

基于数据增强技术的藏汉机器翻译方法研究

索朗措姆, 索朗旺堆

西藏日报社, 西藏 拉萨 850000

DOI: 10.61369/ETR.12312

摘 要 : 藏汉机器翻译在促进藏汉文化传播和交流、推动藏区信息化建设、促进藏区经济发展等方面发挥着重要的作用, 扮演着重要的角色。然而由于缺乏藏汉平行语料、藏汉机器翻译性能不足等问题的影响, 导致藏汉机器翻译的发展受到一定限制。而数据增强技术能够有效解决上述问题, 通过对原始数据进行扩展、变换等操作, 生成大量训练数据, 从而有效提升藏汉机器翻译质量和性能。对此, 本文就基于数据增强技术的藏汉机器翻译方法进行简要分析, 希望为广大读者提供一些有价值的借鉴和参考。

关 键 词 : 数据增强技术; 藏汉机器翻译; 伪平行语料

Research on Tibetan Chinese Machine Translation Method Based on Data Enhancement Technology

Solang Cuomu, Solang Wangdui

Xizang Daily, Lhasa, Xizang 850000

Abstract : Tibetan Chinese machine translation plays an important role in promoting the dissemination and exchange of Tibetan Chinese culture, advancing information technology construction in Tibetan areas, and promoting economic development in Tibetan areas. However, due to the lack of parallel Tibetan Chinese corpora and insufficient performance of Tibetan Chinese machine translation, the development of Tibetan Chinese machine translation has been somewhat limited. Data augmentation technology can effectively solve the above problems by expanding and transforming the original data to generate a large amount of training data, thereby effectively improving the quality and performance of Tibetan Chinese machine translation. In this regard, this article briefly analyzes the Tibetan Chinese machine translation method based on data augmentation technology, hoping to provide valuable references and guidance for readers.

Keywords : data augmentation technology; Tibetan-Chinese machine translation; pseudo-parallel corpus

引言

众所周知, 藏汉两种语言存在着显著的差异, 分别属于不同语系, 在语法结构、表达方式、词序等方面存在着巨大的差异, 这也给藏汉机器翻译增加了多种挑战。其中, 平行语料匮乏、翻译质量不高等问题严重影响藏汉机器翻译的发展。作为一种创新性技术方法, 数据增强技术能够通过生成伪平行语料、数据变换和拓展等方式, 增加训练数据, 有效提升藏汉机器翻译模型的性能, 有助于推动藏汉文化的交流和发展, 推动藏区信息化建设和经济发展。

一、数据增强技术概述

(一) 含义

数据增强技术是一种创新型技术, 它能够对已有数据进行变换和处理, 生成大量新的训练数据, 扩展训练数据集, 从而提升藏汉翻译模型的泛化能力和性能。

(二) 常见方法

1. 同义词替换

在文本数据中, 可以通过查找同义词词典或使用词向量模型

等方式, 将句子中的一些词替换成同义词, 从而生成大量新的句子。^[1]例如, 在具体翻译实践中, 可以将藏文中的某一些低频词汇换成语义相近的同义词, 通过这样的方式, 增加藏语句子的多样性, 不仅能够丰富词汇的表达, 同时, 还能够增加句子的数量。但在此过程中需要注意的是, 同义词替换需要充分考虑词语之间的差异, 确保替换后的句子在使用情境和含义方面保持一致, 避免引入歧义, 使整个语境发生改变。

2. 回译

回译是一种将目标语言先翻译成另一种语言, 之后再将其翻

译回原始语言的方法。在藏汉机器翻译过程中，通过运用回译方法，能够生成大量伪平行语料。^[2]具体来讲，可以先将藏文句子翻译成汉语句子，之后再将其翻译回藏文句子。通过这样的方式，能够得到两个句子，分别是藏文句子和原始藏文句子。尽管这两个句子可能并不完全一致，但两者在含义上相近或相同，能够作为平行语料进行使用。^[3]这种回译方法能够有效增加藏汉平行语料的数量，从而有效提升藏汉机器翻译模型的训练效果。然而，在回译的过程中也可能会存在一些翻译错误，导致句子含义存在较大偏差。因此，在使用回译方法时，需要对生成的伪平行语料进行有效筛选和检查，确保其质量。

2. 随机插入与删除

在句子中随机插入或删除一些词汇，能够有效改变句子的长度和结构。例如，在具体实践中，可以在藏文句子中随机插入一些词语或删除一些不影响句子含义的词汇，从而生成新的藏语句子。这种方法能够改变句子的长度和结构，但也可能会对句子的整体含义和语义的完整性进行破坏。^[4]因此，需要对删除和插入的比例进行合理控制。

3. 句式变换

句式变化主要是指对藏语句子的句式结构进行合理变换，比如说将被动语态转换为主动语态、将疑问句转变为陈述句等。在藏汉机器翻译过程中，通过将藏文语句中的一些句式变换，生成一些新的藏文句子。这样做能够使藏汉机器翻译模型学习多种表达方式和语句结构，从而有效提升翻译的准确性和流程性。

二、基于数据增强技术的藏汉机器翻译现状

（一）研究成果

近些年来，随着数据增强技术的不断发展和广泛运用，基于数据增强技术的藏汉机器翻译也取得了一些研究成果。^[5]例如，一些研究团队通过运用回译、同义词替换等方法，生成了大量的伪平行语料，从而有效提升了藏汉机器翻译模型的训练效果；还有一些研究团队通过利用随机插入与删除方法，生成大量新的藏语句子，从而有效增加了训练数据的多样性，提升了藏汉机器翻译的准确性。这些研究成果为基于数据增强技术的藏汉机器翻译发展奠定了坚实的基础。

（二）存在的问题

当前，基于数据增强技术的藏汉机器翻译依旧存在一些问题。对此，本文就以下几个方面进行简要阐述。

1. 数据质量较低

与汉语相比，藏语存在鲜明的特点，如语法机构较为复杂、词形变化更为丰富等，这导致在生成伪平行语料过程中，可能会存在语法错误、句子含义发生改变、语义偏差等问题，这些问题严重影响数据增强效果。^[6]例如，在同义词替换过程中，可能会因为同义词词典并不健全或者词向量模型的局限性，导致替换后的藏语句子在句意上存在一定的偏差，从而导致数据质量较低。

2. 缺乏针对性

经过笔者实践调查得知，当前所运用的数据增强方法大部分

都是通用的，藏汉语言之间的差异性和独特性并未充分考虑，这导致基于数据增强技术的藏汉机器翻译缺乏针对性，从而影响翻译质量。例如，藏语与汉语之间存在着显著的差异，在句式结构、词汇、文化背景等多个方面存在着明显的不同。^[7]而运用的数据增强技术针对这些差异无法有效进行处理，从而导致藏汉机器翻译的效果有限。

3. 计算资源消耗较大

在具体实践中，一些数据增强方法，如回译、随机删除与替换等，需要进行多次翻译操作，这会消耗大量的计算资源。尤其是在处理规模庞大的数据集时，需要耗费的计算资源更为庞大。^[8]因此，如何降低计算资源消耗，已经成为当前基于数据增强技术的藏汉机器翻译研究的重要方向之一。

三、基于数据增强技术的藏汉机器翻译的具体应用

（一）在模型训练中的应用

1. 拓展训练数据

可以利用数据增强技术，生成大量伪平行语料，从而有效拓展藏汉平行语料库。这样做能够为模型训练提供大量且多元的数据，能够使其学习到更多语言特征。例如，通过使用回译、同义词替换等方法，能够生成大量的平行句子，并将其添加到藏汉平行语料库中，为模型训练提供大量数据，从而有效提升模型的泛化能力。

2. 提升模型鲁棒性

数据增强技术在模型训练中应用能够有效提升模型鲁棒性。鲁棒性主要是指系统、模型在面对噪声、干扰或者环境变化时，依旧具备可靠性、稳定性的能力。在模型训练过程中，常常会因为数据分布不均或数据不足而出现过拟合、欠拟合等问题。^[9]然而，通过运用数据增强技术，能够对原始数据进行处理，生成更多数量的多样化数据，能够有效避免上述问题的产生。通过全面运用数据增强技术，模型能够学习多种不同的语言特征，从而能够针对不同的输入给出更加准确、可靠的翻译结果。

总之，将数据增强技术运用在模型训练之中具有重要的价值，能够显著提升模型的泛化能力和鲁棒性，从而提升藏汉机器翻译质量。

（二）在特定领域中应用

数据增强技术还能够运用在多个特定领域之中，并且发挥着重要的作用。

1. 医学领域

藏医藏药具有悠久发展历史，是藏族文化的重要组成部分。然而，由于藏语与汉语之间存在着较大的差异，导致藏医藏药的应用和推广受到了一定的限制。^[10]而将数据技术应用于藏汉机器翻译之中，能够生成大量伪平行语料，能够为藏医藏药提供更为准确的翻译服务，这不仅能够将藏医藏药文化进行大范围推广，同时还能够使更多病患群体享受到更多医疗资源。^[11]同时，数据增强技术还能够对医学领域中的专业术语、名词等进行处理，提升藏语翻译的准确性和专业性。

2. 法律领域

法律语言具有专业性、规范性以及严谨性特点，对翻译的准确性要求较高，任何一处细小的翻译错误都可能会导致法律条文存在歧义或误导，从而影响法律的公平性。^[12]然而，藏文法律文献资源贫瘠，建设平行语料库面临较大的困难，这也在一定程度上增加了藏汉法律机器翻译的难度。而应用数据增强技术，能够针对法律领域中的专属词汇、特殊用语、句式结构等进行特殊处理，从而提升翻译的准确性和专业性。

（三）与其他技术结合应用

1. 与预训练模型相结合

将数据增强技术与预训练模型相结合，能够有效提升藏汉机器翻译的性能。具体来讲，在预训练阶段，可以利用数据增强技术生成大量的伪平行语料，这些语料能够为预训练模型提供类型丰富、多样的学习材料。^[13]之后，对预训练模型进行科学调整和优化，使其能够更好地捕捉藏汉两种语言之间的特征。这种处理方式能够有效地提升藏汉机器翻译的准确性和流畅性。

2. 与主动学习相结合

主动学习是一种机器学习方法，主要通过选择性地标注最有

信息量的数据来训练模型，从而有效提升模型的性能。将数据增强技术与主动学习相结合，能够利用该技术的强大功能，生成大量伪平行语料，之后运用主动学习方法，从大量的伪平行语料中选择出最具信息量的数据进行标注。^[14]这样做，能够在有限的标注资源下，提升藏汉机器翻译性能。

四、结束语

总之，数据增强技术在藏汉机器翻译中应用具有重要的价值，能够有效解决平行语料匮乏问题，进一步提升翻译模型的鲁棒性和泛化能力。^[15]然而，当前，基于数据增强技术的藏汉机器翻译依旧存在一系列问题，如，数据质量较低、缺乏针对性以及计算资源消耗巨大等。未来，还需要对数据增强技术进行全面、深入的研究，并针对藏汉语言之间的差异性和独特性，开发出更具针对性的、实效性的数据增强方法，从而有效提升翻译质量，进一步推动藏汉机器翻译技术发展，为推动藏汉文化交流沟通、藏区经济发展作出应有贡献。

参考文献

- [1] 格桑加措, 尼玛扎西, 群诺, 等. 基于长尾词分布的藏汉机器翻译数据增强方法研究 [J/OL]. 计算机科学, 1-10[2025-04-22].
- [2] 格桑加措, 尼玛扎西, 嘎玛扎西, 等. 基于领域术语词典和句式框架的藏汉机器翻译领域数据增强方法研究 [J]. 高原科学研究, 2024, 8(03): 92-101+132.
- [3] 仁青卓玛, 拥措, 唐超超. 面向藏汉神经机器翻译的数据筛选方法 [J]. 计算机与现代化, 2024, (06): 19-24.
- [4] 刘文静. 面向中医文化传播的中医古籍机器翻译研究 [D]. 南京中医药大学, 2024.
- [5] 仁青卓玛. 藏汉神经机器翻译研究综述 [J]. 西藏科技, 2024, 46(02): 76-80.
- [6] 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.
- [7] 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]. 青海师范大学, 2023.
- [8] 汪超. 基于数据增强技术的藏汉机器翻译方法研究 [D]. 西藏大学, 2023.
- [9] 申彰利, 周毛克, 赵小兵. 基于多任务学习的民汉神经机器翻译数据增强方法 [J]. 中文信息学报, 2023, 37(02): 97-106.
- [10] 杨丹, 孙义栋, 拥措. 基于数据增强的藏汉神经机器翻译研究 [J]. 计算机与数字工程, 2022, 50(11): 2473-2477.
- [11] 鲍薇, 董建, 徐洋, 等. 藏汉跨语言文本删窃检测数据集 [J]. 中国科学数据 (中英文网络版), 2022, 7(02): 50-58.
- [12] 严威. 中小学教育资源汉藏翻译系统的设计与实现 [D]. 西北民族大学, 2022.
- [13] 色差甲. 藏文律诗自动生成研究 [D]. 青海师范大学, 2022.
- [14] 吴春壮. 基于数据增强的低资源神经机器翻译研究 [D]. 北京交通大学, 2022.
- [15] 蔡子龙, 杨明明, 熊德意. 基于数据增强技术的神经机器翻译 [J]. 中文信息学报, 2018, 32(07): 30-36.