语言分析与机器翻译
课程报告

学 院：计算机科学与工程学院

专 业：计算机科学与技术

班 级：计硕 2004

学 号：2001889

姓 名：周泽帆

2020 年 12 月 20 日

经过这几周机器翻译课程的学习和自己之前的学习积累，我将主要从统计机器翻译和神经机器翻译两方面对所学内容进行简单总结。同时，我使用 RNN 在语言模型上进行了简单应用，实践了课程内容。本报告的组织结构如下：

1 机器翻译知识背景

机器翻译的发展主要经历了基于规则的方法、基于实例的方法、统计机器翻译、神经机器翻译四个阶段。其中，基于规则的机器翻译大多依赖人工定义及书写的规则。主要有两类方法：一类是基于转换规则的机器翻译方法，简称转换法。另一类是基于中间语言的方法。基于实例的机器翻译是在双语句库中找到与待翻译句子相似的实例，之后对实例的译文进行修改，如对译文进行替换、增加、删除等一系列操作，从而得到最终译文。

接下来，我会对课程学习的主要内容，统计机器翻译和神经机器翻译，进行简单总结。

1.1 统计机器翻译

统计机器翻译包括了两个主要过程：训练，即从双语平行数据中学习翻译模型，记为 P(t|s)，其中 s 表示源语言句子，t 表示目标语句子。P(t|s) 表示把 s 翻译为 t 的概率；解码，即使用学习到的模型进行预测，也就是尽可能搜索更多的翻译结果，然后用训练好的模型对每个翻译结果进行打分，最后选择得分最高的翻译结果作为输出。而求最大翻译概率的过程也可以总结为给定源语言句子 s，寻找这样的目标语言译文 t，它使得翻译模型 P(s|t) 和语言模型 P(t) 乘积最大。统计机器翻译主要可分为基于单词、短语、句法的模型。

统计机器翻译的 IBM 模型基于词对齐假设，句子之间的对应就可以由单词之间的对应进行表示。于是翻译句子的概率就转换成了词对齐生成概率：$\mathrm{P(s|t)} = \sum_a \mathrm{P(s,a|t)}$。IBM 模型相当于对这个词对齐的假设再附加不同的条件，以达到优化计算的目标。不同的 IBM 模型化简的层次和复杂度不同。

对于 IBM 模型 1，它做了如下假设：源语言句子长度的生成概率服从均匀分布；对齐概率仅依赖于译文长度，即每个词对齐连接的生成概率也服从均匀分布；源语单词的生成概率仅依赖与其对齐的译文单词。

IBM 模型 2 和 HMM 模型则是通过对扭曲度进行建模来得到翻译模型。相较于 IBM 模型 1，IBM 模型 2 抛弃了词对齐的生成概率服从均匀分布这一假设，认为词对齐是有倾向性的，它与源语言单词的位置和目标语言单词的位置有关。针对 IBM 模型 2 只考虑到了单词的绝对位置，并未考虑到相邻单词间的关系这一问题，基于 HMM 的词对齐模型抛弃了 IBM 模型 12 的绝对位置假设，将一阶隐马尔可夫模型用于词对齐问题。HMM 模型中，对齐概率取决于对齐位置的差异而不是本身单词所在的位置。

由于 IBM 模型 1 和模型 2 对于多个源语言单词对齐到同一个目标语单词的情况并不能很好地进行描述。所以引入了新的翻译模型：首先，确定每个目标语言单词生成源语言单词的个数，称为繁衍率；其次，决定目标语言句子中每个

单词生成的源语言单词都是什么，每个目标语言单词就对应了一个源语言单词列表；最后把各组源语言单词列表中的每个单词都放置到合适的位置上，完成目标语言译文到源语言句子的生成。IBM 模型 3-5 就是对这一基本模型进行了进一步的化简。其中 IBM 模型 4 修正了 IBM 模型 3 不能很好地处理一个目标语言单词生成多个源语言单词的情况。IBM 模型 5 修正了 IBM3、4 会把一部分概率分配给一些根本就不存在的句子这一缺陷。

IBM 模型的贡献在于将扭曲度和繁衍率的概念引入了机器翻译中，为后续的研究提供了新思考。

相比于基于单词的模型，基于短语的模型可以更好地对单词之间搭配和小范围依赖关系进行描述。如今，基于短语的模型仍然是机器翻译的主要框架之一，其中的思想和很多技术手段对今天的机器翻译研究仍然有很好的借鉴意义。句法信息的使用同样是领域主要研究方向之一，使用句法分析可以很好地处理翻译中的结构调序、远距离依赖等问题。这也产生了很多基于句法的机器翻译模型及方法。

1.2 神经机器翻译

神经机器翻译的核心就是编码器解码器框架，它是一种典型的基于“表示”的模型。编码器的作用是将输入的文字序列通过某种转换变为一种新的“表示”形式，这种“表示”包含了输入序列的所有信息。之后，解码器把这种“表示”重新转换为输出的文字序列。这其中的一个核心问题是表示学习，即：如何定义对输入文字序列的表示形式，并自动学习这种表示，同时应用它生成输出序列。

早期神经机器翻译模型的探索集中于对循环神经网络和注意力机制的应用建模。注意力机制的作用就是在生成目标语言单词时能够有选择地获取源语言句子中更有用的部分。神经机器翻译中，注意力机制的是对编码器各层输出进行加权求和来获得一个上下文向量。这个上下文向量是一种包含目标语言单词与源语言单词对应关系的源语言表示。

而随着 Transformer 的问世，传统的循环神经网络和卷积神经网络的结构完全被自注意力机制替代。相较于注意力机制存在长距离依赖的问题，自注意力机制摆脱了顺序传递信息的方式，直接对不同位置单词之间的关系进行建模。具体来说自注意力模型通过计算当前位置与所有位置的匹配程度，也就是在注意力机制中提到的注意力权重，来对各个位置进行加权求和。得到的结果可以被看作是在这个句子中当前位置的抽象表示。这个过程，可以叠加多次，形成多层注意力模型，对输入序列中各个位置进行更深层的表示。Transformer 已经成为了机器翻译中最先进的架构之一。

2 项目实践

本次课程项目参考的论文是 Recurrent neural network based language model，即基于 RNN 的语言模型。n-gram 语言模型指的是根据前 n-1 个单词来预测下一个单词的模型。

论文的主要方法如下：

$$x(t) = w(t) + s(t-1)$$

$$s_j(t) = f(\sum_i x_i(t) u_{ji})$$

$$y_k(t) = g(\sum_j s_j(t)v_{kj})$$

其中 x(t)表示 t 时刻的输入，它是由当前词的表示 w(t)和 t-1 时刻的隐藏层输出 s(t)连接起来的。y(t)表示 t 时刻的输出，s(t)表示 t 时刻的隐藏层状态。f(x)是 sigmoid 激活函数，g(x)是 softmax 函数。u，v 表示权重矩阵。这个计算过程实际上就是还原了 RNN 的结构，如图 1 所示。

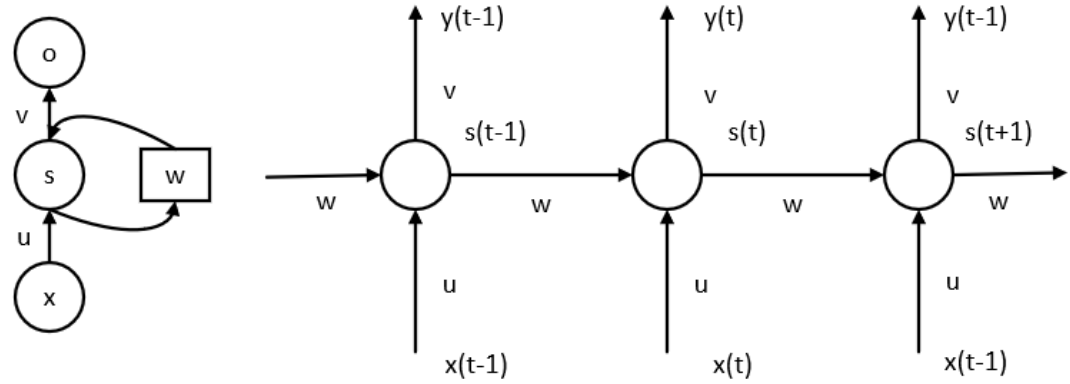

图 1 RNN 结构示意图

我对 RNN 语言模型进行了分 batch 训练（1000），使用 3-gram，模型隐藏层维度为 50，最终在 PTB 数据集上获得了 210 的 ppl 值，训练过程的 ppl 变化图如图 2 所示。

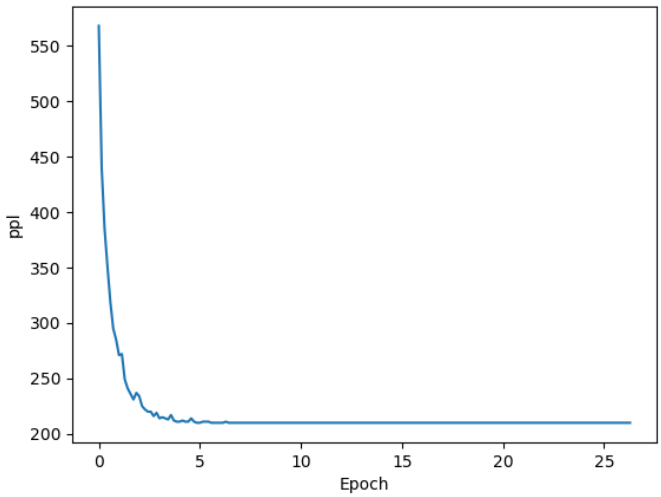

图 2 训练过程 ppl 的变化

3 课程感悟与思考

通过语言分析与机器翻译课程的学习，我对机器翻译的发展和具体方法有了更深刻的认识，并且有了切实的实践过程。

其中不得不提的是，课堂的教材——《机器翻译：统计建模与深度学习方法》，内容十分详尽，通过对本书的阅读和课堂的学习，我能高效地理解之前可能看很多天也一知半解的 Transformer 结构，望而却步的 IBM 模型等等。

虽然之前有过一些相关的知识背景，但是，此次课程对于统计机器翻译的介

绍，又让我对机器翻译这一方向有了更深的认识。无论是基于词、基于短语、基于句法的思维方式还是 IBM 引入的扭曲度、繁衍率的概念，都让我思考传统的语言分析策略、建模方法、优化策略等对于现阶段机器翻译的重要性。即使现在已经是神经机器翻译的天下，之前长达数十年的统计机器翻译的研究成果仍然是举足轻重的。我们需要好好利用这些成果，让他们成为现阶段方法的先验知识或是融入方法中去，更有意义地传承前人。

在课程之余，我也阅读了一些相关论文，其中，给我留下深刻印象的是一篇将 Transformer 结构移植到图像分类领域的文章，其方法称为 Vision Transformer。作者将图片分成一个个 patch，即子图像块，再进行压缩连接，形成类似于自然语言处理中 token 串的形式，输入进 Transformer 编码器结构中，输出结果作为图像的表示，进行分类。最终在大规模数据和模型下取得了优于传统方法的结果。在这篇论文中，我们看到了 Transformer 结构，或者说自注意力机制，对于传统 CNN 的替代作用。起源于机器翻译的模型，也同样能在图像领域大有可为。同样地，机器翻译的未来发展方向——多模态机器翻译，也是两个领域研究成果的融合，这会是未来机器翻译研究的核心内容。我也会继续好好在这一方向学习，保持探索求知的精神。

对课程的建议：

（1） 可能对于 IBM 模型部分的介绍可以再简略一些，到后面感觉大家会有些更不上的情况；

（2） 对于上机，除了复制命令行，体会整个训练解码过程之外，感觉可以再多一些自己调整程序的实践过程；

（3） 对于项目内容，感觉多一些对于简单模型的应用的题目吗，RNN，Seq2Seq，加 Attention 之类的。