中国科学院大学
University of Chinese Academy of Sciences

自 然 语 言 处 理

技术报告

班级：2023 春季自然语言处理班

组号：--

姓名：唐嘉良

学号：2020K8009907032

报告主题：译文质量评估——BLEU

2023 年 7 月 3 日

一．报告摘要

本次实验中，我基于机器翻译的常用译文评估方法 BLEU，选取 WMT-18 的中英新闻领域测试集部分平行句对，编写针对中译英的 BLEU 计算程序，挑选百度翻译、谷歌翻译和 ChatGPT 翻译三种翻译引擎进行评测，并对计算数据进行了对比和关联分析。得到的主要结论为谷歌翻译的整体性能最为出众，但译文质量最不稳定；相反，尽管百度翻译和 ChatGPT 翻译在平均翻译水平上不如谷歌翻译，但在翻译任务上的翻译质量具备较好的稳定性。

二．实验过程与方法

1.数据集准备与实验流程

在 WMT-18 的中英新闻领域测试集中随机挑选 30 个中英文平行句对，将参考译文句子存放在 refx.txt 文件中（x=1，2，...，30），其对应的待评估译文则存放在 sourcex.txt 文件。

针对每个翻译引擎，我们将全部 30 个句子在 WMT-18 中对应的的中文版本作为输入，得到 30 个英语翻译句子文本并全部存入 sourcex.txt，使用编写的一键测试脚本文件计算每个句子的 BLEU 值，统计百度翻译、谷歌翻译和 ChatGPT 翻译三种引擎所得到的结果。脚本文件如下：

run.sh

```
for i in {1..30..1}
do
    python ./bleu.py source$i.txt ref$i.txt
done
```

其中命令参数 source$i.txt 为待评测文件，ref$i.txt 为参考译文。

2.BLEU 计算

我们采用如下的 BLEU 计算方法。提取待评测句子中的 n-gram 词组，进行在参考译文中出现频率的统计，记录频率并维护短译文惩罚因子。随后计算频率的几何平均，考虑惩罚因子得到最终 BLEU 值，计算公式如下：

$$\mathbf{BLEU} = \mathbf{BP} \times \sqrt[\frac{1}{n}]{\prod_{i=1}^{n} p_i}$$

其中 **BP** 为惩罚因子。

此外，我们还进行了概率平滑处理，避免奇异点将整体 BLEU 值归零。由于输入可能是多重集，我们取待评测句子和参考句子出现频次的较小值作为词组真实出现频次参与计算。

作为对比，我们还采取了另外一种 BLEU 计算方法，该方法的不同之处在于对频率进行了加权对数化并采用 exp 指数正规化评估值，计算公式如下：

$$\mathbf{BLEU} = \mathbf{BP} \times \exp\left(\sum_{i=1}^{n} w_i \log p_i\right)$$

具体实现如下：

```python
def BLEU(source, ref):
    precisions = []
    gram_n = [1, 2, 3, 4] # 考虑1-gram到4-gram
    for i in gram_n:
        pr, penalty = count_ngram(source, ref, i)
        # precisions.append(pr)
        precisions.append((float)(1/4)*math.log(pr))

    # 采用几何平均
    # bleu = (reduce(operator.mul, precisions)) ** (1.0 / len(precisions)) * penalty
    bleu = math.exp((reduce(operator.add, precisions))) * penalty
    return bleu
```

```python
def penalty(c, r):
    if c > r:
        bp = 1
    else:
        bp = math.exp(1-(float(r)/c))
    return bp
```

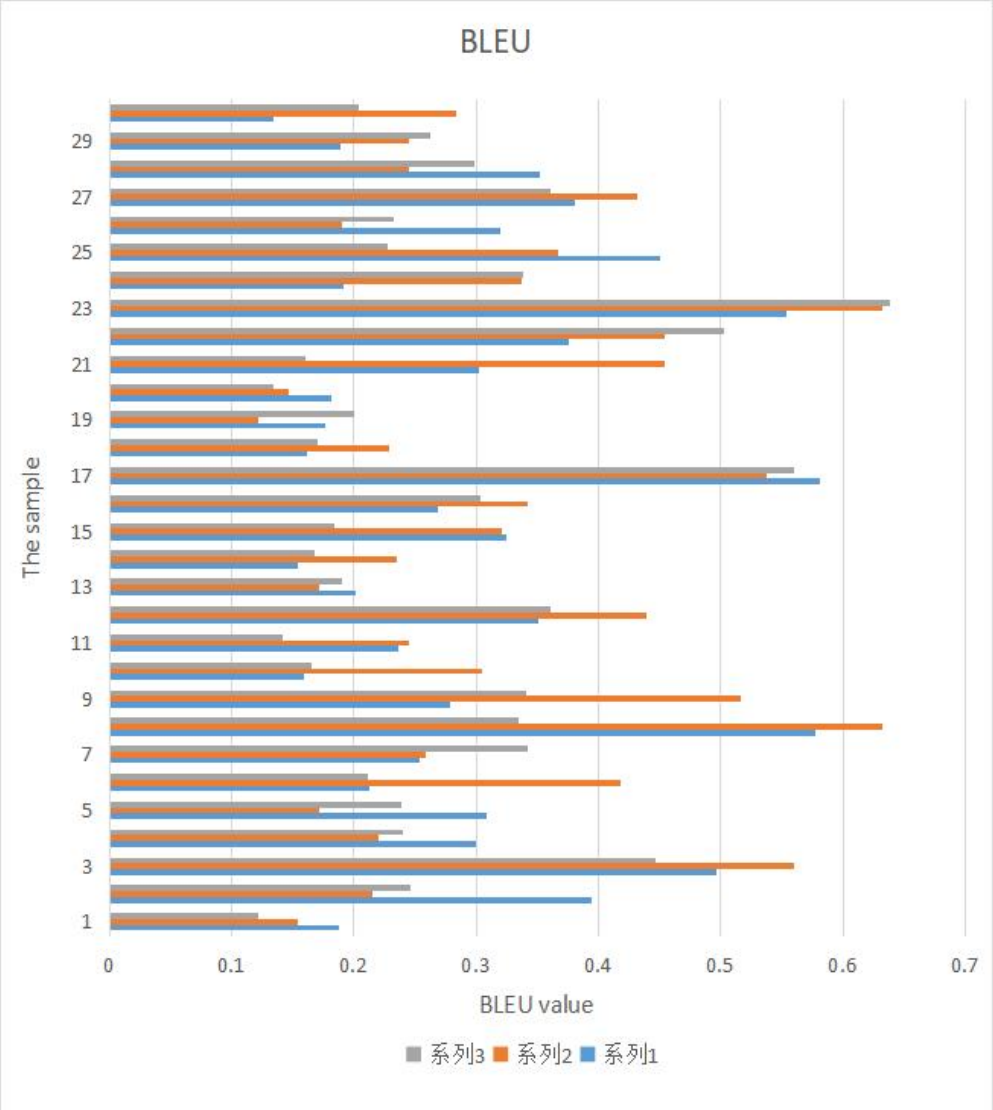
3.平均值、方差分析与翻译引擎性能对比

一方面，我们对百度翻译、谷歌翻译和 ChatGPT 翻译引擎分析 BLEU 值的均值和方差，对比三者在中译英领域的译文质量。另一方面，我们使用 ChatGPT 翻译，针对同一数据集采取两种 BLEU 计算方法，对比其计算结果的差异。

三．结果对比与分析

1.三种翻译引擎计算结果与性能对比

三种翻译引擎的 BLEU 值计算结果如下（采取几何均值的 BLEU 计算方法）：



其中系列 1、系列 2、系列 3 分别对应百度翻译、谷歌翻译和 ChatGPT 翻译。

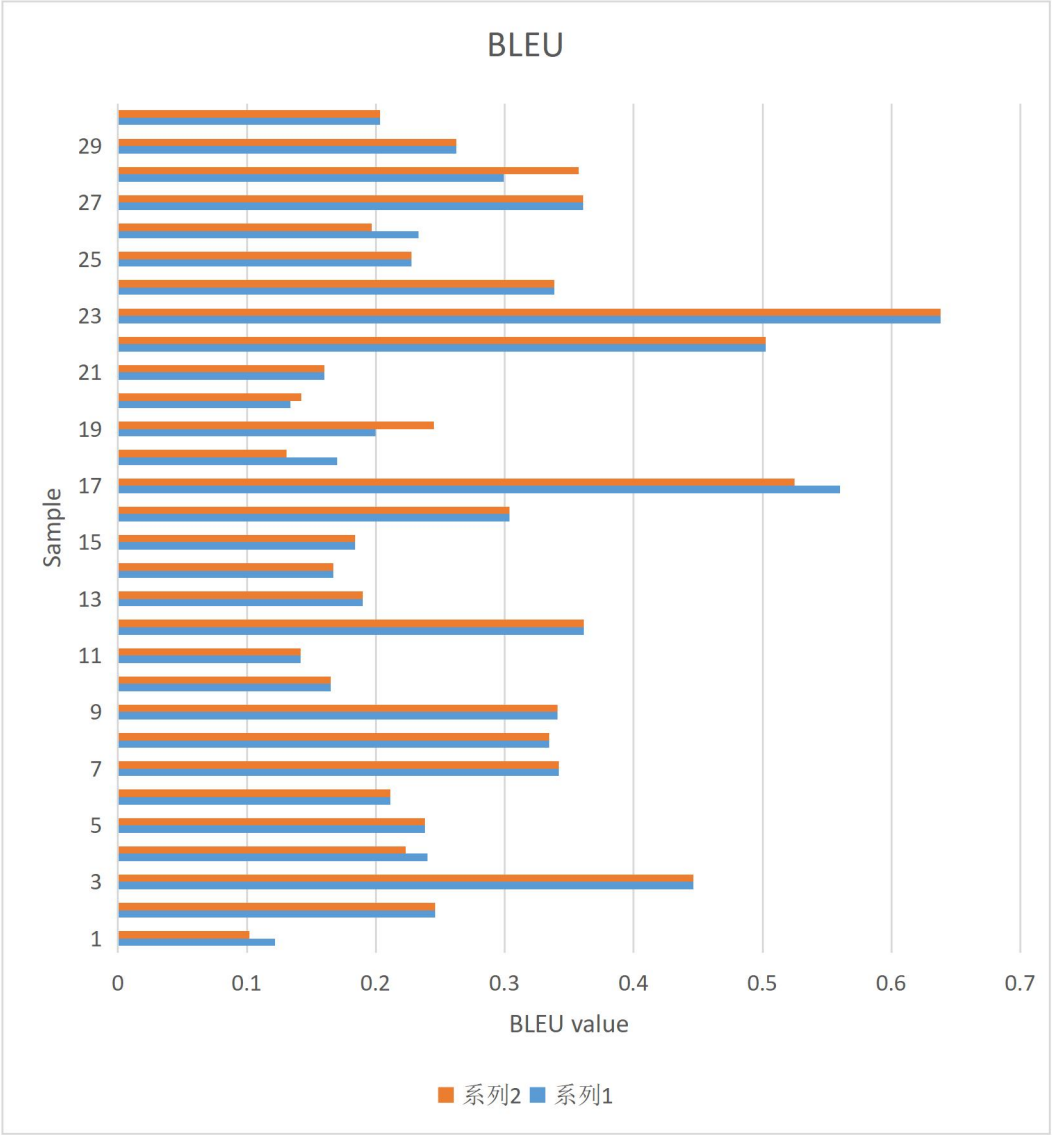
对结果进行均值和方差分析如下：

	百度翻译	谷歌翻译	ChatGPT 翻译
BLEU 均值	0.3017	0.3292	0.2775
BLEU 方差	0.017	0.022	0.016

从上述结果中可以看出：谷歌翻译 BLEU 均值最高，其次是百度翻译，而 ChatGPT 翻译的平均水平最低。另一方面，通过方差结果可以看出 ChatGPT 翻译的性能稳定性最好，百度翻译的稳定性与之相差无几，而平均性能最好的谷歌翻译却最不稳定。推测谷歌翻译可能采取了某种对稳定性有破坏的翻译技术，牺牲一定的稳定性以获取整体更好的翻译表现。

2.两种不同 BLEU 计算方法与结果对比

对于 ChatGPT 翻译，我们采取加权对数方法来计算 BLEU，与前文方法所得结果对比如下：



可以看到,两种不同的BLEU计算方法在同一批样例上所得到的计算结果整体相差很小。由此可见，两种BLEU计算方法具有相似的评估译文的能力，结果具有合理性。这符合我们的预期。

3.实验特征与缺陷分析

观察测试样例，容易发现短句的BLEU值时常不高，这可能是因为词组出现频率受数据集中“坏词”影响较大，也就是说翻译引擎得到的与参考译文不同的词组会对相邻区间的词组统计造成影响，并且由于整体区间较小，这种影响被放大。

另一方面，由于我们选取的数据集是专业化数据集，其在不同引擎所体现的能力无法非常精确地反映翻译引擎的总体能力，而受其专业化翻译能力影响较大。此外，数据集规模仅有30个平行句对，因此实验结果具有一定偶然性。