一、题目

使用 TPU 检测多语言文本中的矛盾和蕴涵

二、组员

方紫婵 2006500012
徐瑞禧 2006500017
许中兰 2006500040

三、kaggle 链接：

https://www.kaggle.com/competitions/contradictory-my-dear-watson

四、科学意义

1. 背景

“...当你消除了不可能的事情时， 无论剩下的，无论多么不可能，都必须是真相” ——亚瑟·柯南·道尔爵士

我们的大脑处理这样的句子的含义相当快。

我们能够推测：

- 有些事情是真实的：“你可以通过淘汰的过程找到正确的答案。
- 其他可能有道理的人：“不可能的想法并非不可能！
- 有些说法显然是矛盾的：“你排除为不可能的事情才是真相所在。

自然语言处理（NLP）在过去几年中变得越来越复杂。机器学习模型可处理问答、文本提取、句子生成和许多其他复杂任务。但是，机器能否确定句子之间的关系？如果 NLP 可以在句子之间应用，这可能会对事实检查，识别假新闻，分析文本等提供重要帮助。此外，在自然语言处理领域，探究多语言文本中的逻辑关系也是一个重要的研究方向，这样的技术突破将有助于实现文本自动化处理和智能化交互，具有重要的科学意义和现实意义。

2. 选题原因

在多语言文本生成的过程中，可能会出现翻译不准确、语义不对等问题，这就涉及到多语言文本中存在的矛盾和蕴涵等逻辑关系。于是我们想使用 TPU 技术检测多语言文本中的矛盾和蕴涵，用以提高文本生成的质量和精度，从而实现自然语言处理领域的自动化处理和智能化交互。

五、科学问题

使用 TPU 技术检测多语言文本中的矛盾和蕴涵，用以提高文本生成的质量和精度，从而实现自然语言处理领域的自动化处理和智能化交互。

六、研究内容

1. 数据集介绍

①介绍

在本次入门竞赛中，我们将成对的句子（由前提和假设组成）分为三类（蕴含 entailment 、中立 neutral、矛盾 contradiction）。

让我们看一下以下前提下每个案例的示例：

他来了，他打开了门，我记得回头看他脸上的表情，我看得出他很失望。

- 假设 1：光是看他进门时脸上的表情，我就知道他很失望。
  根据前提中的信息，我们知道这是真的。因此，这对是蕴涵（entailment）
- 假设 2：他试图不让我们感到内疚，但我们知道我们给他带来了麻烦。
  这很可能是正确的，但我们不能根据前提中的信息得出结论。所以，这种关系是中性（neutral）
- 假设 3：他是如此兴奋和喜悦，以至于他几乎把门从门框上撞了下来。
  我们知道这不是真的，因为它与前提所说的完全相反。因此，这对是矛盾（contradiction）

此数据集包含 15 种不同语言的前提-假设对，包括：阿拉伯语、保加利亚语、中文、德语、希腊语、英语、西班牙语、法语、印地语、俄语、斯瓦希里语、泰语、土耳其语、乌尔都语和越南语。

②文件

train.csv：此文件包含 ID，前提、假设和标签，以及文本的语言及其两个字母的缩写。

test.csv：此文件包含 ID、前提、假设、语言和语言缩写，不带标签。

sample_submission.csv：这是一个正确格式的示例提交文件。

id:每个样本的唯一标识符

label:0 表示蕴涵，1 表示中性，2 表示矛盾

3. 计划使用的算法介绍

（1）基于相似度方法：将对象表示为向量或矩阵，使用各种距离度量方法来计算前提和假设之间的相似度，以此判断是否构成蕴含关系。

（2）基于逻辑演算方法：将文本表示成数学逻辑表达式，构成事实集合，利用逻辑推理规则判断能否根据前提推出假设。

（3）基于深度学习模型：LSTM+attention

LSTM 是一种用于处理序列数据的循环神经网络。其核心的结构单元称为记忆细胞，可以存储和检索过去的信息，并根据需要添加或删除信息。每个记忆细胞包含三个门：输入门、遗忘门和输出门。这些门可以控制信息的流动，从而实现对记忆细胞内部状态的修改和选择性输出。

Attention 机制核心目标是从众多信息中选择出对当前任务目标更关键的信息，在此实验中会把此模型看作是输出和输入的对齐模型。其重要框架中 Encoder 对输入句子进行编码，转换为中间语义；Decoder 则根据中间语义和之前已经生成的历史信息来生成目标值。

七、预期目标（预期达到的目标介绍）

目标是预测给定的假设是否与其前提相关，或者这两个假设是否都不成立。对于测试集的每个样本，必须预测变量的 0、1、2 值，对应蕴含、中性、矛盾。