doi:10.3772/j.issn.2095-915x.2022.06.010

基于语义信息共享 Transformer 的古文机器翻译方法

开放科学（资源服务）标识码（OSID）

周成彬[1] 刘忠宝[1,2,3]

1. 中北大学软件学院 太原 030051；
2. 北京语言大学语言智能研究院 北京 100083；
3. 泉州信息工程学院软件学院 泉州 362000

摘要：[目的 / 意义] 中国古籍浩如烟海，承载了古人的精神和智慧，对古文进行翻译有利于继承和发扬中华传统文化。随着人工智能技术的发展，利用计算机实现古文的自动翻译具有重要意义。[方法 / 过程] 然而，目前对古文进行机器翻译的研究还比较少。因此，本文根据古文与现代文属于同一语言的特点，通过共享词表与嵌入层参数的方法，提出了基于语义信息共享的 Transformer 模型实现古文到现代文的自动翻译，使用 BLEU 作为评价指标。[结果 / 结论] 实验表明，该模型的 BLEU 值达到了 31.43，相比传统的基于 GRU 和 LSTM 的 seq2seq 模型分别提高了 26.94 和 15.19 个 BLEU 值，比基准 Transformer 模型提高了 13.41 个 BLEU 值，证明了模型的有效性。

关键词：机器翻译；Transformer 模型；BLEU；古文翻译

中图分类号：G35; TP391

Machine Translation of Ancient Chinese Text Based on Transformer of Semantic Information Sharing

ZHOU Chengbin[1] LIU Zhongbao[1,2,3]

1. School of Software, North University of China, Taiyuan 030051, China;
2. Institute of Language Intelligence, Beijing Language and Culture University, Beijing 100083, China;
3. School of Software, Quanzhou University of Information Engineering, Quanzhou, Fujian 362000, China

Abstract: [Objective/Significance] Chinese ancient books are as vast as a vast ocean, carrying the spirit and wisdom of the ancients, and translating ancient texts is conducive to inheriting and carrying forward traditional Chinese culture. With the development of artificial intelligence technology, it is important to use computers to realize automatic translation of ancient

基金项目 教育部哲学社会科学研究后期资助项目“大数据环境下数字人文理论、方法与应用研究”（21JHQ081）；福建省社会科学基金项目“大数据驱动的古籍故事化表达与情景化再现研究”（FJ2022A018）。

作者简介 周成彬（1997-），硕士研究生，研究方向为自然语言处理；刘忠宝（1981-），博士，教授，研究方向为文本挖掘、自然语言处理，E-mail：liuzb@nuc.edu.cn。

引用格式 周成彬 , 刘忠宝 . 基于语义信息共享 Transformer 的古文机器翻译方法 [J]. 情报工程 , 2022, 8(6): 114-127.

texts. [Methods/Processes] However, there are still relatively few studies on machine translation of these ancient texts. Therefore, according to the characteristics that ancient texts and modern texts belong to the same language, this paper proposes a Transformer model based on semantic information sharing to realize automatic translation from ancient texts to modern texts by sharing vocabulary and embedding layer parameters, using BLEU as an evaluation index. [Results/Conclusions] Experiments show that the model achieves a BLEU value of 31.43, which is 26.94 and 15.19 BLEU values higher than the traditional GRU-based and LSTM-based seq2seq models, respectively, and 13.41 BLEU values higher than the benchmark Transformer model, so the model is effective.

Keywords: Machine translation; Transformer model; BLEU; ancient Chinese translation

引言

中国古代历史典籍不仅是中国独特的文化遗产，也是世界文明的瑰宝。然而，随着语言在漫长的历史演变过程中，现代人要理解甚至创作古代作品变得相当困难。首先，现代文追求通俗易懂与口语化，而古文追求行文简练且有许多特殊句式，两者的语法顺序大不相同。其次，古文单词多为单音节词，而现代文单词多为双音节。因此，翻译在弥合这两个时代方面发挥着关键作用。对于传统的人工翻译而言，中华古籍的翻译是一项非常困难且耗时的工作，虽然能够取得高质量的译文，但是对翻译者的文化水平要求很高，且需要耗费大量的人力和时间，成本太高。而随着计算机科学技术的发展，为了顺应时代的需要，利用计算机进行自动翻译的机器翻译（Machine Translation，MT）技术越来越成熟。近年来，随着大数据、云计算、人工智能等技术的快速发展，基于语料库的机器翻译逐渐占据主流，其中，又以基于深度学习的神经网络机器翻译最为典型。通过使用基于深度学习的神经机器翻译技术对古籍文献进行自动翻译对我们了解历史、学习优秀传统文化，继承和发扬中华民族精神具有重要意义。

机器翻译是利用计算机自动将一种自然语言（源语言）转换为另一种自然语言（目标语言）的过程[1]。机器翻译是自然语言处理的一个研究分支，也是人工智能的终极目标之一，具有重要的科学研究价值。机器翻译于 20 世纪 30 年代初露端倪，如今已取得突破性进展。从历时视角来看，机器翻译大致经历过两种范式：语言学范式和语料库学范式，前者为基于规则的机器翻译，后者则为基于语料库的机器翻译。目前主流的机器翻译为的神经网络机器翻译[2]（Neural Machine Translation，NMT），相比于传统的统计机器翻译[3]（Statistical Machine Translation，SMT）而言，NMT 能够训练一张从一个序列映射到另一个序列的神经网络，输出的可以是一个变长的序列，NMT 其实是一个 Encoder-Decoder[4] 系统，Encoder 把源语言序列进行编码，并提取源语言中信息，通过 Decoder 再把这种信息转换到另一种语言即目标语言中来，从而完成对语言的翻译。随着基于循环神经网络[5]（Recurrent Neural Network，RNN）的序列到序列框架[6]（Sequence to Sequence，seq2seq）、注意力机制[7]（Attention Mecha-

doi:10.3772/j.issn.2095-915x.2022.06.010

nism）以及基于纯注意力机制的 Transformer 模型 [8] 的提出，NMT 的翻译质量得到了巨大的提升，彻底取代了 SMT 成为了主流的机器翻译技术。

目前 NMT 在许多双语翻译研究上取得了巨大的成效，如中英、中俄，英德等不同语种间的互译，然而在古文到现代文之间的语内互译的研究还比较少。本文根据古文与现代文属于同一语言，共享大量词汇的特点，提出一种基于语义信息共享的 Transformer 机器翻译模型。使用由 NiuTrans Open Source（NOS）①开源的古今平行语料数据集，并使用 BLEU（Bilingual Evaluation Understudy）[9] 值对模型进行评估，取得 31.4 的 BLEU 值，译文质量良好，达到了机器翻译的效果。

1 相关工作

目前古文到现代文的翻译还是以人工翻译为主，人工翻译需要专家准确的理解古文，这就要求翻译者有较高的文化水平，翻译者的数量有限。人工翻译需要耗费大量的人力和时间，且难以实现即时翻译。随着计算机和互联网的发展，利用计算机技术实现即时自动的翻译成为了研究热点。Liu[10] 提出了利用知识图谱理论从古文到现代文的翻译思路，知识图谱具有很强的句子表达能力，对人工翻译有巨大帮助，但是难以实现自动翻译。王爽等 [11] 利用基于规则的机器翻译技术实现了一个古文自动翻译系统，能够实现部分古文献的翻译和标注。韩芳等 [12] 利用句本位句法相关规则构造知识库，使用词义消岐算法，对古文进行基于规则和统计相结合的机器翻译研究。基于规则的机器翻译需要构建大量的规则，成本高，且翻译的句子比较生硬，译文质量低。郭锐等 [13] 建立古文句子的全文索引，基于汉字的信息熵完成最相似古文句的高效检索，为基于实例方法的古今汉语的机器翻译奠定了基础。王爽等 [14] 提出了基于实例的古文机器翻译系统，该系统需要构建一个句对齐和字对齐的语料库，系统根据语言学知识对源语句进行语法分析，然后在语料库中进行匹配分析，结合一定的语法规则输出译文。基于实例的机器翻译方法有明显的缺点，系统复杂，效率低，翻译质量不高，对于语料库中没有的句子难以实现正确的翻译。杨钦 [15] 基于对统计机器翻译系统 Moses 进行优化实现古汉语到现代汉语的翻译。统计机器翻译有很强的词汇和短语翻译能力，但是对译文语序的调序能力较差，难以实现对长句子和复杂句子的顺畅翻译。Zhang 等 [16] 提出具有复制机制和局部注意力机制的端到端的神经网络模型来实现古文到现代文的互译。由于使用神经网络模型需要大量的基于句子对齐的平行语料数据进行训练，他们提出一种无监督算法，该算法利用古今对齐的句子对之间存在许多相同的汉字的特点来构建古今句子对齐的语料库。为了获得更多的古今句对齐语料来提高翻译质量，Liu 等 [17] 根据古现代汉语的特点提出了一种古今从句对齐方法，该方法将词法信息与统计信息相结合，数据集的构建包括平行语料库的爬取和清洗、段落对齐、基于对齐段落的子句对齐以

①https://github.com/NiuTrans/Classical-Modern

及通过合并对齐的相邻子句来扩充数据等四步，创建了一个包含 1.24 万对古今双语句对齐的语料库。在平行语料匮乏的情况下，Chang 等 [18] 将翻译任务构建为多标签预测任务，模型除了预测出翻译外还预测出古文的年代信息，将年代信息作为辅助，再加上按时间顺序的上下文作为辅助信息来提高模型的翻译质量。魏家泽 [19] 构建了基于外部知识协同的古文到现代文的机器翻译模型，通过句内片段协同、注释信息协同和语言知识协同的三维外部知识的联合使用，有效提升了古文到现代文的机器翻译性能。

对以上相关研究进行梳理可以发现，古文的机器翻译研究还存在以下几点问题：一是平行语料匮乏，缺乏大规模高质量的古今平行语料集，难以在深度学习 [20]（Deep Learning，DL）模型上获得令人满意的效果。本文使用的数据集是由 NOS 开源的古今平行语料集，该语料集大小为 151 MB 共 97.5 万条古今平行句子对，基本涵盖了大部分经典古籍著作，是目前已知的最大规模的古今平行语料集，可以满足本模型的需求。二是未能有效的利用古文与现代文属于同一语种的优势。基于此，本文提出了一种简单有效的方法，通过共享词表和嵌入层参数实现古文与现代文中相同词汇的语义信息共享，通过实验证明这有效的提升了机器翻译性能。三是汉语时代差异的问题，我国汉语历史悠久，每个朝代的语言风格不同，不同朝代的文字，即使相同，含义相差也很大。在已有研究中，大多是通过以年代信息或者引入外部知识作为辅助来提高翻译的准确性，但是在对未知年代信息和没有外部知识的古文进行翻译时，效果就会下降。本文旨在训练一个对古文进行翻译时无需引入任何外部知识，可以直接对原始古文句子进行高质量翻译的模型，故本文使用完全基于注意力机制的深度学习模型 Transformer 进行训练，相比于传统的基于 RNN 或 CNN[21] 的 seq2seq 模型，Transformer 对句子有更强的全局信息感知能力，可以学习到句子的内部依赖关系，只要各个朝代的平行语料足够，模型就可以学到各个朝代的翻译风格，在对多义词进行翻译时，模型会结合上下文信息对该词进行最符合句子语境的翻译。

2 语义信息共享的 Transformer

Transformer 是一个 Encoder-Decoder 架构，其整体结构如图 1 所示，它由编码器和解码器两部分组成。输入古文，通过编码器输出对应的上下文向量，与 RNN 将源语句压缩为一个上下文向量不同，Transformer 的编码器将会得到与输入序列等长的上下文向量，且可以对输入的每个单词进行并行处理。解码器将编码器输出的上下文向量作为输入，最终输出现代文。

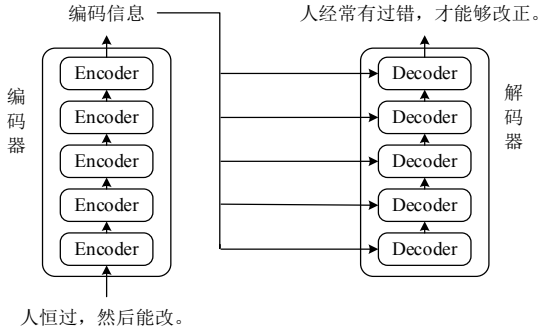

图 1 Transformer 整体结构

Transformer 的编码器部分和解码器部分都是由多个相同的层叠加组成的，其内部结构如图 2 所示，左边是编码器部分，右边是解码器

doi:10.3772/j.issn.2095-915x.2022.06.010

部分。每个编码器层由两个子层组成，第一个子层是多头自注意力层，第二个子层是一个基于位置的全连接前馈网络。为了更好的优化深度网络防止网络退化和避免梯度爆炸或消失，每个子层都添加了残差连接 [22] 和进行了层归一化 [23]。每个解码器层由三个子层组成，且每个子层也都使用了残差连接和层归一化。第一个子层为掩码多头自注意力层，其掩码机制使得模型保留了自回归属性，确保预测仅依赖于已生成的输出词元 [8]。第二个子层为多头注意力层，用来接收编码器的输出。解码器的输出结果最后还需要通过一层线性层和 softmax 层，线性层输出一个大小等于目标语言词汇表大小的向量。然后通过 softmax 层得到每个词向量的概率，并选择概率最高的词向量作为当前位置的输出。

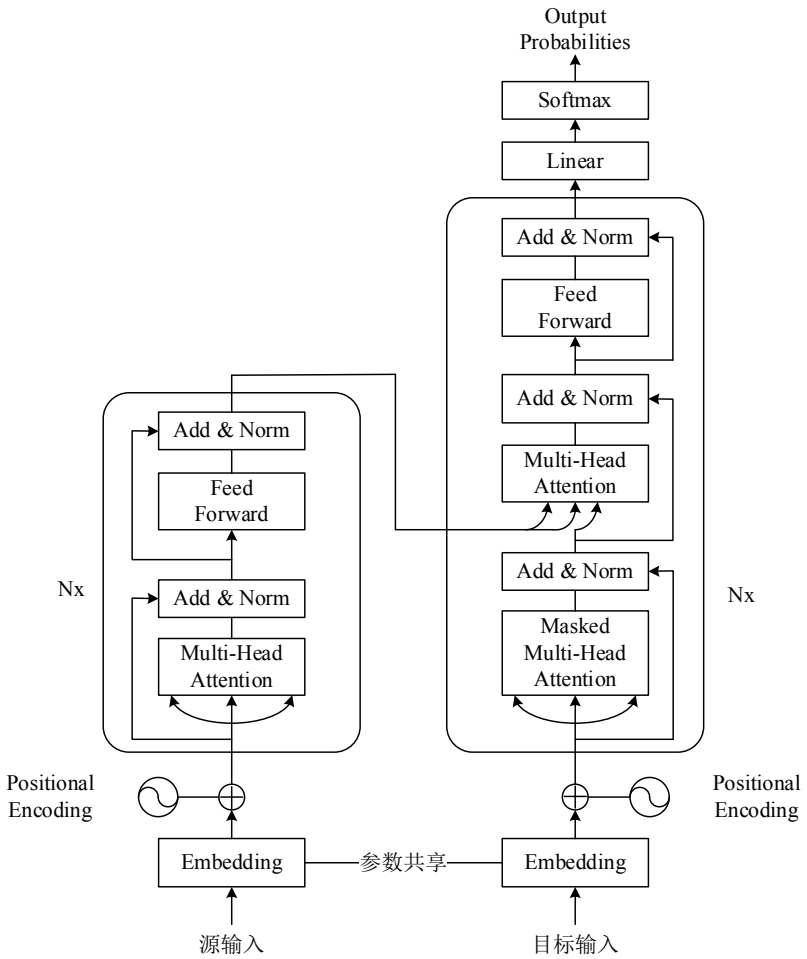

图 2　**Transformer** 内部结构

2.1　语义信息共享的嵌入层

古文与现代文之间存在大量语义相同或相近的词元，为了使模型的编码器和解码器能共享这些语义信息，本文提出了两个简单有效的方法。一是在构建词表时，古文与现代文共用

一个词表，在对句子做嵌入时，编码器只激活古文的 Embedding，解码器只激活现代文的 Embedding，虽然词表变大了但是并不会影响模型的性能，且可以帮助编码器和解码器共享语义相同的词。针对汉语中同一个文字在不同词组中含义不同的问题，本文在构建词表时以词为粒度对句子进行切分，模型对多义字进行翻译时会根据其所组成的词组进行对应的翻译。而针对同一个词在不同朝代含义不同的问题，则是通过模型对该朝代大量平行句子对进行训练从而获得该朝代的表述特征和特有语义，结合句子上下文信息对多义词进行翻译。共享词表还可以有效的减少低频词，增加模型的泛化能力，提升机器翻译的效果。二是 Transformer 的编码器和解码器共享嵌入层参数，嵌入层的作用是通过向量化对句子进行表征，共享嵌入层可以有效利用古文与现代文属于同一语种的优势，对于语义信息相同的词在翻译过程中的一致性可以更好的建模。共享嵌入层参数还可以减少模型的参数数量，加快模型的收敛速度。

2.2 位置编码

Transformer 模型没有使用任何循环神经网络或卷积神经网络，而是完全依赖于注意力机制进行计算。而注意力机制是对序列进行并行计算，不是进行顺序计算，不能获取到序列间的位置信息。为了使模型知道输入序列的位置信息，所以在模型底部的嵌入层后加上了位置编码层。位置编码向量的维度与嵌入表示的词向量维度相同，二者可以直接进行相加。位置编码可以是可学习也可以是固定的，本文使用的是基于正弦函数和余弦函数的固定位置编码，位置编码的计算如公式（1）、（2）所示。

$$PE_{(pos,2i)} = \sin\left(pos / 10000^{2i/d_{model}}\right) \quad (1)$$

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

其中 PE 指的是位置编码矩阵，pos 表示当前词在序列中的具体位置，i 表示词向量的第 i 个维度，d_{model} 表示词向量的维度大小。假设输入序列的长度为 60，向量维度为 32，其位置编码如图 3 所示，横坐标表示序列中每个词的位置，纵坐标表示位置向量的值，每条曲线表示每个维度的位置向量。

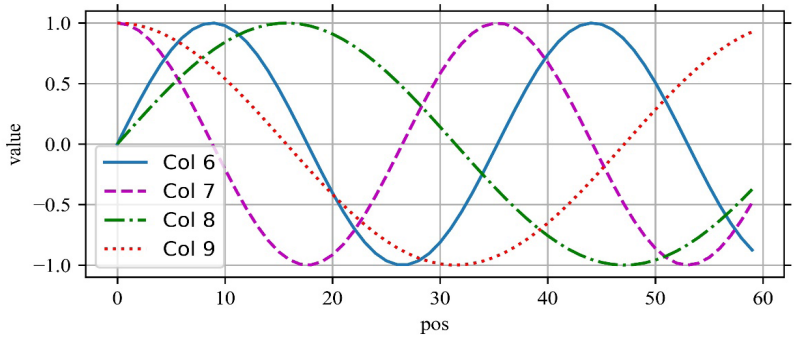

图 3　长度为 60，维度为 32 的正余弦位置编码示例图

2.3 多头注意力机制

Transformer 的核心就是注意力机制。注意力机制可以使神经网络有选择性的关注重要信息，忽略无关信息。其核心内容为根据 query、

doi:10.3772/j.issn.2095-915x.2022.06.010

key 和 value，通过注意力函数后输出 value 的加权和，query、key 和 value 都是向量。注意力函数的具体计算过程可以分为三步：第一步使用注意力评分函数 sim 计算出 query、key 二者的相似度得到注意力分数，第二步使用 softmax 函数将注意力分数进行归一化得到一个概率分布作为注意力权重，注意力权重系数之和为 1，第三步根据注意力权重对 value 做加权求和。在计算时，一般将一组 query、key、value 分别组合成矩阵 Q、K、V。注意力函数的计算公式如式（3）所示。

$$Attention\ (Q,K,V)=\text{softmax}(\text{sim}(Q,K))V \quad (3)$$

Transformer 使用的是缩放点积注意力，其结构如图 4 所示，注意力评分函数使用的是点积函数。为了避免点积的值太大，在计算出 Q 和 K 的注意力分数后要除以进行缩放。d_k 是矩阵 K 的维度，如果 d_k 太大，点积得到的值会较大，经过 softmax 操作之后产生的梯度太小，不利于模型的训练。

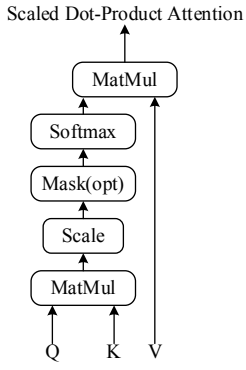

图 4　缩放点积注意力

缩放点积注意力的公式如式（4）所示。

$$Attention(Q,K,V)=\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4)$$

对于自注意力而言，Q、K、V 都是由同一个输入序列 $X(a,\cdots\cdots,a_n)$ 分别乘以三个权重矩阵 W_q、W_k、W_v 进行线性变换得到的，如公式（5~7）所示。

$$Q=W_q\cdot[a_1a_1\cdots a_i] \quad (5)$$

$$K=W_k\cdot[a_1a_1\cdots a_i] \quad (6)$$

$$V=W_v\cdot[a_1a_1\cdots a_i] \quad (7)$$

由于 Q、K、V 都来自于同一个序列，在对句子进行注意力编码时，当前位置的词会过度关注于自身，而忽略其他位置。为了解决这个问题，模型使用了多头自注意力机制，其结构如图 5 所示。多头自注意力机制就是将 Q、K、V 映射到多组不同的子空间进行自注意力计算，然后再将每个不同的自注意力结果进行拼接，最后进行一次线性变换输出。多头注意力机制可以使模型学习到不同子空间的信息，增强了模型的表达能力。

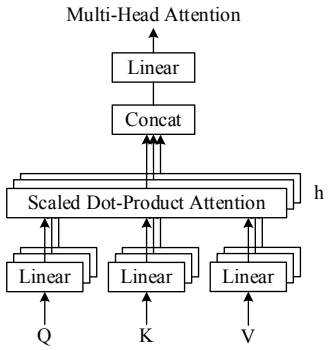

图 5　多头自注意力

多头自注意力机制的公式如式（8~9）所示。

$$MultiHead(Q,K,V)=\text{Concat}(\text{head}_1,\ldots,\text{head}_h)W^O \quad (8)$$

$$head\ i_i=Attention\left(QW_i^Q,KW_i^K,VW_i^V\right) \quad (9)$$

其中，W_i^Q、W_i^K、W_i^V 为参数矩阵，$W_i^Q\in\mathbb{R}^{d_{\text{model}}\times d_k}$，$W_i^K\in\mathbb{R}^{d_{\text{model}}\times d_k}$，$W_i^V\in\mathbb{R}^{d_{\text{model}}\times d_v}$，$W^O\in\mathbb{R}^{hd_v\times d_{\text{model}}}$，$d_k=d_v=d_{\text{model}}/h$。

2.4 位置前馈网络

位置前馈网络（Feed Forward Neural Network，FFN）是由两层线性变换组成的多层感知机，两层线性变换之间使用 Relu 激活函数进行连接，且参数不共享。位置前馈网络的公式如式（10）所示。

$$FFN(x)=\max\left(0,xW_1+b_1\right)W_2+b_2 \quad (10)$$

其中，x 代表输入，W_1 表示第一次线性变换的参数矩阵，b_1 表示第一次线性变换的偏置向量，W_2 表示第二次线性变换的参数矩阵，b_2 表示第二次线性变换的偏置向量。

3 实验分析

3.1 数据集

本文使用的数据集是由东北大学 NLP 实验室和 Niutrans Research 维护的 NOS 开源的中文古今平行语料集，该语料集共约 97 万句对，基本涵盖了大部分经典古籍著作，部分样本数据展示如表 1 所示。

表 1 古今对齐语料展示

古文	现代文
世祖发百里内马，得千五百匹。	世祖征发百里之内的马匹，共得一千五百匹。
至遇事，则别白是非，不少借隐。	至于碰上事情，就分辨是非，没有一点隐讳。
王并依表以遣之。	国王全都依照他上表说的办。
未至越，越杀其王降，汉兵亦罢。	还没有到达越地，越人就杀死了他们的国王向汉朝投降，汉军也就收兵了。

在使用数据集训练模型之前，需要对数据集进行划分。首先将所有数据进行随机打乱以增强模型的泛化能力，然后将模型划分为三组：训练集、验证集和测试集，大致按照 8：2：2 的比例进行划分，三个数据集的统计信息如表 2 所示，训练集共 955096 条句对，验证集和测试集都为 10000 条句对。

表 2 数据集划分

设置	句子对条数
训练集	955096
验证集	10000
测试集	10000

3.2 数据集处理

机器翻译模型不能识别原始数据，在进行机器翻译训练前需要对数据集进行以下几步处理：分词、制作词表、将文本序列转化为数字序列、构造小批量数据。

第一步分词也叫做词元化，就是将文本切分为一个个独立的词。本文对古文使用的 Jiayan[②]分词，Jiayan 是一款专注于古汉语处理的 NLP 工具包，Jiayan 分词有两种分词算法，一是利用无监督、无词典的 N 元语法和隐马尔可夫模型进行古文自动分词，二是利用词库构建功能产生的文言词典，基于有向无环词图、句子最大概率路径和动态规划算法进行分词，本文使用的是第一种分词算法。本文对现代文使用的是 Jieba[③]分词，Jieba 分词是一款优秀的中文分词工具，其分词原理为首先基于前缀词

②https://github.com/jiaeyan/Jiayan
③https://github.com/fxsjy/jieba

doi:10.3772/j.issn.2095-915x.2022.06.010

典实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图（DAG），然后采用了动态规划查找最大概率路径，找出基于词频的最大切分组合，对于未登录词，采用了基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法。古文使用 Jiayan 分词后的效果如表 3 所示，现代文使用 Jieba 分词的效果如表 4 所示。

表 3　古文分词

古文	Jiayan分词
世祖发百里内马，得千五百匹。	[‘世祖’, ‘发’, ‘百里’, ‘内’, ‘马’, ‘，’, ‘得’, ‘千五百’, ‘匹’, ‘。’]
至遇事，则别白是非，不少借隐。	[‘至’, ‘遇事’, ‘，’, ‘则’, ‘别’, ‘白’, ‘是非’, ‘，’, ‘不’, ‘少’, ‘借’, ‘隐’, ‘。’]
王并依表以遣之。	[‘王’, ‘并依’, ‘表’, ‘以’, ‘遣’, ‘之’, ‘。’]
未至越，越杀其王降，汉兵亦罢。	[‘未至’, ‘越’, ‘，’, ‘越’, ‘杀’, ‘其’, ‘王’, ‘降’, ‘，’, ‘汉兵’, ‘亦’, ‘罢’, ‘。’]

表 4　现代文分词

现代文	Jieba分词
世祖征发百里之内的马匹，共得一千五百匹。	[‘世祖’, ‘征发’, ‘百里’, ‘之内’, ‘的’, ‘马匹’, ‘，’, ‘共得’, ‘一千五百’, ‘匹’, ‘。’]
至于碰上事情，就分辨是非，没有一点隐讳。	[‘至于’, ‘碰上’, ‘事情’, ‘，’, ‘就’, ‘分辨是非’, ‘，’, ‘没有’, ‘一点’, ‘隐讳’, ‘。’]
国王全都依照他上表说的办。	[‘国王’, ‘全都’, ‘依照’, ‘他’, ‘上表’, ‘说’, ‘的’, ‘办’, ‘。’]
还没有到达越地，越人就杀死了他们的国王向汉朝投降，汉军也就收兵了。	[‘还’, ‘没有’, ‘到达’, ‘越地’, ‘，’, ‘越人’, ‘就’, ‘杀死’, ‘了’, ‘他们’, ‘的’, ‘国王’, ‘向’, ‘汉朝’, ‘投降’, ‘，’, ‘汉军’, ‘也’, ‘就’, ‘收兵’, ‘了’, ‘。’]

第二步制作词表，就是根据分词后的结果对源语言和目标语言别分建立一个字典。由于古文与现代文使用的都是简体汉字，属于语内翻译，所以本文中使用古文分词后的结果和现代文分词后的结果共同建立了一个词表。在机器翻译中，只需要使用训练集的数据进行词表的创建而不使用验证集和测试集的数据。还指除此之外，定了三个额外的特殊词元，<unk> 用以统一表示低频词和未知词，<pad> 用来进行填充，<bos> 作为开始词元，<eos> 作为结束词元。统计去除低频词后的训练集的词表大小如表 5 所示。

表 5　词表统计

最少出现次数	合并	source	Target
1	627618	210053	491326
2	329530	132531	250191
3	230229	102610	171461
4	188609	87448	138709

第三步将文本序列转化为数字序列，就是将分词后的文本序列通过查表转化为数字序列。数字化的效果如表 6 所示。

表 6　文本序列数字化

文本序列	数字序列
[‘世祖’, ‘发’, ‘百里’, ‘内’, ‘马’, ‘，’, ‘得’, ‘千五百’, ‘匹’, ‘。’]	[536, 308, 2157, 221, 247, 4, 46, 7844, 2469, 5]
[‘至’, ‘遇事’, ‘，’, ‘则’, ‘别’, ‘白’, ‘是非’, ‘，’, ‘不’, ‘少’, ‘借’, ‘隐’, ‘。’]	[55, 9140, 4, 52, 681, 444, 3081, 4, 9, 323, 2557, 1513, 5]
[‘王’, ‘并依’, ‘表’, ‘以’, ‘遣’, ‘之’, ‘。’]	[44, 40390, 810, 12, 192, 7, 5]
[‘未至’, ‘越’, ‘，’, ‘越’, ‘杀’, ‘其’, ‘王’, ‘降’, ‘，’, ‘汉兵’, ‘亦’, ‘罢’, ‘。’]	[3550, 699, 4, 699, 120, 21, 44, 305, 4, 4891, 104, 495, 5]

第四步是构建小批量数据。由于计算机的内存有限，模型进行训练时不能一次性进行全局训练，而是将数据打包成批量大小相同的小批量数据（batch）进行训练。由于每条句子的长度不相同，在构造小批量数据时，以每个 batch 中最长的序列为标准，使用 <pad> 进行补齐，同时在序列的开始添加 <bos>，在结尾添加 <eos>。

3.3 实验环境

实验环境如表 7 所示。

表 7 实验环境与配置

实验环境	环境配置
操作系统	ubuntu18.04
CPU	Intel(R) Xeon(R) Gold 6330
GPU	RTX A5000
显存	48GB
内存	50GB
编程语言	Python3.8
深度学习框架	PyTorch 1.9.0

3.4 参数设置

模型的相关参数设置如表 8 所示。模型训练时使用带掩码的交叉熵损失函数（CrossEntropyLoss），在计算损失时会自动忽略所有的 <pad>。模型使用的优化器为 Adam。模型进行预测时使用搜索策略为束搜索（beamsearch），束宽（beam_size）设置为 5。

表 8 模型参数设置

参数类型	值	参数解释
vocab_size	200000	词表大小
batch_type	tokens	批次类型
batch_size	4096	训练批次大小
valid_batch_size	1024	验证批次大小
model_dtype	fp32	模型数据类型
learning_rate	0.0002	学习率
warmup_steps	8000	自定义衰减的预热步骤数
label_smoothing	0.1	平滑率
average_decay	0.0005	移动平均衰减
enc_layers	6	编码器层数
dec_layers	6	解码器层数
heads	8	多头注意力的头数
d-model	512	编-解码器隐藏节点数
word_vec_size	512	嵌入层隐藏节点数
transformer_ff	2048	前馈层隐藏节点数
dropout	0.1	丢弃率

3.5 评价指标

BLEU 是机器翻译中常用的评价指标，它可以对模型生成的句子和实际的目标句子进行相似度评估，BLEU 的取值在 0 到 1 之间，0 表示两个句子完全不匹配，1 表示两个句子完美匹配。通常也可以将 BLEU 的值乘以 100 来作为指标。BLEU 的原理是通过对预测句子的 N-grams 的与实际句子的 N-grams 词进行匹配，计算出各阶 N-grams 词的精度，由于预测的句子越短，匹配的难度越容易，所以引入了简短惩罚因子（Brevity Penalty，bp）。BLEU 的计算公式如式（11）所示。

$$BLEU = BP * \exp\left(\sum_{i=1}^{N} w_n \log P_n\right) \quad (11)$$

doi:10.3772/j.issn.2095-915x.2022.06.010

其中，N 最大值取 4，即 4-grams。BP 为惩罚因子表达式如式（12）所示，P_n 为 N-grams 精度表示式如式（13）所示。

$$\mathrm{BP}=\begin{cases}1 & \text{if } c>r\\ e^{(1-r/c)} & \text{if } c\le r\end{cases} \quad (12)$$

其中 c 是机器翻译句子的长度，r 是实际句子的长度。

$$P_n=\frac{\sum_{n-gram\in\hat{y}} CounterClip\ (n-gram)}{\sum_{n-gram\in\hat{y}} Counter\ (n-gram)} \quad (13)$$

其中，分母为机器翻译句子中 n-gram 词组的总数，分子为机器翻译句子中的词组能与实际句子匹配得上的个数。

3.6 实验结果

本文使用的模型为基于语义信息共享的 Transformer 模型，为了验证本模型的有效性，使用相同的数据集在以下几个经典模型上进行实验，此外本文还对比了相关研究中所提文献的实验结果。

（1）以单层 GRU[24] 构成 seq2seq 模型。

（2）以两层 LSTM+Attention 构成的 seq2seq 模型。

（3）基准的 Transformer 模型。

（4）seq2seq+Copy+Local Attention 模型 [16]。该模型是由 Zhang 等提出具有复制机制和局部注意力机制的端到端的神经网络模型。

（5）Transformer+Augment 模型 [17]。该模型是由 Liu 等提出的基于数据增强的 Transformer 模型。

（6）Time-Aware Ancient Chinese Text Translation 模型 [18]。该模型是由 Chang 等提出的具有时间感知的古文翻译模型。

（7）Seg+sub+stage-M5 模型 [19]。该模型是由魏家泽提出的基于外部知识协同的古文翻译模型，是通过注释信息协同（Sub）、句内片段协同（Seg）和语言知识协同（Stage）综合起来的三维联合协同的翻译模型。

（8）语义信息共享的 Transformer 模型，即本文所提出的模型。

各模型的实验结果如表 9 所示。从实验的结果可以看出，两层 LSTM+Attention 模型比单层的 GRU 模型的 BLEU 值高出了 11.57，证明更深的网络和 Attention 机制的引入能给模型带来巨大的提升。基准 Transformer 的机器翻译模型比两层 LSTM+Attention 模型的高了 1.78 的 BLEU 值，这是因为 Transformer 对比 LSTM 有更强的序列建模能力和全局信息感知能力，Transformer 能更好的提取句子的语义信息，在语义表达上也更强。语义信息共享的 Transformer 比基准 Transformer 的 BLEU 值高了 13.41，能取得如此大幅度提升的原因在于，共享词表和嵌入层参数能有效利用古文与现代文是同一语种的优势。此外，本文所提出的模型效果也高于其他文献中所提出的模型，证明了使深度学习模型自动学习和利用古文与现代文之间语义相同的特征比通过对模型使用复制机制或引入外部知识更有效果。从人工翻译的角度来看，不同语种之间的翻译比同一语种内的翻译要困难得多，因为同一语种不但使用同一种文字，还共享很多意思相

同或相近的词。共享词表与嵌入层参数即对源语言和目标语言使用了同一种表达，更容易获得两者之间的联系。

同一句古文使用不同的模型进行翻译，更能看出各模型的翻译效果，由于缺少其他文献的样本数据，本文只展示传统模型及本文所提模型的翻译样本。样本展示如表 10 所示。

表 9　实验结果

Model	Bleu
GRU	4.49
LSTM+Attention	16.24
基准Transformer	18.02
seq2seq+Copy+Local Attention	26.95
Transformer+Augment	27.16
Time-Aware Translation	24.51
Seg+sub+stage-M5	25.51
语义信息共享的Transformer	31.43

表 10　模型翻译样本展示

Source	众议始定。
Reference	这样，众公卿才作了决定。
GRU	众人商议。
LSTM+Attention	众人商议确定。
基准Transformer	众人的议论才确定。
语义信息共享的Transformer	众人的议论这才确定。
Source	孝文三年坐不敬，国除。
Reference	孝文三年因不敬罪。被除国。
GRU	孝文帝三年三年不<unk>，，封国被废除。
LSTM+Attention	孝文帝三年不敬，封国被废除。
基准Transformer	孝文帝三年因不恭敬获罪，封国被废除。
语义信息共享的Transformer	孝文帝三年前，他犯有不敬之罪，封国被废除。
Source	出，使子贡脱骖而赙之。
Reference	出来后，让子贡解下一匹骖马来给他作丧礼。
GRU	子贡出来，子贡子贡<unk><unk><unk><unk><unk>。
LSTM+Attention	出去后，派子贡脱去骖马送给他。
基准Transformer	出来后，让子贡解下一匹骖<unk>给他作丧礼。
语义信息共享的Transformer	出来后，让子贡解下骖马送给丧家。
Source	孤曰：吾兄居长，自应继位，我安可越次而处大业。
Reference	互巫说：我的哥哥年长，自然应该继位，我怎能超越次序而继大业?
GRU	<unk>说：<unk><unk>，，<unk>，，<unk>，，<unk><unk><unk>。
LSTM+ATTENTION	我说：我哥哥年纪长了，自然应该继承王位，我怎么能超越次序而居大业。
基准Transformer	我说：我的哥哥是年长的，自然应该传位，我怎么能越次序居大业呢?
语义信息共享的Transformer	孤说：我兄长年长，自然应当继承职位，我怎么能超越等级而成就大业呢?
Source	耳不可赡，目不可厌，口不可满；身尽府种，筋骨沈滞，血脉壅塞，九窍寥寥，曲失其宜，虽有彭祖，犹不能为也。
Reference	耳朵的欲望不可满足，眼睛的欲望不可满足，嘴巴的欲望不可满足；当身体全浮肿，筋骨沉滞，血脉阻塞，九窍空虚，全都丧失了它们正常的机能，虽然有彭祖在世，也是无能为力的。
GRU	<unk>不能<unk>，，<unk><unk><unk>，<unk><unk><unk><unk>，<unk><unk><unk><unk><unk><unk><unk><unk><unk><unk><unk><unk><unk><unk><unk><unk><unk><unk><unk><unk><unk><unk><unk>，<unk><unk><unk><unk>，<unk><unk><unk><unk>，也
LSTM+Attention	耳朵不可以满足，眼睛不能满足，口不能满足；身体<unk>，筋骨<unk>，筋骨<unk>，九窍<unk>，<unk><unk>，即使有彭祖，也不能做。
基准Transformer	耳朵不可能满足，眼睛不可以满足，嘴不可以满足；人身上皮肤<unk>，筋骨<unk>，血脉凝滞，九窍寥寥，架设不了几年即使有彭祖，他也不会去做。。
语义信息共享的Transformer	人的耳朵不能满足，眼睛不能厌恶，嘴不能让它满足；人的身体可以全部于官府播种，筋骨长期衰弱，血脉壅塞，九窍稀少，<unk>的功能不相适应，即使有彭祖，也不能用。

doi:10.3772/j.issn.2095-915x.2022.06.010

从表中可以看出基于语义信息共享的Transformer模型的翻译效果最佳，特别是在对长句子的翻译上。虽然本文提出的模型在古文翻译上要优于其他模型，但是还是存在一些缺点。如在对句子“孝文三年坐不敬，国除。”的翻译中，模型翻译为“孝文帝三年前，他犯有不敬之罪，封国被废除。”，其中由于缺了上文，模型未能识别出主语，而是将“孝文”作为了主语翻译成了“孝文帝”，而实际上“孝文”在此处是和“三年”一起组成“孝文三年”指代时间。此外，模型在处理通假字时也会出现误翻，如“身尽府种”，模型将其翻译为“人的身体可以全部于官府播种”，实际上，“府种”为“浮肿”的意思，府，通“肘”。种，通“肿”。出现这些情况的原因是模型往往会将古汉语按照最常用的方式去进行翻译，除了古文中的这些特殊情况外，总体来说，本模型的翻译效果良好，可以帮助人们更好的阅读古文。

4 结论

本文提出的基于语义信息共享的Transformer机器翻译模型在对古文的翻译中，优于传统seq2seq模型、基准Transformer模型以及相关研究中所提文献的模型。本文主要基于基准Transformer做了两点改变，一是在数据处理阶段源语言和目标语言共享同一个词表，二是在模型训练阶段编码器和解码器共享嵌入层参数。通过对实验结果的分析，本模型的翻译效果良好，但在处理古文的通假字和缺乏上下文等特殊情况上还有待改进。在后续研究中应针对上述问题进行探讨，以更进一步提高译文的质量与准确性。

参考文献

[1] 李业刚，黄河燕，史树敏，等．多策略机器翻译研究综述[J]. 中文信息学报，2015, 29(2):1-9.

[2] Bengio Y, Ducharme R, Vincent P. A neural probabilistic language model[J]. Advances in Neural Information Processing Systems, 2000, 13: 1137-1155.

[3] Della Pietra V J. The mathematics of statistical machine translation: Parameter estimation[J]. Using Large Corpora, 1994:223.

[4] Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv preprint arXiv: 1406. 1078, 2014.

[5] Elman J L. Finding structure in time[J]. Cognitive science, 1990, 14(2):179-211.

[6] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks [J]. Advances in Neural Information Processing Systems, 2014, 27.

[7] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv: 1409. 0473, 2014.

[8] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30.

[9] Papineni K, Roukos S, Ward T, et al. Bleu: a method for automatic evaluation of machine translation[C]. Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002:311-318.

[10] Liu X, Hoede C. Translation: an example from ancient Chinese to modern Chinese[J]. 2002.

[11] 王爽，熊德兰，王晓霞．古文翻译系统的设计与实现[J]. 电脑知识与技术，2009, 5(4):855-856+867.

[12] 韩芳，杨天心，宋继华．基于句本位句法体系的古汉语机器翻译研究[J]. 中文信息学报，2015, 29(2):103-110+117.

[13] 郭锐，宋继华，廖敏．基于自动句对齐的相似古文句子检索[J]. 中文信息学报，2008, (2): 87-91+105.

[14] 王爽，熊德兰，王晓霞．基于实例的古文机器翻译

设计与实现 [J]. 许昌学院学报 , 2009, (5): 88-91.

[15] 杨钦 . 文言文翻译及阅读理解关键技术的研究 [D]. 哈尔滨：哈尔滨工业大学 , 2015.

[16] Zhang Z, Li W, Su Q. Automatic translating between ancient Chinese and contemporary Chinese with limited aligned corpora[C]. CCF International Conference on Natural Language Processing and Chinese Computing. Springer, Cham, 2019:157-167.

[17] Liu D, Yang K, Qu Q, et al. Ancient–modern chinese translation with a new large training dataset[J]. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 2019, 19(1):1-13.

[18] Chang E, Shiue Y T, Yeh H S, et al. Time-Aware Ancient Chinese Text Translation and Inference[J]. arXiv preprint arXiv: 2107. 03179, 2021.

[19] 魏家泽 . 基于外部知识协同的古文到现代文机器翻译研究 [D]. 北京：中国科学技术信息研究所 , 2020.

[20] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553):436-444.

[21] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86(11):2278-2324.

[22] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016:770-778.

[23] Ba J L, Kiros J R, Hinton G E. Layer normalization[J]. arXiv preprint arXiv:1607. 06450, 2016.

[24] Chung J, Gulcehre C, Cho K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[J]. arXiv preprint arXiv: 1412. 3555, 2014.