到90%。这里准确率可以根据任务改为F1等其他指标。

**5.2 Basline**

本文使用预训练模型T5和BART作为baseline，采用F1值作为单个任务的评价指标，下面是模型结果



可以看出T5的性能要比BART的好，但是仍然都远低于人类的水平。本文还用其他的QA数据集对T5的模型进行训练，结果是w/MTL那行，可以看出使用额外的训练数据并没有让模型性能提升太多

**6、个人想法**

这篇论文提出的数据集和QA任务的数据集几乎一样，只不过换了一种描述方法，并且使用了一个新的评价指标，但本质上来讲，还是QA的任务。所以这里的task description和我读论文前预想的不太一样。

这篇论文提到，通过对任务描述的学习，进行zero-shot学习，再QA任务中，input passage和question都是输入，在这篇论文中，input passage成了description，然后作者就把在该数据集上的模型学习当作zero-shot（这里也可能是我理解偏差）。作者为了体现zero-shot的特点，baseline都选取了预训练模型，这一点有待商榷。