基于古籍白话译本的古文机器翻译研究[1]

魏家泽[1]，何彦青[1]*，董诚[1]，洪涛[2]，苏瑞欣[2]

（1. 中国科学技术信息研究所情报理论与方法研究中心，北京，100038）
（2. 古联（北京）数字传媒科技有限公司，北京，100073）

【摘要】古文献具有极高的文学价值与历史价值，古文机器翻译有助于促进文化传播和古文外译。目前古文机器翻译研究仍然存在语料稀缺、语言风格迥异以及一词多义等问题。本文借助古籍白话译本中蕴含的丰富信息从三个维度来协同提升古文机器翻译效果：从古籍白话译本中提取高质量的句内互译片段用来扩充训练语料实现**句内片段协同**；对古籍所属朝代信息进行语言分期，借助 Fine-tune 微调技术训练上古期、中古期和近古期三个不同时期的翻译模型实现**语言分期协同**；从双语注释信息中提取古文词汇的注释信息，构建多义注释词典帮助翻译引擎实现**注释信息协同**。最后以翻译效果提升为标准对每个协同方法分别进行实验验证，在语料有限的情况下，三种协同方法均可以有效提升翻译效果。

关键词：注释信息 语言分期协同 句内片段
中图分类号： **G355** **文献标志码：** A

1. 前言

作为一个文明未有间断的国家，我国源远流长的历史，为后人留下了丰富而大量的历史文化遗产，其中包含卷帙浩繁的古文典籍。这些古文献具有极高的文学价值与历史价值。由于古人与现代人的语言习惯有较大差异，古文与现代文之间存在的语言障碍限制了古文献的广泛传播，因此人们通过古文今译来跨越这种古今语言的鸿沟。尽管学者已经对常见古文进行了翻译与注解，但仍存在大量古文没有白话译文，例如考古出土的典籍、族谱或佛道经文等。仅仅依靠人工翻译就会存在标准不一致、词语与句式错误等问题，因此开展对古文机器翻译的研究具有现实意义和应用价值。

在已有的古文机器翻译研究中，多数的研究着力于古文到现代文之间平行语料库的构建，以及将各种机器翻译模型应用到古文到现代文的翻译中，研究分布在对句子进行分词或分字的粒度选择以及加入外部字典知识指导翻译等。这些成果对古文今译的机器实现进行了卓有成效的探索。目前古文机器翻译研究仍然存在三个问题。1）**语料稀缺**。由于现在流行的机器翻译方法比较依赖双语平行语料，当语料不足时翻译系统生成译文的顺畅性、可读性都将受到影响。在通用领域的机器翻译研究中已有研究者借助单语语料来帮助机器翻译，此类方法在古文到现代文的翻译中还鲜有尝试。2）语言风格差异。我国历史悠久，各时期汉语语言风格迥异，无论是遣词还是造句，不同时期的古文差别很大。3）**一词多义（活用）**。如古文词汇“师”，在“三人行，必有我师”中译为“老师”，但在“齐师伐我”中应译为军队；再如古文词“三象”，有“三只大象”、“三象乐（古乐）”、“三位翻译人员”等三种词义相差甚远的译法。这样的一词多义仅仅依靠翻译系统很难进行甄别常常导致错译。

古籍白话译本是现代人对古籍所作的白话注解，译本中通常包含古文篇章及其现代文译文，另附有注解性信息，如词语注释、点评等。针对古文机器翻译的上述问题，本文借助古籍白话译本中蕴含的丰富信息从三个维度来协同提升古文机器翻译效果：双语句对、句内互译片段、双语注释信息以及古籍所属朝代信息等。高质量的句内互译片段用来扩充训练语料实现**句内片段协同**；对古籍

[1]**基金项目：**中国科学技术信息研究所重点工作项目 “俄汉跨语言知识发现与服务研究”（项目编号：ZD2020-10）与“面向垂直领域应用场景的机器翻译研究”（项目编号：ZD2020-18）。

* **通信作者：**heyq@istic.ac.cn

所属朝代信息进行语言分期，借助 Fine-tune 微调技术训练上古期、中古期和近古期三个不同时期的翻译模型实现**语言分期协同**；从双语注释信息中提取古文词汇的注释信息，构建多义注释词典帮助翻译引擎实现**注释信息协同**。最后以翻译效果提升为标准对每个协同方法分别进行实验验证。

本文将在第 2 章介绍古文机器翻译相关研究，第 3 章介绍古籍白话译本及抽取出的信息类型，第 4 章描述三维协同古文翻译方法，第 5 章展示实验结果与分析。最后在第 6 章进行总结。

2. 相关研究

自上世纪 60 年代，机器翻译的发展可大致分为四个阶段：规则机器翻译、实例机器翻译、统计机器翻译、神经机器翻译。目前效果最好的是神经机器翻译[1,2]，采用端到端的思想，分别对源语言和目标语言进行学习，最终通过具有数千万参数的神经网络将源语言句子直接翻译为目标语言句子，神经机器翻译模型训练的过程，即为找到该模型在整个平行语料上翻译概率最大化的参数的过程。

随着机器翻译方法的推进与改善，各种翻译模型逐次被应用到了古代汉语到现代汉语的翻译当中。基于规则的翻译方法[3,4]，翻译分为源语分析部分和译文生成部分，遵循语言学知识。源语分析部分，结合句法规则、词义消歧算法等对源语进行分析，分析结果表示为内部表示形式，译文生成部分将这种内部表示形式转化为目标语中的合法语句。这种规则方法针对特定的古代汉语文献准确率很高，但由于古代文献在不同的年代用词差异性很大，需要构建大量的规则，代价昂贵，且可能出现规则冲突情况。基于实例的方法[5, 6]需要有效的实例模板，找到较相似的实例模板，并对实例修改后得到译文，但当实例库的覆盖率无法保障时，效果较差。基于统计的方法、神经机器翻译方法都需要大量的语料进行训练，在古文今译的任务中语料较少，因而实际翻译效果一般。由于基于统计的方法需要单独训练多个功能模块完成翻译，语料较少时相较神经机器翻译效果更差。因此后两种方法应用在古文今译研究上，大多集中在特定词语翻译和单语语料的使用。张引、 陈琴菲[7]等通过 GIZA++工具训练得到古今词典后用其在神经机器翻译的译文上完成特定词语译文替换，高升等人[8]引入对《中医名词词典》的使用，确定古汉语专有名词的现代汉语解释后将其映射为词向量并与神经机器翻译的源端输出向量拼接形成指导向量完成翻译，该类方法具有启发性，但需要增加约束使《中医名词词典》词典信息切实引入到翻译过程中。吕建成[9]等人在古今汉语神经机器翻译模型训练中，训练 word2vec 语言模型构建古文近义词词典，完成源语言的近义词增广。杨钦[10]在统计机器翻译中对领域内外语料不同混合方式进行尝试后得出单语增益有限的结论。

学者对古代汉语的历史分期问题论述众多，但未曾达到一致意见，目前主要有“二分”,“三分”,“四分”,“五分”等几种观点。单侠[11]将分歧的产生总结为四个问题：分期标准（文体转变、语言内部发展规律等）、“文言”与“白话”纠葛、分期粗细差别（分期数量、大小时期）、是否需要划分出过渡期。以王力先生[12]和向熹先生[13]的划分方法为例，他们都对汉语采用了四分法，其中古代汉语占有三期：上古期、中古期、近古期，他们对各期的年代起止大致相同，但王力先生对每一期细分出一个过渡期，而向熹先生对每一期细分出三个小期。

与上述通过单语或词语优化的古文机器翻译方法不同，本文充分利用古籍白话译本的信息，分别提取出高质量句内片段、古文词汇注释词典来分别实现训练语料和词语翻译优化；借鉴王力先生和向熹先生的三分法对古汉语分期为：上古期、中古期、近古期，其中上古期为两汉及其之前各朝代，近古期为宋代之后各朝代，中古期为两期之间的各朝代，将训练语料分期来改善古文翻译的效果。

3. 古籍白话译本

古籍白话译本的原始信息格式，如图 1 所示。译本主要标签为：标题、正文、注释，其中标题

标签用于确定篇章边界或译文边界，正文标签表示古文或译文内容，注释标签指明词语注释内容边界。

```
<?xml version="1.0" encoding="UTF-8"?>
<语料 xmlns="http://shangyuan/shuju_yuliao" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
    <原书页面>ZSK10290-000009-L00001</原书页面>
    <标题 level="1" type="卷次" id="ZSK10290-000001-BT0000002" 提取标题="第一篇　立命之学" 是否提取="是" 唯一标识符=""
     是否显示="是">第一篇　立命之学</标题>
    <段落 first_indent="2" succeed_indent="" id="ZSK10290-000001-DL0000017">
        <正文>余童年丧父，老母命弃举业学医<注号 id="1">（1）</注号>，谓可以养生<注号 id="2">（2）</注号>，可以济人
        <注号 id="3">（3）</注号>，且习一艺以成名，尔父夙心也<注号 id="4">（4）</注号>。</正文>
    </段落>
    <标题 是否提取="否" 是否显示="是" type="校注标题" level="8" id="ZSK10290-000001-BT0000003">【注释】</标题>
    <段落 first_indent="2" succeed_indent="" id="ZSK10290-000001-DL0000018">
        <注文 type="注"><注号 id="5">（1）</注号>
        举业：为应科举考试而准备的学业，目的在于求取功名。科举考试是古代选拔官吏的制度。隋代设置进士科以试策取士。
        唐代沿用并增设秀才等十多种科目。因设立科目，以考试举士，故称科举。明、清时专指习八股文。</注文>
    </段落>
    <段落 first_indent="2" succeed_indent="" id="ZSK10290-000001-DL0000019">
        <注文 type="注"><注号 id="6">（2）</注号>养生：养活自己及家庭。这里的"生"指生活。</注文>
    </段落>
    <标题 是否提取="否" 是否显示="是" type="校注标题" level="8" id="ZSK10290-000001-BT0000004">【译文】</标题>
    <段落 first_indent="2" succeed_indent="" id="ZSK10290-000001-DL0000022"><正文 type="2">
    我童年时期就失去了父亲，老母亲让我放弃科举考试的学业而去学医，说学医可以养活家庭，同时也可以用医术来救济别人，
    而且精通一门手艺并以此成名，也是父亲平素的心愿。</正文>
    </段落>
    <标题 是否提取="否" 是否显示="是" type="校注标题" level="8" id="ZSK10290-000001-BT0000005">【点评】</标题>
    <段落 first_indent="2" succeed_indent="" id="ZSK10290-000001-DL0000023">
        <注释 type="注">
        《了凡四训》四篇是了凡先生对其子所说的话，即诫子家训，是长辈对晚辈的勉励和劝行之语。了凡先生起首便从自己的
        生平讲述，可谓"现身说法"。</注释>
    </段落>
```

图 1 古籍白话译本页面

Fig. 1 Page of Vernacular Translation of Ancient Books

以古籍白话译本《了凡四训》作为示例，从中抽取的信息如图 2 所示。按照标签提取互译篇章对进而得到互译句对，分别为古文句子和现代文译文。对于古文句子，句内通常含有多义词和稀缺词，译本中会对这些词汇给出今人注释。例如古文句子“余童年丧父，老母命弃举业学医，谓可以养生，可以济人”，其中古文词汇“养生”的今人注释为“养生：养活自己的家庭。这里的‘生’指生活。”。收集这些注释信息构建多义注释词典，可以作为外部知识有效补充信息的缺失。通过句子背景信息，如古籍书名或作者，确定其朝代及时期，完成分期协同。该句子所属为明代古籍《了凡四训》，明代为近古期，这样的朝代信息作为句子背景信息可以帮助语言分期。该古文句子和其现代文译文中包含有多个句内片段，例如“老母命弃举业学医”<=>“老母亲让我放弃科举考试的学业而去学医”，用这样的句内对译片段可以扩增训练语料。

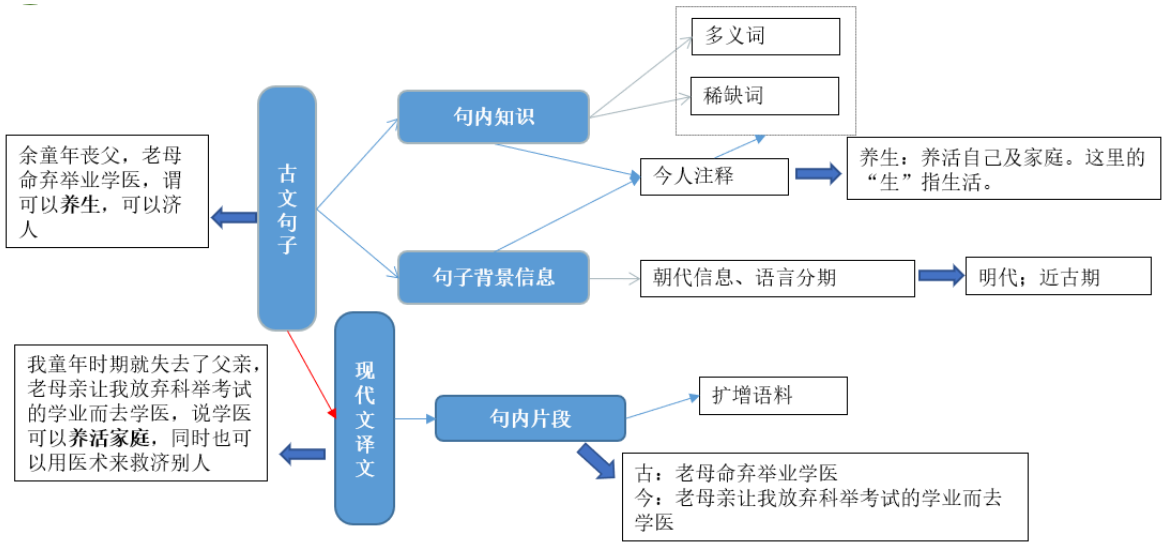

图 2 译本信息抽取

Fig. 2 Information Extraction of the Books

4. 三维协同古文翻译

三维协同翻译模型的流程图见图 3。首先从古籍白话译本中按照标签提取互译句对和注释信息。从注释信息中提炼注释，采用注释信息协同来生成训练语料。互译句对中包含有多个句内片段，利用句内片段协同来扩增训练语料。对所有古籍白话译本涉及的古籍，采用语言分期协同训练分期翻译引擎。对于待翻译译文，判断分期来调用分期翻译引擎，得到翻译译文。

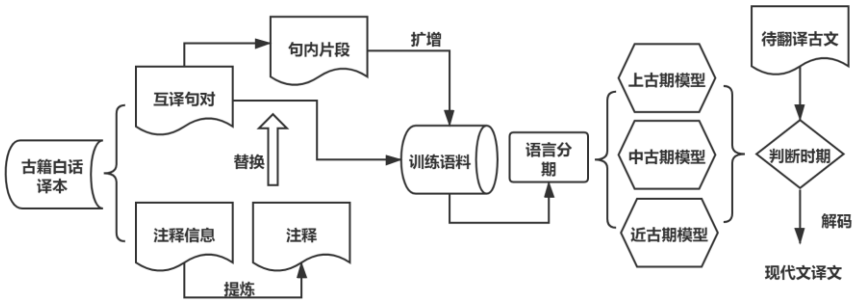

图 3 翻译模型架构

Fig. 3 Translation Model Architecture

4.1 注释信息协同

古籍白话译本中蕴含了大量的今人对古文词汇的注释信息，对这些注释信息进行提取与提炼，是对翻译系统的一个有益的外部知识的补充。以图 2 中的今人注释为例，古文词汇“养生”的注释信息为“养生：养活自己的家庭。这里的‘生’指生活。”。通常注释信息中的首句包含最有用的信息，因此为了确保对于古文词汇“养生”的译文的准确提取，把注释信息进行切句，仅保留首句“养生：养活自己的家庭。”。现代文译文中通常会增益，除词语译文外补充一些隐含的信息，这些信息在机器翻译中难以通过古文词汇学习得到，因此通过查找古文词汇“养生”和注释信息首句的最大公共子序列的方式，保留注释句中最有用信息，完成提炼。经过提炼后古文词汇“养生”得到其准确翻译为“养活家庭”。图 4 为注释信息协同流程。

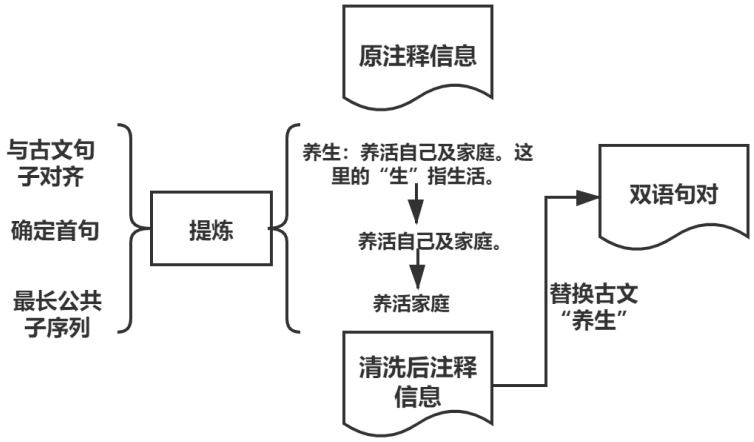

图 4　注释信息协同流程

Fig. 4 Collaborative Process of Annotation Information

对所有古籍白话译本中的古文词汇及其注释进行提取，进而归并去重，得到多义注释词典。该

词典中的每一个古文词汇，包含多个义项，每个义项具备清洗后注释信息、古文句、现代文句。对双语句对的古文词汇进行替换，用于训练翻译引擎。对于待翻译的译文中出现多义注释词典中的古文词汇时，如果被注释词含有多个义项，计算待翻译古文 s_1 与每个义项的古文句 s_2 的相似度来确定唯一义项，进而用提炼过的古文词汇注释进行替换。相似度计算选取 dice 方法。

表 1 注释词典

Tab1. Annotated Dictionary

古文词汇	义项 ID	注释信息	古文句	现代文句
师	1	军队	齐师伐我	齐国军队攻击我国
	2	老师	三人行，必有我师	三个人同行,一定有可以当我老师的
	3	学习	师夷长技以制夷	学习西方的先进技术来抵制西方

4.2 句内片段协同

由于训练语料不足，所以采用句内片段协同来扩充训练语料。句内片段协同流程，如图 5 所示。

利用互译句对训练一个古文到现代文的翻译模型 MTB。首先将互译句的古文句子和现代文句子都按逗号切分成短句，仅保留古文句子和现代文句子的短句数相等的互译句对。逗号为古今文中最常见符号，切分出的短句数量最多，短句数量相同，可以降低句内片段对齐的难度。其次，提取包含任意数量连续短句的古文片段 sega 和现代文片段 segm，利用正向翻译模型 MTB 将古文片段 sega 翻译为现代文 $trans_{sega}$。然后进行第一次筛选，将片段机器译文 $trans_{sega}$ 与所在句对的每个现代文片段 segm 依次计算相似度，保留其中相似度最大的现代文片段，将其作为一组候选互译片段，其中相似度仍为 Dice 相似度。每个互译句对通过古文片段可得到多组候选互译片段，仅保留其中相似度最高的一组，作为第二次筛选。第三次筛选，仅保留相似度阈值在 simi_t 上的互译片段。最终将原始互译句对与筛选出的互译片段 segam 共同作为语料进行训练。

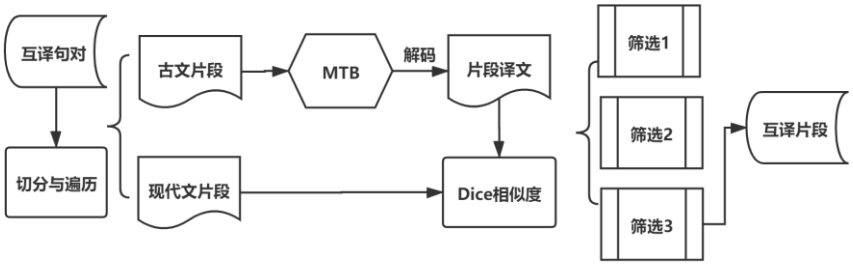

图 5 句内片段协同流程

Fig. 5 Collaborative Process of Intrasentence Fragment

4.3 语言分期协同

语言分期协同流程见图 5。在数据准备时期，首先确定各古籍白话译本所属朝代，进而将双语

句对按时期划分为三个分期子集，其中上古期为两汉及其之前各朝代，近古期为宋代之后各朝代，中古期为两期之间的各朝代。把所有双语语料训练得到翻译模型-父模型 MTF。对每个分期子集的双语句对，用各时期语料分别训练从现代文到古文的反向翻译模型，将其他期语料的现代文部分反向翻译得到该分期的伪语料。真实语料和伪语料混合后，通过微调（Fine-tune）父模型训练得到各分期的翻译模型-子模型 MTC。在解码时通过语言模型 LM 或语言分类工具确定待翻译句的时期信息，进而选定当期子模型用于翻译。

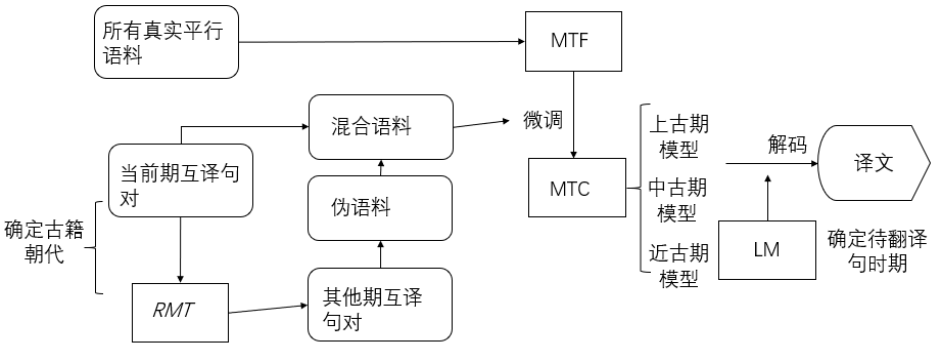

图 6. 语言分期协同流程

Fig. 5 Collaborative Process of Language Staging

5. 实验

5.1 数据准备与实验设置

本文选用《中华经典古籍库》中的 159 本古籍白话译本作为实验数据，从先秦至明清等各个时期，体裁未进行限定，诗、辞、杂说均有涵盖。对语料进行清洗时选取共同标签对比方法确定其中互译的篇章或段落，然后用哈工大自然语言处理工具 LTP[2]进行分句，并合并过短句子；进而进行句对齐步骤，共得到 35w 句对。

将 159 本古籍白话译本进行分期后，各期的书籍及句对数量如表 2 所示。由于中古期数据比较少，这里仅对上古期和近古期进行实验验证。将上古期数据和近古期数据都留出部分句对作为测试语料，其余数据作为训练语料；在训练 MTF 时将两份数据共同作为开发集，微调得到子模型 MTC 时，分别用当期测试集作为开发集。

表 2 分期语料数量统计

Tab2. Statistics on the Number of Staged Corpus

时期	书籍数	训练集	测试集
上古	56	13w	1.9k+
中古	31	5w	-
近古	72	16w	2k+
总计	159	34w	4k+

[2] https://www.ltp-cloud.com/

翻译系统为开源项目 fairseq-pytorch，所选神经机器翻译架构为 Transformer 架构，翻译基于分字进行，翻译模型训练的超参数选取常见设置，未对其进行进一步实验。每个模型使用 2 块 GPU 核进行训练，每个 batch 设置为 6000 字，词向量维度为 512，隐层状态维度为 2048。模型中采用 dropout 机制，dropout 设置为 0.1。初始学习率为 0.1，warmup 步数为 1000。模型训练轮数为 13，微调时额外用混合语料训练多个轮次；片段筛选的相似度阈值 simi_t 初始设置为 0.6，其代表片段对齐效果具有可信度。对实验结果进行评价，选取基于字的 BLEU 打分方式。

5.2 注释信息协同效果

语料中词汇替换规模如表 3 所示，经过替换的古文词汇具有一定规模。159 本古籍白话译本中提取注释信息经提炼后，统计信息如表 4 所示。表 5 展示了清洗后的注释信息。可以看出注释信息得到了有效提炼，与原注释信息相比，噪音少，信息明确。从表 4 字数上看，清洗前后注释信息总字数发生了较大变化。从表 5 实际效果上看，注释信息得到了有效提炼，与原注释信息相比，噪音少，信息明确；注释不单单是词与词的对应关系，包含其对古文词引申出额外的定语信息。

表 3 替换词汇规模

Tab3. Replaced words Size

语料集名称	抽取句对数量（句）	替换古文词汇数
训练集	34w+	19w+
测试集	4k+	5k+

表 4 注释规模

Tab4. Annotation Size

被注释词数量	清洗前注释字数	清洗后字数
92077	659w+	51w+

表 5 注释信息展示

Tab4. Annotation Exhibition

注释词	清洗前注释信息	清洗后注释信息
举业	为应科举考试而准备的学业，目的在于求取功名。	科举考试的学业
进学	科举时，童生参加岁试，被录取入府县学肄业，称为进学。	参加岁试入学
暗室屋漏	指别人看不见的地方，隐私之室。	别人看不见的私室

表 6 展示了基于注释信息协同的训练结果。其中 M1 为利用原始语料训练得到结果，M2 位利

用注释替换后的训练结果，M3 为将原始句子和替换后句子之间加标签作为源端句子进行训练结果。结果表明注释信息具有实际效果，且在本实验中注释信息能够替换原古文词汇。

表 6 注释信息协同训练结果

Tab6. Results of annotation information Collaborative Experiment

实验设置	Bleu4-char	
	上古期	近古期
M1	16.45	24.54
M2	**19.23**	**25.44**
M3	18.39	24.86

5.3 语料分期协同效果

语料分期协同实验结果如表 7 所示。其中 M1 为将三期真实语料共同训练结果；M2 为将混合语料直接进行训练结果， M3：在 M1 基础上进行微调后结果。表 6 结果显示，将语料分期进行针对性训练取得一定提升效果。其一，虽然古汉语应该进行语言分期，但各时期古汉语并非完全割裂，语言习惯存在继承现象，语句内容存在引用现象，因此近古期增益相对较少；其二，M3 相较 M2 结果进一步提升，证明真实语料仍具有重要价值。

表 7 语料分期训练结果

Tab7. The Results of Training by Stages

实验设置	Bleu4-char	
	上古期	近古期
M1	16.45	24.54
M2	17.13	24.64
M3	**18.83**	**25.36**

5.4 句内片段扩增效果

句内片段协同的实验结果如 8 表所示。M1 为基线系统，同样将三期真实语料共同训练；M2 系统将相似度阈值 simi_t 设置为 0.6，未限定各互译句对的互译片段组数，即未进行 4.3 节句内片段扩增的筛选 2 步骤；M3 系统相似度阈值 simi_t 设置为 0.6，并限定各互译句对的互译片段组数为 1；M4 系统限定各互译句对的互译片段组数为 1，保留相似度在[0.6, 0.7]之间句对；M5 系统限定组数，保留相似度在 0.7 以上句对。M6 进行了随机片段的对比试验，其中随机片段是指对任一片段随机选取子句作为译文，且为与 M2 进行区别，随机选取组内片段，同样限定互译片段组数为 1。表 8 结果说明句内片段在合适的设置下对翻译效果有提升。其一，M2 设置下，翻译效果不增反降，句内片段总数过多使模型更倾向于句内片段，失去了原有数据分布的真实性；其二，M4 与 M5 结果翻译效果均有所提升，差异不大，从数量上可知两者有 7w 句子为交集。其三，M6 设置结果同样提升，分析认为是目标端片段发挥了单语增强方式[14]类似作用。

表 8 句内片段协同实验结果

Tab8. Results of Intrasentence Fragment Collaborative Experiment

实验设置	片段数量	Bleu4-char	
		上古期	近古期
M1	-	16.45	24.54
M2	77w	15.91	24.38
M3	16w	16.43	24.81
M4	10w	**16.62**	**24.82**
M5	13w	16.58	24.80
M6	20w	16.59	24.79

5.5 联合效果

表 9 为将注释信息、句内片段和语料分期协同方法（Sub+seg+stage）联合使用后的结果，即将句内片段与注释信息替换后语料联合训练得到模型后，用分期混合语料进一步微调，其中 M3、M4、M5 均沿用表 8 中的相似度阈值设置。其中将两者联合使用时，注释信息与语言分期（Sub+stage）仅在上古期继续提升，句内片段与语言分期（Seg+stage-M4）仅在近古期继续提升，注释信息与句内片段（Sub+seg-M4）未能继续提升。三者联合使用后在上古期和近古期均进一步提升，从而更有效验证本文方法的有效性。

表 9 联合实验结果

Tab9. Joint experiment results

实验设置	Bleu4-char	
	上古期	近古期
Sub+stage	19.66	25.44
Seg+stage-M4	18.30	25.41
Sub+seg-M4	18.81	24.65
Seg+sub+stage-M3	19.57	25.23
Seg+sub+stage-M4	20.18	**25.06**
Seg+sub+stage-M5	**19.83**	25.51

6. 总结

本文提出基于古籍白话译本的古文到现代文翻译模型，利用从古籍白话译本中获取的三维信息系统来完成古文今译。在信息使用上，注释信息协同提炼古文词汇的精准注释，句内片段协同可以在句对齐性能有限的情况下筛选出高质量互译片段提升高质量片段在模型中的重要性；语言分期协同将分期翻译模型结合回翻方法构成混合数据，再辅以分期微调模型的设计，均提升了古文到现代

文的翻译效果。

由于古籍数量有限，研究还有很大提升空间。神经机器翻译对注释信息的吸收仅通过前处理替换方式，后续研究可以针对明确的外部知识设计不同方案。增加互译语料数量，在保证有效训练的前提下，通过一致性分析等，进一步比较上古期和近古期语料在实验中的不同表现。

参考文献

[1] Gehring J, Auli M, Grangier D, et al. Convolutional Sequence to Sequence Learning[J]. arXiv: Computation and Language, 2017.
[2] Vaswani A, Shazeer N, Parmar N, et al. Attention is All you Need[C]. Neural Information Processing Systems, 2017: 5998-6008.
[3] 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(04):855-856+867.
[4] 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.
[5] 郭锐. 基于实例的古汉语机器翻译研究[D]. 北京: 北京师范大学, 2007.
[6] 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, 28(05): 88-91.
[7] 浙江大学. 一种多特征融合的古今汉语自动翻译方法:CN201910033155. 8[P]. 2019-04-26.
[8] 北京邮电大学. 一种基于词典和 seq2seq 预训练机制的中医古籍翻译方法:CN201910020459. 0[P]. 2019-05-10.
[9] 四川大学. 一种基于神经网络的古文翻译方法:CN201910012805. 0[P]. 2019-05-21.
[10] 杨钦. 文言文翻译及阅读理解关键技术的研究[D]. 黑龙江: 哈尔滨工业大学, 2015.
[11] 单侠. 关于汉语史分期的一点思考[J]. 前沿, 2009, 000(002): 186-188.
[12] 王力. 汉语史稿[M]. 北京: 中华书局, 1980.
[13] 向熹. 简明汉语史[M]. 北京: 高等教育出版社, 1993.
[14] Sennrich R, Haddow B, Birch A. Improving Neural Machine Translation Models with Monolingual Data[C]. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016.

Research on machine translation of ancient Chinese based on modern Chinese versions of ancient books

Wei Jiaze[1], He Yanqing[1]，Dong Cheng[1]，Hong Tao[2], Su Ruixin[2]

(1. Institute of Scientific and Technical Information of China, Beijing, 100038, China; 2. Gulian (Beijing）Media Tech Co., Ltd., Beijing, 100073, China)

Abstract: The ancient literature has high literary value and historical value, Machine translation of ancient Chinese helps to promote cultural communication and foreign language translation of ancient texts. At present, there are still some problems in the research of machine translation of ancient Chinese, such as the scarcity of corpus, different language styles and polysemy. This paper makes use of the rich information contained in the modern Chinese versions of ancient books to improve machine translation of ancient Chinese from three dimensions. High quality inter-sentence bilingual fragments are extracted from the modern Chinese versions of ancient books to expand the training corpus and realize intra-sentence fragment collaboration. Language stages are carried out by using the dynasties information of ancient books , the translation models of ancient, middle ancient and near ancient times ， were trained to realize language staged coordination by the help of fine-tune technology. The annotation information of ancient Chinese words is extracted from the bilingual annotation information, and the polysemous annotation dictionary is built to help the translation engine realize the annotation information coordination. Each collaborative method was verified by experiments with translation effect enhancement. In the case of limited corpus, all the three collaborative methods could effectively improve the translation effect.

Keywords: annotation information; language staged collaboration; intra-sentence fragment