

基于长短时注意力机制的汉藏党政公文机器翻译

赞拉公^{1,2} 安见才让^{1,2}
ZAN Lagong AN Jiancairang

摘要

在研究汉藏党政公文机器翻译时,由于语言资源匮乏,在对党政公文领域的术语、短语及句子进行翻译时,专用术语和短句方面均暴露出错译、漏译、解释性翻译、口语化翻译以及用词不精准等问题。这些问题严重影响翻译质量,对汉藏党政公文的准确传递和理解起着负面作用。因此,构建高质量的汉藏党政公文语料库的同时融合和优化模型性能来提高翻译任务至关重要,文章通过融合 LSTM 和 Transformer 的长短时注意力机制的翻译模型来分别在音节、词、音节和词融合的分词方式上进行实验,实验结果表明,融合 LSTM 和 Transformer 的长短时注意力机制的翻译模型在音词融合分词方式上准确率分别在词和音节上提升了 3.14% 和 1.92%,此外,将翻译软件与阳光翻译、班智达翻译软件在翻译任务上进行对比后得知,翻译软件在汉藏党政公文领域的翻译效果达到了预期的目标。

关键词

长短时注意力机制;模型融合;汉藏党政公文;音词融合;翻译效果

doi: 10.3969/j.issn.1672-9528.2025.01.021

0 引言

汉藏翻译的历史悠久,自公元 823 年,蕃赞普赤祖德赞与唐朝唐穆宗时期立下长庆会盟碑文(又称“唐蕃会盟碑”)起,汉藏翻译活动日益频繁^[1]。这对促进汉藏两族人民间的文化交流,推动地区繁荣发展起到了积极的推动作用。

新中国成立后,为加速中国特色社会主义事业的发展进程,在西藏自治区和青海、甘肃、四川、云南四省设立了专门的汉藏翻译部门。然而,这些藏区的汉藏翻译部门面临着诸多难题,如公文翻译任务量大、对翻译速度要求高,同时还存在翻译专业人才短缺、人工成本高昂等问题。

此外,市面上汉藏机器翻译软件的研发成果颇为丰富,可在现有的汉藏机器翻译实际应用里,专用术语、短句方面存在着错译、漏译、解释性翻译、口语化翻译、用词不精准等问题。因此,本文利用 LSTM 的长期依赖处理能力和 Transformer 的高效并行计算能力来提出了融合 LSTM 和 Transformer 模型的基于长短期注意力机制的方法来提高汉藏机器翻译模型。

1 相关研究

如今,众多研究人员积极开展汉藏机器翻译的研究工作,

使翻译性能与效果均取得显著提升。且研究主要围绕藏文文法知识的融合展开^[2]。华果才让等人^[3]在分析针对译文中的藏文虚词错误问题,提出藏文虚词纠错模型,提高了汉藏机器翻译中的藏文虚词的纠错性能。孙义栋等人^[4]针对翻译效果差的问题,通过改进藏汉词表来提升汉藏双向的翻译性能。杨丹等人^[5]针对稀疏的藏汉平行语料,利用回译的方法来增强藏汉机器翻译的数据集。严松思等人^[6]应用藏文文法特点,提出了基于藏文虚词知识的融合方法。唐超超等人^[7]针对考虑语言和数据资源的差异性,通过比较方式对 3 种不同的子词处理方法在汉藏机器翻译上进行实验,提出了不同子词处理策略对翻译性能。桑杰端珠^[8]提出应用基于词典注入方法来提升预训练模型。汪超^[9]针对缺少高质量的平行语料稀缺问题,提出了同义词替换和回译方法来缓解语料稀缺问题,并在 Transformer 和 XLM-R 实验中准确率分别提升了 0.86、0.83。刘丁^[10]通过 Transformer 上训练好的三千万中英文语料模型迁移到汉藏机器翻译上提升了翻译模型效果。

待翻译的专业性文本在通用的汉藏数据集上训练出的模型中进行测试时,翻译出来的译文效果存在美中不足的问题,因此,本文从国家党政机构发布的政府汉藏党政公文语料中,采集国家层面的如中国共产党第二十次全国代表大会报告汉藏双语版等和省级层面的如 2023 年青海省人民政府工作报告汉藏双语等,因收集的藏文译文由国家相关行政部门发布,故译文的安全性、规范性和准确性都有可靠的保障。通过汉藏党政公文领域收集了大量的高质

1. 青海民族大学智能科学与工程学院 青海西宁 810007
2. 青海民族大学藏文信息处理与软件研究所 青海西宁 810007
[基金项目]青海民族大学 2023 年度校级本硕博(学生)项目“基于深度学习的汉藏政府公文机器翻译技术研究”(09M2023005)

表 3 本实验的翻译软件与其他软件在汉藏党政公文领域的翻译任务上的效果对比实例

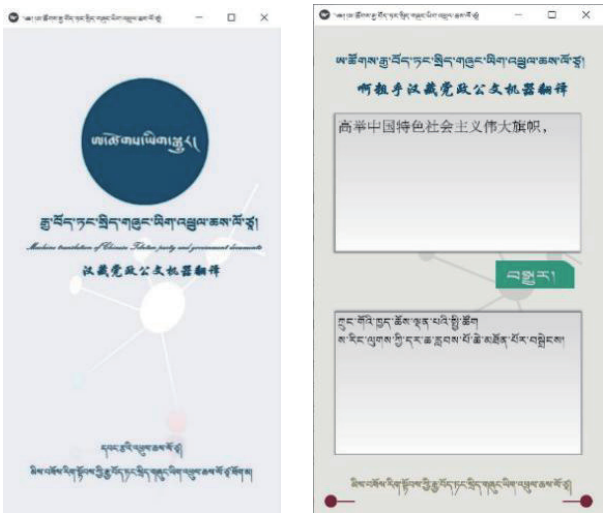
实例	原文	阳光藏汉机器翻译	班智达翻译	汉藏公文党政机器翻译
1	我们坚持加强党的全面领导和党中央集中统一领导	ང་ཚོས་ཏང་གི་ཚྭ་གསལ་ཡོངས་ཀྱི་འགོ་ཁྲིད་དང་ཏང་ཁྲུང་དབྱེང་གི་གཞིགས་སྒྲིག་ལྟི་བུ་འགོ་ཁྲིད་ལ་གྲགས་ཐྱོན་ཆུག་ཆུག་འཕམ་འཕྲོང་ལ།	ང་ཚོས་ཏང་གི་ཚྭ་གསལ་ཡོངས་ཀྱི་འགོ་ཁྲིད་དང་ཏང་ཁྲུང་དབྱེང་གི་གཞིགས་སྒྲིག་ལྟི་བུ་འགོ་ཁྲིད་ལ་གྲགས་ཐྱོན་ཆུག་ཆུག་འཕམ་འཕྲོང་ལ།	ཏང་གིས་ཚྭ་གསལ་ཡོངས་ནས་ཏང་ཁྲུང་དབྱེང་འགོ་ཁྲིད་དང་ཏང་ཁྲུང་དབྱེང་གི་གཞིགས་སྒྲིག་ལྟི་བུ་འགོ་ཁྲིད་ལ་གྲགས་ཐྱོན་ཆུག་ཆུག་འཕམ་འཕྲོང་ལ།
2	我们对新时代党和国家事业发展作出科学完整的战略部署	ང་ཚོས་དུས་རབས་གསར་པའི་ཏང་དང་རྒྱལ་ཁབ་ཀྱི་བྱ་གཞི་འཕེལ་རྒྱས་གཏོང་བར་ཚན་རིག་དང་མཐུན་ཞིང་ཆ་ཚང་པའི་འཕམ་ཆུས་ཀྱི་བོད་སྒྱུ་བྲལ་ཡོད།	ང་ཚོས་དུས་རབས་གསར་པའི་ཏང་དང་རྒྱལ་ཁབ་ཀྱི་བྱ་གཞི་གཏོང་བར་ཚན་རིག་དང་མཐུན་ཞིང་ཆ་ཚང་པའི་འཕམ་ཆུས་ཀྱི་བོད་སྒྱུ་བྲལ་ཡོད།	དུས་རབས་གསར་པའི་ཏང་དང་རྒྱལ་ཁབ་ཀྱི་བྱ་གཞི་འཕེལ་རྒྱས་གཏོང་བར་ཚན་རིག་དང་མཐུན་ཞིང་ཆ་ཚང་པའི་འཕམ་ཆུས་ཀྱི་བོད་སྒྱུ་བྲལ་ཡོད།
3	构建高水平社会主义市场经济体制	ཆུ་ཚད་མཐོ་པའི་སློ་ཚོགས་རིང་ལུགས་ཀྱི་ཚོང་རེའི་དཔལ་འབྱོར་སློག་སྒྲིག་འཛུགས་པ་དང་།	ཆུ་ཚད་མཐོ་པའི་སློ་ཚོགས་རིང་ལུགས་ཀྱི་ཚོང་རེའི་དཔལ་འབྱོར་སློག་སྒྲིག་འཛུགས་པ་དང་།	ཆུ་ཚད་མཐོ་པའི་སློ་ཚོགས་རིང་ལུགས་ཀྱི་ཚོང་རེའི་དཔལ་འབྱོར་སློག་སྒྲིག་འཛུགས་པ་དང་།
4	坚持致力于为非公经济发展营造良好环境和提供更多机会的方针政策	སློལ་དབང་བ་མེད་པའི་ལམ་ལུགས་ཀྱི་དཔལ་འབྱོར་གོང་དུ་སྒྲིལ་བར་ཁོར་ཡུག་བཟང་པོ་སྒྲུབ་པ་དང་གོ་སྐབས་ཐུང་ལས་མང་བ་མཁོ་འདོན་བྱེད་པར་འབད་བརྩོན་བྱ་བའི་བྱེད་སྒྲུབ་དང་ཕྱིར་ཐོང་ཆུས་ཆུག་འཕམ་འཕྲོང་ལ་བྱ་དགོས།	སློལ་དབང་བ་མ་ཡིན་པའི་ལམ་ལུགས་ཀྱི་དཔལ་འབྱོར་འཕེལ་རྒྱས་གཏོང་བར་ཁོར་ཡུག་བཟང་པོ་སྒྲུབ་པ་དང་གོ་སྐབས་ཐུང་ལས་མང་བ་མཁོ་འདོན་བྱེད་པའི་བྱེད་སྒྲུབ་དང་ཕྱིར་ཐོང་ཆུས་ཆུག་འཕམ་འཕྲོང་ལ་བྱ་དགོས།	དཔལ་འབྱོར་ལ་དབང་བ་མ་ཡིན་པའི་ལམ་ལུགས་ཀྱི་དཔལ་འབྱོར་འཕེལ་རྒྱས་གཏོང་བར་ཁོར་ཡུག་བཟང་པོ་སྒྲུབ་པ་དང་གོ་སྐབས་ཐུང་ལས་མང་བ་མཁོ་འདོན་བྱེད་པའི་བྱེད་སྒྲུབ་དང་ཕྱིར་ཐོང་ཆུས་ཆུག་འཕམ་འཕྲོང་ལ་བྱ་དགོས།
5	中国式现代化是物质文明和精神文明相协调的现代化	ཁྱེད་ཀྱི་རྒྱལ་པོ་ལྟར་པའི་དེང་རབས་ཅན་གྱི་དངོས་པོའི་དཔལ་ཡོན་དང་དབང་པོའི་དཔལ་ཡོན་ལས་ཚན་མཐུན་ཆུང་བྱེད་པའི་དེང་རབས་ཅན་གྱི་ལུ་ལ།	ཁྱེད་ཀྱི་རྒྱལ་པོ་ལྟར་པའི་དེང་རབས་ཅན་གྱི་དངོས་པོའི་དཔལ་ཡོན་དང་དབང་པོའི་དཔལ་ཡོན་ལས་ཚན་མཐུན་ཆུང་བྱེད་པའི་དེང་རབས་ཅན་གྱི་ལུ་ལ།	ཁྱེད་ཀྱི་རྒྱལ་པོ་ལྟར་པའི་དེང་རབས་ཅན་གྱི་དངོས་པོའི་དཔལ་ཡོན་དང་དབང་པོའི་དཔལ་ཡོན་ལས་ཚན་མཐུན་ཆུང་བྱེད་པའི་དེང་རབས་ཅན་གྱི་ལུ་ལ།

政公文领域的术语、短语、句子翻译时，专用术语、短句方面未存在着错译、漏译、解释性翻译、口语化翻译、用词不精准等问题。从而汉藏党政公文领域的汉藏机器翻译的翻译效果得到了预期的效果。

3.4 App 功能实现

本文汉藏党政公文机器翻译 App 是基于 Python 的一种 GUI 图形用户界面工具来开发完成^[13]，kivy 是 Python 内置函数其应用于开发 App 程序的模块，目前应用 Python 编程语言里能够开发 App 应用程序的只有 Kivy 模块^[14]。它不仅语法简单，而且可移植性较强^[15]。

启动开发的汉藏党政公文翻译 App 后，就会出现汉藏党政公文机器翻译的首页，首页上显示着有关翻译的软件的信息，用户点击首页任意位置进入 App 主界面，也就是实现翻译功能的页面，主界面上有一个输入框、按钮、输出框，用户将待翻译的内容输入至输入框，随后点击翻译按钮（བསྒྲུབ་པ་），软件把内容发送到融合 LSTM 和 Transformer 的长短时注意力机制的翻译模型中进行翻译，最后把最优的译文内容输出到翻译结果显示框中。如图 2 所示。



(a) App 首页 (b) App 主界面

图 2 汉藏党政公文翻译 App 软件界面

4 总结

针对贫语言资源条件有限而翻译效果欠佳的汉藏机器翻译的研究，本文应用融合 LSTM 和 Transformer 的长短时注意力机制的翻译模型在汉藏党政公文领域的数据集上采用了音节、词、音节和词融合的分词方式进行译文效果比较，从实验的结果得知融合 LSTM+Transformer 的长短时注意力机制在音节、词、音节和词融合的实验结果的准确率分别是达到 72.9%、74.12%、76.04%。音词融合的分词模式跟音节和词分词的模式相比，实验中融合 LSTM 和 Transformer 的长短时注意力机制的翻译模型在音词融合分词方式上准确率分别提升了 3.14% 和 1.92%，而且本实验的翻译软件与西藏大学的阳光汉藏翻译软件和青海师范大学的班智达翻译软件对比得知，党政公文领域的汉文句子在翻译任务上进行测试时，本翻译软件所译的译文内容中存在极个别漏译和一个错译之外，其余的内容都能准确翻译。在之后的研究中，为了能够更好地提高融合 LSTM+Transformer 的长短时注意力机制的汉藏翻译模型性能，以扩建高质量的汉藏党政公文语料库为基础，融合多种高性能的模型来提高和辅助翻译质量。

参考文献：

[1] 达哇才让. 党政文献民族语文翻译难点及策略研究 [J]. 民族翻译, 2023(2): 22-27.

[2] 头旦才让. 汉藏神经机器翻译关键技术研究 [D]. 西藏: 西藏大学, 2021.

[3] 华果才让, 班玛宝, 桑杰端珠, 等. 面向汉藏机器翻译后处理的藏文虚词纠错模型 [J]. 计算机仿真, 2021, 38(12): 391-396.

[4] 孙义栋, 拥措, 杨丹. 基于 VOLT 的藏汉双向机器翻译 [J]. 计算机与现代化, 2022(5): 28-32+39.

[5] 杨丹, 拥措, 仁青卓玛, 等. 基于 mRASP 的藏汉双向神经机器翻译研究 [J]. 计算机技术与发展, 2023, 33(12): 200-206.

[6] 严松思, 珠杰, 汪超, 等. 藏文虚词知识融合的藏汉机器

一种宽带距离像互相关测速的改进方法

柴许杨^{1,2} 张健³ 刘野³
CHAI Xuyang ZHANG Jian LIU Ye

摘要

距离像互相关作为宽带测速领域的有效方法之一,能够精准地实现测速功能。然而,其对于测速精度有着严苛要求,必须依托大点数快速傅里叶变换技术予以支撑,方能达成预期的高精度测速目标,基于此,文章提出了一种基于线性调频Z变换算法的宽带距离像互相关测速方法,首先建立宽带去斜回波模型;然后推导相邻脉冲的距离像互相关算法原理以及峰值点位置和速度的映射关系;最后利用雷达预测速度的先验信息,在小速度区间内进行速度测量,仿真和实测结果验证了该方法不仅降低了快速傅里叶变换点数,同时也改善了测速精度。

关键词

测速;去斜处理;距离互相关;线性调频Z变换;FFT

doi: 10.3969/j.issn.1672-9528.2025.01.022

0 引言

随着雷达技术的迅猛发展,信号带宽也随之拓展,为雷达系统性能的提升注入强大动力,同时对精确测量也提出了更高要求^[1]。因此,如何提高雷达测速精度是掌握目标运动特性的关键因素之一^[2]。

文献[3]利用最小波形熵算法对速度进行最优搜索,然

而,由于其计算过程繁杂、工作量巨大,并且在低信噪比的恶劣条件下,进行全局寻优时极易遭遇困境,往往难以达成预期目标,最终导致失败结局。文献[4]分析了时域下互包络互相关和复包络互相关算法的区别以及影响测速性能的因素,得出增加FFT点数可以提高测速精度。文献[5]创新性地提出了针对正弦波频率估计的修改Rife算法。这一算法在理论层面展现出了极高精度优势,其均方根误差能够极为接近克拉美-罗限,为高精度的频率估计提供了有力支撑,但其运算量颇为庞大。文献[6]和[7]对比了频域、时域互相关算法。频域互相关精度高,适慢速目标,但其无模糊速度范

1. 中国电子科技集团公司第三十八研究所 安徽合肥 230088
2. 孔径阵列与空间探测安徽省重点实验室 安徽合肥 230088
3. 中国人民解放军 63726 部队 宁夏银川 750000

- 翻译方法研究[J]. 中央民族大学学报(自然科学版), 2024, 33(1): 20-27.
- [7] 唐超超, 拥措, 仁青卓玛. 不同子词处理方法对机器翻译的影响研究[J]. 电脑知识与技术, 2024, 20(12): 4-7+11.
- [8] 桑杰端珠. 基于深度学习的藏汉(汉藏)机器翻译关键技术研究[D]. 青海: 青海师范大学, 2023.
- [9] 汪超. 基于数据增强技术的藏汉机器翻译方法研究[D]. 西藏: 西藏大学, 2023.
- [10] 刘丁. 基于迁移学习的低资源神经机器翻译方法研究[D]. 兰州: 西北民族大学, 2021.
- [11] 郑鑫, 陈海龙, 马玉群, 等. 融合依存句法和LSTM的神经机器翻译模型[J]. 哈尔滨理工大学学报, 2023, 28(3): 20-27.
- [12] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]//NIPS'17: Proceedings of the 31st Interna-

tional Conference on Neural Information Processing Systems. New York: CAI, 2017: 6000-6010.

- [13] 明日科技. Python+Kivy (App 开发) 从入门到实践[M]. 长春: 吉林大学出版社, 2020.
- [14] 赞拉公, 安见才让. 基于Python+Kivy的汉英词典实现[J]. 电脑编程技巧与维护, 2023(8): 29-31.
- [15] 刘凌霄, 郝宁波, 吴海涛. 21 天学通 Python[M]. 北京: 电子工业出版社, 2016.

【作者简介】

赞拉公(1997—), 男, 青海海南州人, 硕士, 研究方向: 藏文信息处理及应用。

安见才让(1969—), 男, 青海西宁人, 硕士, 教授, 研究方向: 藏文信息处理和分布式网络。

(收稿日期: 2024-10-13)