期末作业实验报告：更好的日-英翻译器

郭晟毓 赵亦阳

January 12, 2025

1 实验概述

我们在作业一中的日-英翻译任务的基础上，通过增大数据量、改变模型结构等方法训练了一个更好的日-英机器翻译模型，并以之为基准进行了一系列对比实验探究不同因素对模型表现的影响。

2 实验动机

在作业一中，我们在给定数据集（大小约为六万条平行语料对，以下称为task_1数据集）上训练了一个LSTM模型。然而，无论是从测试集上的各项指标，还是从模型翻译样例文本的直观结果来看，这都是一个表现很差的翻译模型。具体结果如下：

Performance: LSTM trained on task_1

Loss: 1.75
BLEU: 12
私の名前は爱です。 → My name is my name.
昨日はお肉を食べませんでした。 → I didn't eat meat yesterday.
いただきますよう。 → I'll do this.
秋は好きです。 → I like green.
おはようございます。 → Good morning.
ドアを闭める。 → Lock the door.
ドアが闭まる。 → The door is down.
ドアが闭められる。 → You can to open door door
英语を教えてくれてありがとう。 → Thank you for me me me
英语を教えていただけませんか。 → Can you teach English speak?

为了得到一个更好的日-英翻译模型，我们收集了更多的日-英翻译数据，并且使用了基于Transformer结构的模型。在此基础上，我们用控制变量的办法探究不同因素的模型翻译表现的影响，并根据实验的结果，进一步改进翻译模型。

3 实验过程

3.1 准备数据

除了作业一中的task_1数据集外，本次实验中还使用了Opus[1]与Tatoeba[2]两个数据集。其中，Opus 是一个全面的开源平行语料库集合，汇集了官方文件、维基百科、电影字幕等多来源的平行文本；Tatoeba 是由成千上万志愿者维护的大规模句子与翻译数据库。我们将大小为约一百万条平行语料对的Opus用于训练和测试，对于Tatoeba，我们只使用其测试集（大小约为一万条平行语料对）用于模型的测试。

3.2 搭建分词器

使用`Helsinki-NLP/opus-mt-ja-en`[3]中预训练的分词器。

3.3 搭建模型

对于序列到序列的机器翻译任务，基于Encoder-Decoder架构的`MarianMTModel`应该比较合适。部分参数如Listing 1。

Listing 1 Model Configuration

```
MarianConfig {
  "d_model": 512,
  "decoder_attention_heads": 8,
  "decoder_ffn_dim": 2048,
  "decoder_layers": 6,
  "encoder_attention_heads": 8,
  "encoder_ffn_dim": 2048,
  "encoder_layers": 6,
  "max_position_embeddings": 512,
  "num_hidden_layers": 6,
  "vocab_size": 60716
}
```

3.4 训练模型

部分训练参数如Listing 2。

3.5 评估模型

我们从测试集上的指标和对样例文本的翻译结果两个方面评估翻译模型的表现。我们使用的测试集包括task_1、Opus和Tatoeba，在每个测试集上分别计算交叉熵损失和BLEU。样例输入则包括了来自日语教材、文学作品和我们自己的造句的日文文本，涵盖了不同的长短、复杂程度以及侧重点。

Listing 2 Training Arguments

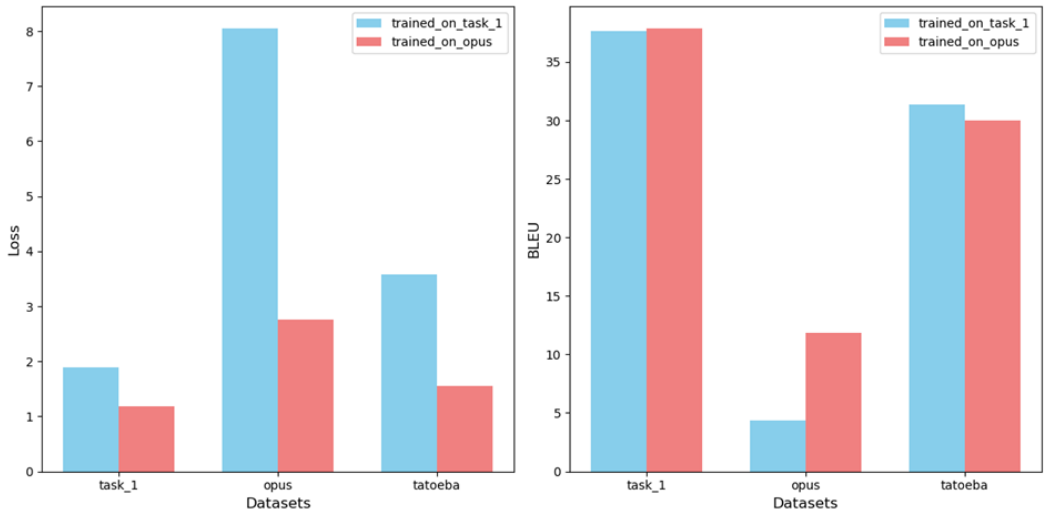
```
Seq2SeqTrainingArguments(
    learning_rate=2e-5,
    per_device_train_batch_size=32,
    per_device_eval_batch_size=64,
    weight_decay=0.01,
    num_train_epochs=20,
    predict_with_generate=True,
    generation_num_beams=6,
    fp16=True,
    generation_max_length=128
)
```

4 实验结果与分析

4.1 在不同数据集上训练至收敛

在这部分实验中，我们分别在 assignment 1 提供的日英训练集、opus训练集，以及二者的合并训练集上训练模型到收敛，然后分别在 assignment 1 中的日英测试集、opus测试集与tatoeba测试集上进行验证。

最初，我们只尝试在两个数据集上分别训练，本意是为了对比观察opus数据集的质量，得到如下图所示的实验结果。



同时我们收集了一些翻译结果，用于示意数据集性能：

Performance: Marian trained on task_1

私の名前は愛です。 → My name is love.
昨日はお肉を食べませんでした。 → I didn't eat meat yesterday.
いただきますよう。 → I envy you.
秋は好きです。 → Do favouite is here.
おはようございます。 → Good morning.

Performance: Marian trained on opus

私の名前は愛です。 → My name is love.
昨日はお肉を食べませんでした。 → I didn't eat meat yesterday.
いただきますよう。 → Let's eat.
秋は好きです。 → I like autumn.
おはようございます。 → Good morning.

两相比较之下，可以看出在opus数据集上训练得到的模型，不仅能够正确翻译在训练集中有多个对应词的文段（如"秋"对应"autumn"和"fall"，"爱"对应"like"和"love"），也能够正确翻译缺乏上下文、缺乏主语、语义多样的日常用语（如"いただきますよう。"）。这佐证了扩大数据集的效果。

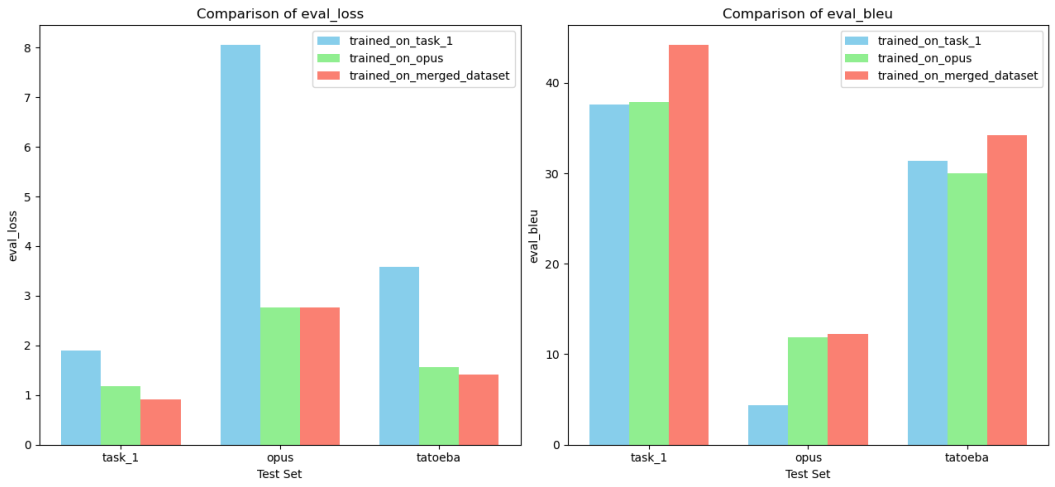
然而，对于语法结构较为复杂的句子，即使是在opus上训练的模型也不能做到完全翻译正确，如下句：

A failed translation of a double negative sentence by Marian trained on opus

日本には社会保険制度があり、国民は必ず公的な医疗保险に入らなくてはいけない。

There is a social insurance system in Japan, and the people must never enter a public medical insurance policy.

这里模型没能识别"なくてはいけない"这个意为"不得不"的文法片段，而只识别了表示否定的"ない"。这说明简单增大数据规模，并不能保证包罗日语中的重要特征。为此我们决定在数据集中专门增加强调语法结构的翻译条目，用以补足opus数据中存在的语法不严谨的问题。恰好在这方面assignment 1 的数据集表现较好，因此我们尝试将两个数据集拼合起来，重新训练至收敛，得到结果如下图所示。



A success translation of a double negative sentence by Marian trained on merged dataset

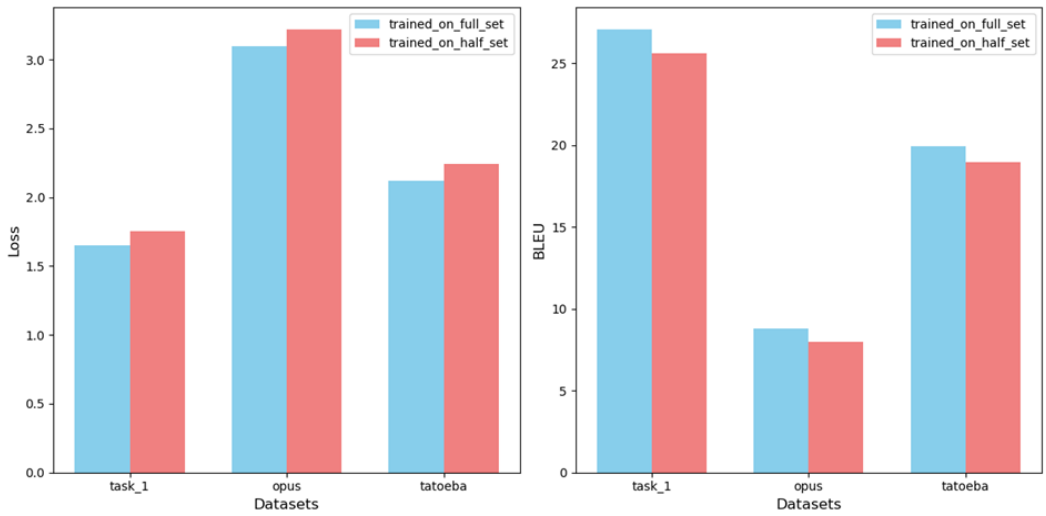
日本には社会保険制度があり、国民は必ず公的な医疗保险に入らなくてはいけない。

Japan has a social insurance system, and the people have to enter a public medical insurance policy.

由该图以及翻译样例可以看出，新模型的能力相比于单独在两个数据集上训练的能力都有所提升。

4.2 在同一数据集的不同部分中训练相同FLOPs

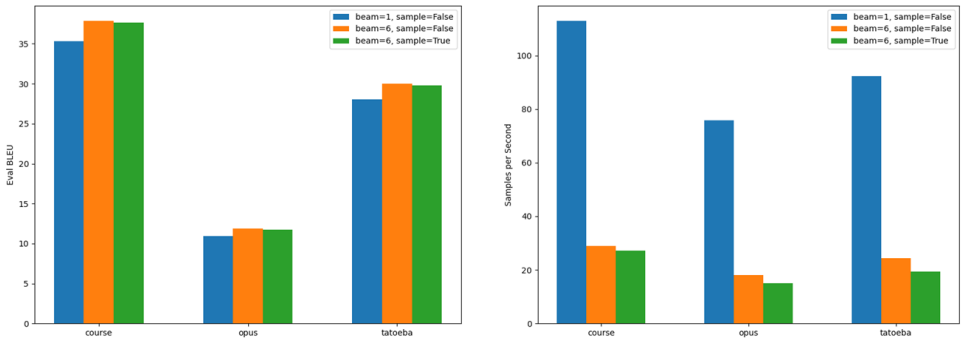
在这部分实验中，我们受到 scaling law 的启发，探究了数据集大小对模型能力的影响，从而决定如何最优化计算资源。通过控制训练过程的FLOPs不变，我们测试了在opus训练集全集以及一半训练集上训练 120000 batch 得到的模型分别在 assignment 1 测试集、opus测试集、tatoeba测试集上的表现，得到结果如下图所示。



由图可知，在每种测试集下，更见多识广的数据集总是水平更高，这说明当采取Marian为模型结构时，应当尽可能增大数据集规模。这与上一部分中的增加语法数据集达到了相乘效果。

4.3 采用不同decoding策略训练

在这部分实验中，我们针对 beam search 的 beam 数以及是否采样进行调整，探究如何权衡翻译效果与翻译时长。得到的实验结果如下图所示。



由图可知，单就模型效果而言，采用束搜索优于采用贪婪搜索、不采样优于采样；而单就翻译时长而言，贪婪搜索最快，不采样的束搜索次之，采样的束搜索最慢。不过对于单条翻译而言，上述时长的差别微乎其微，因此应当选择不采样的束搜索作为模型策略。

5 总结

本次实验中，我们训练了一个日-英翻译模型，并进行了一系列对比实验。一个可能比较有启发性的发现是，少量（六万条相对于一百万条）较高质量数据（task_1相对于Opus）的引入可以较大程度地提升模型表现。`inference.ipynb`中包含了更多的样例文本及相应的翻译结果。

References

[1] Jörg Tiedemann. Parallel data, tools and interfaces in OPUS. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA).

[2] Jörg Tiedemann. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online, November 2020. Association for Computational Linguistics.

[3] Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato Yves Scherrer, Raul Vazquez, and Sami Virpioja. Democratizing neural machine translation with OPUS-MT. *Language Resources and Evaluation*, (58):713–755, 2023.