[2022.12.20]

自然语言处理导论

机器翻译研究报告

岳禹彤

2000012952

目录

一、摘要

本实验利用 HuggingFace Transformers 工具包进行机器翻译，此工具包基于 transformer 模型。本实验用三个模型对两个不同的数据集进行了机器翻译，并利用 BLEU 评估了模型生成句子和实际句子的差异。最后对翻译结果进行了分析，并总结了翻译过程中出现的问题。

二、实验目的

学习机器翻译软件包的使用，调研工具包所依赖的相关技术，在英文-中文语言对上评价、对比、分析现有模型的翻译结果。利用相关工具在给定的数据集上进行测试并汇报结果，评价、对比、分析旨在揭示现有机器翻译工具包的优缺点。

三、工具以及使用方法

Python3.8.3
HuggingFace Transformers 工具包，它提供了在中文英文数据上训练好的模型。只需要简单地调用封装好的库函数即可。

我们选择 transformer 里的模型有 liam168/trans-opus-mt-en-zh 、Helsinki-NLP/opus-mt-en-zh、BubbleSheep/Hgn_trans_en2zh。

```
from transformers import pipeline

translator = pipeline("translation_en_to_zh",model='Helsinki-NLP/opus-mt-en-zh')
```

```
from transformers import AutoModelForSeq2SeqLM,AutoTokenizer,pipeline

mode_name = 'trans-opus-mt-en-zh'

model = AutoModelForSeq2SeqLM.from_pretrained(mode_name)

tokenizer = AutoTokenizer.from_pretrained(mode_name)

translation = pipeline("translation_en_to_zh", model=model, tokenizer=tokenizer)
```

```
from transformers import AutoTokenizer, AutoModelForSeq2SeqLM,pipeline
```

```
tokenizer = AutoTokenizer.from_pretrained("BubbleSheep/Hgn_trans_en2zh")

model = AutoModelForSeq2SeqLM.from_pretrained("BubbleSheep/Hgn_trans_en2zh")

translation = pipeline("translation_en_to_zh", model=model, tokenizer=tokenizer)
```

四、模型的基本原理

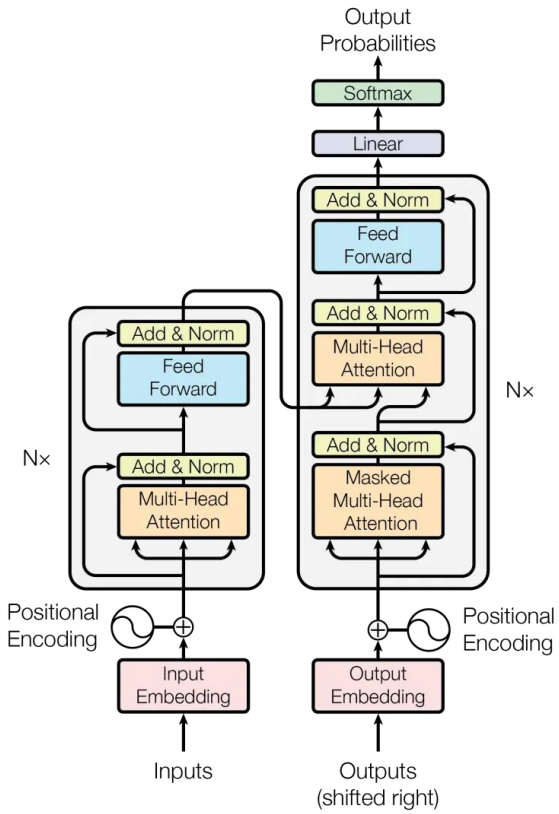
语言模型是自然语言处理的关键， 而机器翻译是语言模型最成功的基准测试。因为机器翻译正是将输入序列转换成输出序列的序列转换模型（sequence transduction）的核心问题。机器翻译（machine translation）指的是将序列从一种语言自动翻译成另一种语言。在翻译的时候，也会有相关问题，比如文本长度、语序变化、文本风格、罕见词、多义词等。我们将通过 Huggingface Transformers 模型来分析译文的通顺程度、准确程度（错译、漏译、过度翻译）、是否自然/地道等，即“信达雅”等。

我们选择 transformer 里的模型有 liam168/trans-opus-mt-en-zh 、Helsinki-NLP/opus-mt-en-zh、BubbleSheep/Hgn_trans_en2zh。OPUS 模型最初是用 Marian[1] 训练的。这是一个高度优化的机器翻译工具包，完全用 C++ 编写。BubbleSheep/Hgn_trans_en2zh:该模型已经进行了预先的英汉翻译训练，并使用 THUOCL 的数据集对模型进行微调。

Huggingface Transformers 是基于一个开源基于 transformer 模型结构[2] 提供的预训练语言库，它支持 Pytorch，Tensorflow2.0，并且支持两个框架的相互转换。框架支持了最新的各种 NLP 预训练语言模型，使用者可以很快速的进行模型的调用，并且支持模型 further pretraining 和下游任务 fine-tuning。

在 transformer 没有诞生之前，大多数模型（Encoder-Decoder）都是基于 CNN 和 RNN 的，而 Transformer 就是基于 attention 机制的，Attention 可以解决 RNN 及其变体存在的长距离依赖问题，也就是 attention 机制可以有更好的记忆力，能够记住更长距离的信息，另外最重要的就是 attention 支持并行化计算，transformer 模型完全的抛弃了 CNN 和 RNN 的结构。整体结构如下：

[1] https://marian-nmt.github.io/docs/
[2] https://arxiv.org/abs/1706.03762



Transformer 就是一个基于多头注意力机制的模型。

Transformer Encoder 模型的输入是一句话的字嵌入表示和其对应的位置编码信息，模型的核心层是一个多头注意力机制。注意力机制最初应用在图像特征提取任务上，比如人在观察一幅图像时，并不会把图像中每一个部分都观察到，而是会把注意力放在重要的部分，后来研究人员把注意力机制应用到了 NLP 任务中，并取得了很好的效果。多头注意力机制就是使用多个注意力机制进行单独计算，以获取更多层面的语义信息，然后将各个注意力机制获取的结果进行拼接组合，得到最终的结果。Add&Norm 层会把 Multi-Head Attention 层的输入和输出进行求和并归一化处理后，传递到 Feed Forward 层，最后会再进行一次 Add&Norm 处理，输出最终的词向量矩阵。

transformer 中使用了 6 个 encoder，为了解决梯度消失的问题，在 Encoders 和 Decoder 中都是用了残差神经网络的结构，即每一个前馈神经网络的输入不光包含上述 self-attention 的输出 Z，还包含最原始的输入。encoder 是对输入（机器学习）进行编码，使用的是自注意力机制+前馈神经网络的结构，同样的，在 decoder 中使用的也是同样的结构。也是首先对输出（machine learning）计算自注意力得分，不同的地方在于，进行过自注意力机制后，将 self-attention 的输出再与 Decoders 模块的输出计算一遍注意力机制得分，之后，再进入前馈神经网络模块。

以上就是 Transformer 的框架了，但是 Transformer 整个框架下来并没有考虑顺序信息。Transformer 用到了另一个概念：“位置编码”。在输入中做手脚，把输入变得有位置信息。给每个词向量加上一个有顺序特征的向量，发现 sin 和 cos 函数能够很好的表达这种特征，所以通常位置向量用以下公式来表示：

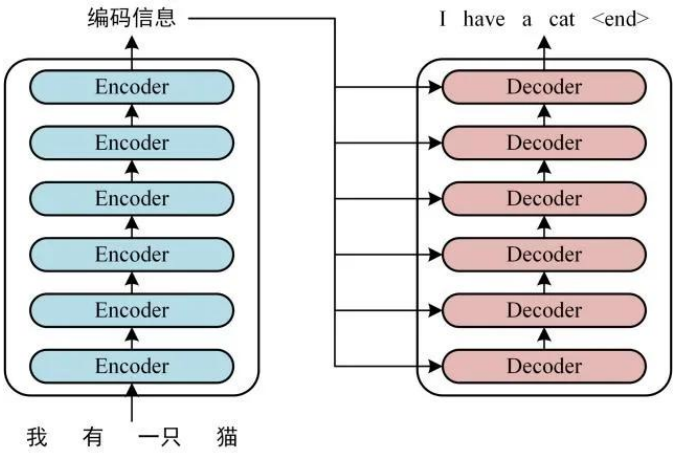
$$PE_{(pos,2i)} = \sin\left(pos/10000^{2i/d}\right)$$
$$PE_{(pos,2i+1)} = \cos\left(pos/10000^{2i/d}\right)$$

• Transormer 的优缺点

优点：（1）虽然 Transformer 最终也没有逃脱传统学习的套路，Transformer 也只是一个全连接（或者是一维卷积）加 Attention 的结合体。但是其设计已经足够有创新，因为其抛弃了在 NLP 中最根本的 RNN 或者 CNN 并且取得了非常不错的效果，算法的设计非常精彩，值得每个深度学习的相关人员仔细研究和品位。（2）Transformer 的设计最大的带来性能提升的关键是将任意两个单词的距离是 1，这对解决 NLP 中棘手的长期依赖问题是非常有效的。（3）Transformer 不仅仅可以应用在 NLP 的机器翻译领域，甚至可以不局限于 NLP 领域，是非常有科研潜力的一个方向。（4）算法的并行性非常好，符合目前的硬件（主要指 GPU）环境。

缺点：（1）粗暴的抛弃 RNN 和 CNN 虽然非常炫技，但是它也使模型丧失了捕捉局部特征的能力，RNN + CNN + Transformer 的结合可能会带来更好的效果。（2）Transformer 失去的位置信息其实在 NLP 中非常重要，而论文中在特征向量中加入 Position Embedding 也只是一个权宜之计，并没有改变 Transformer 结构上的固有缺陷。

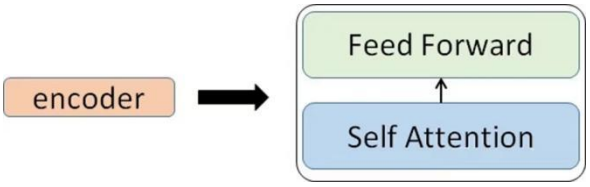
五、翻译过程



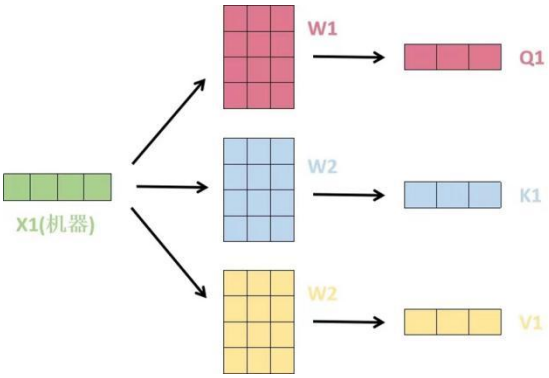
Transformer 的整体结构，左图 Encoder 和右图 Decoder。

当输入一个文本的时候，该文本数据会先经过一个叫 Encoders 的模块，对该文本进行编码，然后将编码后的数据再传入一个叫 Decoders 的模块进行解码，解码后就得到了翻译后的文本，对应的为 Encoders 编码器，Decoders 解码器。一般情况下，Encoders 里边有 6 个小编码器，Decoders 里边有 6 个小解码器。在编码部分，每一个的小编码器的输入是前一个小编码器的输出，而每一个小解码器的输入不光是它的前一个解码器的输出，还包括了整个编码部分的输出。

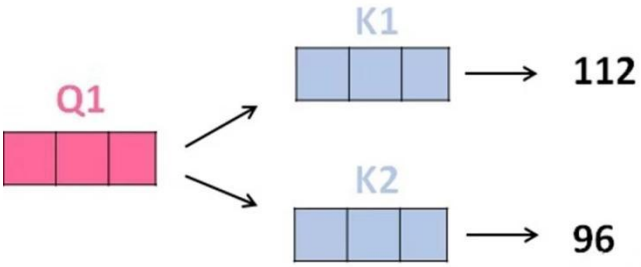
一个 encoder 里边的结构是一个自注意力机制加上一个前馈神经网络。



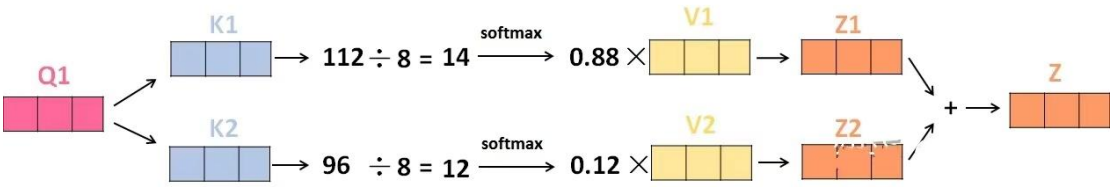
1.self-attention 的输入就是词向量，即整个模型的最初的输入是词向量的形式。自注意力机制就是自己和自己计算一遍注意力，即对每一个输入的词向量，我们需要构建 self-attention 的输入。在这里，transformer 首先将词向量乘上三个矩阵，得到三个新的向量，之所以乘上三个矩阵参数而不是直接用原本的词向量是因为这样增加更多的参数，提高模型效果。对于输入 X1(机器)，乘上三个矩阵后分别得到 Q1,K1,V1，同样的，对于输入 X2(学习)，也乘上三个不同的矩阵得到 Q2,K2,V2。



2. 接下来计算注意力得分，这个得分是通过计算 Q 与各个单词的 K 向量的点积得到的。以 X1 为例，分别将 Q1 和 K1、K2 进行点积运算。

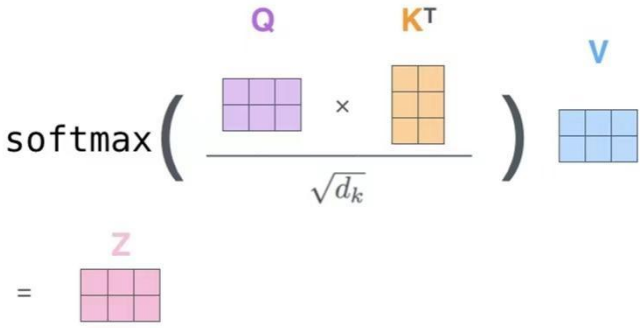


3. 将得分分别除以一个特定数值 8（K 向量的维度的平方根，通常 K 向量的维度是 64）这能让梯度更加稳定。并将上述结果进行 softmax 运算得到，softmax 主要将分数标准化，使他们都是正数并且加起来等于 1。将 V 向量乘上 softmax 的结果，为了保持我们想要关注的单词的值不变，而掩盖掉那些不相关的单词（例如将他们乘上很小的数字）。

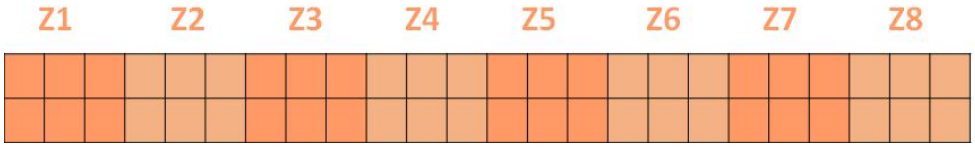


4. 将带权重的各个 V 向量加起来，至此，产生在这个位置上（第一个单词）的 self-attention 层的输出，其余位置的 self-attention 输出也是同样的计算

方式。



•最后，此模型为了进一步细化自注意力机制层，增加了“多头注意力机制”的概念，这从两个方面提高了自注意力层的性能。第一个方面，扩展了模型关注不同位置的能力，这对翻译一下句子特别有用，因为我们想知道“it”是指代的哪个单词。第二个方面，他给了自注意力层多个“表示子空间”。对于多头自注意力机制，我们不止有一组 Q/K/V 权重矩阵，而是有多组（论文中使用 8 组），所以每个编码器/解码器使用 8 个“头”（可以理解为 8 个互不干扰自的注意力机制运算），每一组的 Q/K/V 都不相同。然后，得到 8 个不同的权重矩阵 Z，每个权重矩阵被用来将输入向量投射到不同的表示子空间。经过多头注意力机制后，就会得到多个权重矩阵 Z，我们将多个 Z 进行拼接就得到了 self-attention 层的输出：



我们经过了 self-attention 层，我们得到了 self-attention 的输出，self-attention 的输出即是前馈神经网络层的输入，然后前馈神经网络的输入只需要一个矩阵就可以了，不需要八个矩阵，只需要把这些矩阵拼接起来然后用一个额外的权重矩阵与之相乘即可。最终的 Z 就作为前馈神经网络的输入。

六、数据集的选择及其统计特征

数据集：我们选择了 HuggingFace Transformers 工具包里自带的数据集 opus100 和 news_commentary。

BLEU 自然语言处理中的机器翻译任务中，BLEU 非常常见，它是用于评估模型生成的句子(candidate)和实际句子(reference)的差异的指标.它的取值范围在 0.0 到 1.0 之间，如果两个句子完美匹配(perfect match)，那么 BLEU 是 1.0，反之，如果两个句子完美不匹配(perfect mismatch)，那么 BLEU 为 0.0.虽然这个指标不够完美，但是它有 5 个非常引人注目的好处(compelling benefits)：计算代价小，快.容易理解.与语言无关.与人类评价结果高度相关.被学术界和工业界广泛采用.

BLEU 方法的实现是分别计算 candidate 句和 reference 句的 N-grams 模型[3]，然后统计其匹配的个数来计算得到的. 显然，这种比较方法，是与语序无关的.论文对匹配的 N-grams 计数进行了修改，以确保它考虑到 reference 文本中单词的出现，而非奖励生成大量合理翻译单词的候选结果。本文将其称为修正的 N-grams 精度。sacrebleu 和 nltk 的 bleu 都是两种计算 bleu 的库。两者都提供了 corpus-bleu 和 sentence-bleu 的选项，两个库默认的 bleu 都是 corpus-bleu。

我们调用库函数来对翻译的结果进行评价：

```
from sacrebleu.metrics import BLEU
f = open('2000012952-岳禹彤-enzh1.txt', 'r',encoding='utf-8')
preds = []
lables = []
for line in f:
      pred, lable = line.split('\t')[1:]
      preds += [pred.strip()]
      lables += [[lable.strip()]]
blue = BLEU(tokenize= 'zh')
print(blue.corpus_score(preds, lables))
```

将不同模型和不同数据库的结果进行统计：

Model / Dataset	Helsinki-NLP/opus-mt-en-zh	liam168/trans-opus-mt-en-zh	BubbleSheep/Hgn_trans_en2zh
opus100	BLEU=64.42	BLEU=67.77	BLEU=19.25
news_commentary	BLEU=100	BLEU=100	BLEU=21.36

下面将详细对实验结果进行分析。

七、翻译结果分析：

可以看出opus100数据集在Helsinki-NLP/opus-mt-en-zh、liam168/trans-opus-mt-en-zh两个模型准确率较高，在BubbleSheep/Hgn_trans_en2zh模型上准确率较低。这是因为不同模型使用了不同训练数据集的缘故。

•在opus100上，模型3准确率较低的原因有如下几点：

1. 一些专有名词翻译得准确率较低。

eg：ILOAT, however, appeared to have the edge over UNAT.
然而，劳工组织行政法庭相对于联合国行政法庭似乎有些优势。
然而,ILOAT似乎比UNAT有优势。

2. 语序不能够很好地把握，有时候甚至会忽略掉重要的主语等。

Eg:He liked the dress Mommy bought for me.
他說他喜欢妈妈买给我的裙
喜欢妈妈为我买的衣服

3. 对动词用法把握不准确，以及一些专有名词的不准确

eg：You're gonna lock me up?
你要把我关起来?
你要锁我?
eg:Well, Father is a little worried about you.
父亲有一点为你担心
父王有点担心你

•在news_commentary数据集上，BubbleSheep/Hgn_trans_en2zh模型的准确率较低，总结分析有以下几点：

1. 罕见词专有词分析不透彻：

Eg：Transatlantic Trade for All
所有人的跨大西洋贸易
跨大西洋通商
A Balanced Look at Sino-American Imbalances
中美贸易不平衡的平衡观
平衡观中美平衡

2.长句子翻译会变短

eg：PARIS – As the economic crisis deepens and widens, the world has been searching for historical analogies to help us understand what has been happening. At the start of the crisis, many people likened it to 1982 or 1973, which was reassuring, because both dates refer to classical cyclical downturns.
巴黎-随着经济危机不断加深和蔓延，整个世界一直在寻找历史上的类似事件希望有助于我们了解目前正在发生的情况。一开始，很多人把这次危机比作1982年或1973年所发生的情况，这样得类比是令人宽心的，因为这两段时期意味着典型的周期性衰退。
巴黎—随着经济危机加深和拓展,世界一直在寻找历史类比帮助理解所发生的事情。危机一开始,许多人将它比作1982或1973年,这令人宽慰,因为两个日期都指经典周期滑坡。

3.对语序和句子的主语把握不到位

The tendency is either excessive restraint (Europe) or a diffusion of the effort (the United States). Europe is being cautious in the name of avoiding debt and defending the euro, whereas the US has moved on many fronts in order not to waste an ideal opportunity to implement badly needed structural reforms.

目前的趋势是，要么是过度的克制（欧洲），要么是努力的扩展（美国）。欧洲在避免债务和捍卫欧元的名义下正变得谨慎，而美国已经在许多方面行动起来，以利用这一理想的时机来实行急需的结构性改革。

趋势要么过度约束(Europes ），要么扩散努力(United States)。以避免债务和捍卫欧元为名,欧洲正在谨慎行事,而美国则在许多战线上走动,以免浪费理想契机实施急需的结构性改革。

八、问题总结

（1）英汉语言的不同[3]

从语言形态学分类来说，英语属于印欧语系，是一种综合型语言，而汉语则是一种以分析型为主的语言。英语重形态，汉语轻形态。这种不同语言的形态特点反映在英汉句子结构上，即英语重**形合**,汉语重**意合**。

•英语重形合，句中各意群、成分之间都用适当连接词组成复句，形式上比较严谨，缺乏弹性；汉语则重意合，即更多地依靠语序直接组合复句，由于其句子成分用逻辑意义贯穿起来，结构灵活、简洁，不会引起误解。比如：

a.不入虎穴，焉得虎子。

If one does not enter the tiger’s den, how can he get a tiger’s cub?

b. Never put off till tomorrow what can be done today.

•汉语重意合，呈“隐形”，按照事情发展的先后顺序（时序、因果、时空）逻辑顺序，以短句的形式行文推进，句子之间无过多的形式连接，叙事从容不迫，层层展开。汉语的词汇一般来说并无词性变化，词汇的形态简单，能构成派生词的前缀合后缀的数量有限。它通常要用一个词（词组）来表达英语词缀的意义。语法关系不会死通过词汇自身的形态变化来表达，而是通过虚词、词序等手段来完成的。汉语词汇基本没有词形变化，主要是依靠词语词序及暗含的逻辑关系来表达语义。

•英语是“语调语言”，而汉语是“声调语言”，现代汉语最重要的一个特征在于双音节化和四音节化，因此，叠词，四字短语，双音节字，词组，句子结构很常见，成为一种修辞手段。

如：

It was a day as fresh as grass growing up and clouds going over and

[3] 《英汉对比研究》

butterflies coming down can make it.

绿草萋萋，白云冉冉，彩蝶翩翩，那日子如此清新可爱。

•汉语句子向左，英语句子向右发展，汉民族习惯于先对事情发生的背景进行铺垫，从侧面说明，阐述外围的环境，最后点出话语的焦点信息。如

昨天下午在办公室与她谈话的那个人已经被解聘了。

The man who had a talk with her in the office yesterday afternoon has been dismissed.

•英语重短语；汉语轻短语，这就要求译者在英译汉时既要准确把握数量众多的英语短语的意思，还应考虑汉语行文习惯，使用汉语四字短语和排比句式。

•英语句子的谓语只能由动词担任；而汉语中谓语可由所有的语言单位来承担

由于汉语担任谓语的成分复杂，所以在汉译英时往往要重新确定主语，必须将句子重心局限于谓语动词上，把看似不能担任句子谓语的句子成分转化为英语的动词谓语。

•英语句子的谓语只能由动词担任；而汉语中谓语可由所有的语言单位来承担

由于汉语担任谓语的成分复杂，所以在汉译英时往往要重新确定主语，必须将句子重心局限于谓语动词上，把看似不能担任句子谓语的句子成分转化为英语的动词谓语。

•英语多复合长句；汉语多简单短句

从句可以充当除了谓语动词之外的所有句子成分，使英语句容易出现结构复杂的长句。而汉语词汇的粘合力较差，不宜拖带过多的修饰成分，更多擅长使用流水句式。英译汉时往往需要拆译从句或长句。

• 英语前重心；汉语后重心

英语句子往往先说最近发生的事，再说先前发生的事，基本按时间逆序展开；或者先叙述事实，再说出其发生的时间、地点、方式手段。汉语正好与此相反。

英语句子往往先给出观点、结论、推断，再加以论证；而汉语句子则习惯于先说事，再总结，往往采用“前因后果”的句式。

综上：

英译汉时，应该尽量符合汉语的习惯；句子短小、灵活，敢于打乱原来的形合结构，从连词处拆卸原来英语的严谨、紧凑、四班的结构，进行句式转换、重组，多采用“话题——说明”结构。

（2）英汉的风格问题：

英语造句常用各种形式手段连接词、语、分句或从句，注重显性接应（overt cohesion）,注重结构完整，注重以形显义。

汉语造句少用甚至不用形式连接手段，注重隐形连贯（covert coherence）,注重逻辑事理顺序，注重功能、意义，注重以神统形。

01 原文：近闻夫人健康如常，颇感欣慰[4]。译文：It is a supreme comfort to me when I am informed that you are as healthy as ever.

分析：原文是无主句；而译文考虑到英文形合的特点，首先选择了可以统率行文的“It is…”的主谓句法结构，然后辅以状语从句“when I am informed that you are as healthy as ever”解为“It is a supreme comfort to me”的前

[4] 摘自《1988 年 5 月邓颖超致宋美龄的信》

提，读来非常顺达晓畅。
02 原文：亡羊补牢，犹未为晚。译文：It is not too late to mend the fold even after some sheep have been lost.
分析：首先确定以“It is not too late”作为主线的主谓关系，再通过“to mend the fold ”引出真实主语，使英文更流畅自然、脉络清晰。
03 原文：那时舅舅抱着我，哄着我，我觉得很温暖。译文：Sitting on my uncle’s lap, being humored all the way, I was feeling very good.
分析：原文句式短促，主语从“舅舅”进而转移到了“我”。主语的转移在中文中是自然的。刘士聪先生翻译此句时颇费了一番心思，他以英文的句法手段（这里他用了两个分词结构做并列状语），由一个主语领衔，译得非常巧妙。
英文句子中的形合手段：关系词和连接词、介词，广泛使用代词以及替补词 it 和 there 等。
汉语句子中的意合手段：反复、排比、对偶、对照，紧缩句等。
比如：跑得了和尚，跑不了庙。
Even if the monk can run away, his temple cannot run with him.
谦虚使人进步，骄傲使人落后。
Modesty helps one go forward, whereas conceit makes one lag behind.
不进则退。
Move forward, or you’ll fall behind.
He who does not advance falls backward.
（3）英汉翻译语序变化问题：中英文在表达语序方面存在一定的差异
1. 先表态后叙述
西方人一般先表达某件事的态度，然后再具体表达做某件事，而中国人则相反，因而在翻译这类句子时需要按照英语表达习惯，进行语序调整。
例 若你们能从速发货，以赶上季节开始时的旺季，我们将非常感激。
译文: We shall be very much obliged if you will effect shipment as soon as possible, thus enabling them to catch the brisk demand at the start of the season.
例 如果你方报价具有竞争性，交货期可接受的话，我们愿意向你方订货。
译文：We shall be very glad to place our order with you if your quotation is competitive and delivery date acceptable.
2. 由结果到原因
西方人一般先表达某件事的结果，然后再说明其原因，而汉语表达一般相反，因而在翻译这方面的句子时也需要按照英语表达习惯，进行语序调整。
例 我们认为我们有责任再次提醒贵方，由于进口许可证限制的缘故，信用证展期是不可能的。
译文: We feel it our duty to remind you once again that it is impossible to extend the L/C because of import license restriction.
3. 定语从句后置
汉语属于竹竿型结构，语序是按时间先后顺序；而英语属葡萄型结构，主干突出，较短，一般带有从句，与汉语不同的是其定语从句总是后置。因此，汉译英时必须调整语序，把定语部分调到后面，突出主要部分。
例 1760 年，在英国开始的工业革命大大推动了现代贸易的发展。
译文：The event which most stimulated the development of modern trading

was the Industrial Revolution, which began in England in 1760.

九、参考文献

《英汉对比研究》
https://huggingface.co/models
https://marian-nmt.github.io/docs/
https://arxiv.org/abs/1706.03762