220230944151 严京辉

Task1-Neural Machine Translation with RNNs

(g) (8 points) (written) The generate_sent_masks() function in nmt_model.py produces a tensor called enc_masks. It has shape (batch size, max source sentence length) and contains 1s in positions corresponding to 'pad' tokens in the input, and 0s for non-pad tokens. Look at how the masks are used during the attention computation in the step() function (lines 311-312). First explain (in around three sentences) what effect the masks have on the entire attention computation. Then explain (in one or two sentences) why it is necessary to use the masks in this way

enc_masks 在 attention 计算中的作用是屏蔽源句子中的填充标记。它将输入中的填充标记对应的位置设置为 1，非填充标记对应的位置设置为 0。通过将注意力权重与填充标记对应的位置设置为接近负无穷的小值，模型在注意力计算过程中有效地忽略了这些填充标记。它可以**防止模型关注填充标记。**填充标记没有带有任何有意义的信息，只是为了保证输入长度相同，因此通过屏蔽它们，模型可以更好地集中注意力在真实的输入标记上，提高翻译质量。

(h) (10 points) (written) Once your model is done training (this should take under 3 hours on the VM), execute the following command to test the model:

测试结果如下：

```
load test source sentences from [./br_chr_data/test.br]
load test target sentences from [./br_chr_data/test.chr]
load model from model.bin
Decoding: 100%|                                                                                | 2894/2894 [12:05<00:00,  3.99it/s]
Corpus BLEU: 54.393580840666345
```

(i) i. (5 points) Explain one advantage and one disadvantage of dot product attention compared to multiplicative attention. ii. (5 points) Explain one advantage and one disadvantage of additive attention compared to multiplicative attention.

点乘注意力：计算效率高，特别适用于较大的输入序列。但不具备学习参数，无法对不同位置之间的关联性进行灵活建模。

加法注意力：可以对输入进行非线性转换，提高模型的表达能力。但计算开销大，特别是在处理较大输入序列时不适合使用

Task2-Analyzing NMT Systems

(a) (15 points) In part 1, we modeled our NMT problem at a subword-level. That is, given a sentence in the source language, we looked up subword components from an embeddings matrix. Alternatively, we could have modeled the NMT problem at the word-level or char-level, by looking up whole words or the single char from the embeddings matrix. Try to analyze the advantages and disadvantages of these three methods, and explain which method you think is better, and provide reasons. (Hint: Consider dictionary size and sequence length.)

词级别建模：词级别的建模就是将语料分割成单个单个的词然后构成字典，通常来说它能保留完整的词语语义信息，有利于模型理解和生成更准确的翻译结果。但是如果对于词汇表中不存在的词语，无法找到其表示，这可能导致翻译错误或信息丢失。另外，对于某些语言比如中文，它的分词不像英文是天然按照空格进行划分，需要的额外的分词处理，但是中文的分词并不是简单的事情，会对算法提出挑战。

子词级别建模：这种方法允许我们分解词汇为更小的子词单元，从而捕捉更细微的语言信息。例如，将英文单词划分为前缀、词源、后缀等，每个子词都有其自身的含义。子词级建模的主要优势是它能够更好地处理罕见或词汇表外的单词，通过类似前缀+词源+后缀这样的组合去推理出在字典表中不存在的单词。这使得模型在处理未见过的单词（例如单复数、派生词）时具有更强的泛化能力。子词级的潜在缺点是，它可能会导致输入序列变得过长，因为单词被分解成了多个子词单位。此外，由于子词切分方式的不确定性，可能会导致模糊的表示，特别是对于某些语言或领域特定的词汇。

字符级别建模：字符级建模将每个字符作为划分单元，放到中文里就是单独的字。这种建模方法的优点就是对于任何语言的处理方法都比较统一，能够处理任意字符，包括未登录词或专有名词等。但是由于它是按字符划分的，因此在训练的时候，输入序列的长度会前所未有地大，导致极大的训练开销。同时由于划分的太过琐碎，导致模型很难捕捉到连缀起来的语义信息。

综上所述，我觉得方案的选取可能需要取决于具体的使用场景，如果你是英文这种天然带分割的语言，我觉得使用词级别的建模就挺好的。如果你是中文这种同时又有庞大的语料库，那么使用字符级别的建模也未尝不可。庞大的语料库也能帮助模型去理解字符连缀起来的语义信息。如果是比较居中的场景，比如有人工分词的小数据集之类的，可以使用子词级别的建模，这也是两种相权衡的建模方法。

(b) (15 points) What is the biggest difficulty in translating Braille into Chinese? (Hint: Braille is derived from Chinese pronunciation.)

将盲文翻译为中文时，面临的最大挑战在于汉字的同音异义现象。由于盲文是基于汉语拼音音节的，在没有足够上下文信息的情况下，很难准确地选择出正确的汉字。同时盲文的数据集也是个问题，模型很难通过有限的盲文翻译数据集去学习到汉语拼音在上下文中的语义信息。

(c) (15 points) One challenge of training successful NMT models is lack of parallel corpus, particularly for resource-scarce languages like Braille. One way to address this challenge is to pre-train the model using monolingual data. Train two language models of two source and target languages to initialize the NMT model, and then fine-tune the model using parallel data. How does monolingual training help in improving NMT performance with low-resource languages?

通过单语训练，NMT模型可以从大量单语数据中学习源语言和目标语言的语言特征、结构和规律。这有助于模型更好地理解和生成目标语言的翻译。尽管没有平行语料库，但单语数据包含了丰富的语义信息和上下文，可以帮助模型理解源语言句子的含义。在预训练阶段，模型可以通过单语数据学习如何表达相似的语义概念，并将这些知识应用于翻译任务中。通过单语数据进行预训练，可以有效扩充训练数据量，提供更多样化和丰富的语言信息，从而减轻数据稀缺性对NMT性能的影响。通过单语预训练，模型可以在高资源语言上学习通用语言知识，并将这些知识迁移到低资源语言的翻译任务中。这种迁移学习可以提高模型的泛化能力和翻译质量。

(d) (16 points) Here we present a series of errors we found in the outputs of our NMT model. For each example of a reference (i.e., '字') Chinese translation, and NMT (i.e., '自') translation, please:

1. Identify the error in the NMT translation.
2. Provide possible reason(s) why the model may have made the error (either due to a specific linguistic construct or a specific model limitation).
3. Describe one possible way we might alter the NMT system to fix the observed error. There are more than one possible fixes for an error. For example, it could be tweaking the size of the hidden layers or changing the attention mechanism.

Below are the translations that you should analyze as described above. Only analyze the underlined error in each sentence. Rest assured that you don’ t need to know Braille to answer these questions. You just need to know Chinese! If you want to learn about some braille corresponding to Chinese, you can use the translation function on this website.3

i. 同音异义

1. NMT 的错误是将其翻译为相同读音，但是表型不同的汉字，比如 '字' -> '自'
2. 可能的原因是中国盲文是基于中文读音的，所以这两个字写在盲文中就是相同的表示，导致机器错译，这种错译也是机器没有正确识别到上下文导致的
3. 增加与错译字相关的语料或者人工标注，让机器更好地理解该字相关的上下文。

ii. 重复出现

1. NMT 的错误是将 '抓' 字重复翻译四次

2. 可能的原因是模型的注意力机制可能在处理重复信息时出现问题，导致同一字符被多次生成
3. 通过调整模型的注意力机制，来更好地处理源句子中的重复信息。此外，可限制输出句子中单个字符可以重复的最大次数

iii. 读音相近

1. NMT 的错误是将读音相近的 '翻译' 误译成 '反映'
2. 可能的原因是字典的设置有问题，可能这两个 tokens 的向量表示太过于接近，模型在选择的时候没有选到正确的表示
3. 调整模型的字典对应 tokens 的向量表示，或者是将模型的自由度阈值调低点，这样模型就不能擅自做出推理导致误判

iv. 翻译不全

1. NMT 的错误是过早截断翻译
2. 猜测是模型可能在处理句子边界时出现问题，导致未能正确识别句子的结束位置。
3. 可以在数据集中人工引入 <bos> <eos> 虚标签来标记句子的开头和结尾，这样能帮助模型更加有效地识别句子

(e) (40 points) BLEU score is the most commonly used automatic evaluation metric for NMT systems. It is usually calculated across the entire test set, but here we will consider BLEU defined for a single example.4 (too long... omit...)

i. Which of the two NMT translations is considered the better translation according to the BLEU Score? Do you agree that it is the better translation?

对于翻译 c_1:

$p_1 = 0, p_2 = 0.5, p_3 = 0, p_4 = 0$

$len(c_1) = 18, len(r) = 22, BP = \exp(1 - \frac{len(r)}{len(c_1)}) = 0.672$

$BLEU = BP \times exp(\lambda_1 \log p_1 + \lambda_2 \log p_2) = 0.336$

对于翻译 c_2:

$p_1 = 0, p_2 = 0, p_3 = 0, p_4 = 0$

$len(c_2) = 20, len(r) = 22, BP = \exp(1 - \frac{len(r)}{len(c_2)}) = 0.367$

$BLEU = BP \times exp(\lambda_1 \log p_1 + \lambda_2 \log p_2) = 0$

从 BLEU 指标上来看，c_1 的更高，并且从语义上看，c_1 确实是更加好的翻译

ii. (10 points) Translation tasks in other languages can have multiple translations of a source language sentence, while there is only one correct translation for Braille to Chinese translation. Is BLEU suitable for evaluating Braille to Chinese translation tasks. Why?

不太适合。前面说到，盲文是根据汉语拼音来的，也就是说盲文是可以翻译成唯一的汉语拼音的。而 BLEU 指标是用来衡量机器翻译和多个译文的相似度的，所以不太适合这种单指标的任务。

iii. (10 points) Calculate BLEU based on chars or BLEU based on words to evaluate the translation task from Braille to Chinese. Which is better and explain your reasons.

我觉得采用基于单词的 BLEU 指标更为合适。中文词语的语义内涵很丰富，如果将词语生硬地拆解成简单的字，就会破坏其表达的准确性，导致错译。并且盲文中是存在空格进行分词的，也不存在传统中文语料的分词问题，使用基于单词的 BLEU 指标更加妥当。

iv. (10 points) List two advantages and two disadvantages of BLEU, compared to human evaluation, as an evaluation metric for Machine Translation.

优点：

1. 量化的指标能够让机器理解，同时让人对翻译结果有直观的理解。
2. 其计算效率也较高，且认可度高，可避免繁琐的人工评估，同时给出客观结果。

缺点：

1. 它是基于 n-gram 来进行匹配的，但是有些语言，单词用 n-gram 语法很难把握住整体的句子意思，特别是对于表达效率较低的语言，比如日语。
2. 它也无法考虑语序的差异，并且对长文本的评估存在偏好，因为长文本中匹配的 n-gram 较多，容易获得较高的BLEU分数，而短文本则可能受到惩罚。