文章编号：1003-0077(2019)12-0061-06

融合单语语言模型的藏汉机器翻译方法研究

慈祯嘉措[1,2]，桑杰端珠[1,2]，孙茂松[3]，色差甲[1,2]，周毛先[1,2]

(1. 青海师范大学 藏文信息处理教育部重点实验室，青海 西宁 810008；
2. 青海省藏文信息处理与机器翻译重点实验室，青海 西宁 810008；
3. 清华大学 计算机系，北京 100084)

摘 要：由于藏汉平行语料匮乏，导致藏汉神经网络机器翻译效果欠佳，该文提出了一种将藏语单语语言模型融合到藏汉神经网络机器翻译的方法，首先利用神经网络实现藏语单语语言模型，然后使用 Transformer 实现藏汉神经网络机器翻译模型，最后将藏语单语语言模型融合到藏汉神经网络机器翻译中。实验表明，该方法能显著提升藏汉神经网络机器翻译质量。基线系统藏语到汉语的 BLEU 值为 21.1，汉语到藏语的 BLEU 值为 18.6，融合藏语单语语言模型后，藏语到汉语的 BLEU 值为 24.5，汉语到藏语的 BLEU 值为 23.3，比原有基线系统的 BLEU 值分别提高了 3.4 和 4.7。

关键词：藏语；语言模型；机器翻译；融合；神经网络

中图分类号：TP391　　**文献标识码**：A

Tibetan-Chinese Machine Translation Based on Tibetan Language Model Enhanced Transformer

CIZHEN Jiacuo[1,2], SANGJIE Duanzhu[1,2], SUN Maosong[3], SE Chajia[1,2], ZHOU Maoxian[1,2]

(1. MOE Key Laboratory of Tibetan Information Processing, Qinghai Normal University, Xining, Qinghai 810008, China;
2. Provincial Key Laboratory of Tibetan Intelligent Information Processing and Machine Translation, Xining, Qinghai 810008, China;
3. Department of Computer Science, Tsinghua University, Beijing 100084, China)

Abstract: To better utilize the monolingual Tibetan texts in Tibetan-Chinese neural machine translation(NMT), we propose to pre-train a Tibetan neural language model and then integrate it into a Transformer-based Tibetan-Chinese NMT model. Experiments indicate our approach can boost the Tibetan-Chinese results from 21.1 to 24.5, and the Chinese-Tibetan form 18.6 to 23.3 in terms of BLEU score.

Keywords: Tibetan; language model; machine translation; fusion; neural net

0 引言

早期的语言模型和机器翻译方法受限于人工构建的规则，由于语言的复杂性和多样性，基于规则的方法需要构建规模庞大的规则库才能刻画语言的特性，但规则库的维护和复杂性又依赖于人类专家的经验和知识，无法对语言现象进行完备的描述。为解决规则机器翻译的缺陷和不足，基于统计的机器翻译研究开始涌现，其方法是通过大规模的标注语料学习语言的基本特性，由于统计机器翻译需要大规模的标注数据会消耗大量的人力物力。目前，基于神经网络的机器翻译能够较好地解决规则和统计方法存在的问题。与传统方法相比，以 Transformer 为代表的神经网络方法对数据更加依赖，因为其巨大的网络参数空间需要用大规模数据进行参数估计，从而导致翻译性能并不理想。

为了解决低资源下机器翻译中存在的问题，2016 年 Zoph Barret 等提出了一种迁移学习方法，其主要思想是先训练一个完备的机器翻译系统模

收稿日期：2019-06-18　**定稿日期**：2019-08-01

基金项目：国家自然科学基金(61063033,61662061)；国家重点研发计划(2017YFB1402200)

型,然后将这个模型的参数传递给低资源的机器翻译模型,从而达到低资源模型参数的初始化和约束训练,这样可以显著提高低资源条件下机器翻译的性能[1]。2017 年 Robert Ostling 等利用向量间的依赖关系和单词对齐来解决翻译中的排序问题,并且证明了 NMT 也可用于低资源场景[2]。2018 年 Ebtesam H Almansor 等提出了递归神经网络和卷积神经网络相融合的机器翻译模型,用来解决低资源下阿拉伯语到英语的机器翻译问题[3]。2018 年 Tao Feng 等为了解决低资源下机器翻译的性能问题,提出了两种解决方法,第一种方法采用解码器权重共享来增强低资源 NMT 系统的目标语言模型,第二种方法应用跨语言嵌入和源语言表示空间共享来加强低资源 NMT 编码器[4]。

1 总体框架和相关理论

基于神经网络构架的藏汉机器翻译研究刚刚起步,特别是对低资源条件下的藏汉神经网络机器翻译相关研究很少。本文首先利用 Transformer 作为基线系统搭建藏汉神经网络机器翻译系统,然后在编码器中将源语言置空,也就是说编码器只训练单语的语言模型,然后利用现有资源对解码器中两个语言(藏汉)之间的对应关系进行训练,通过加入不同规模的语料,对比和分析其实验结果,期望得到一个低资源条件下高效的藏汉神经网络机器翻译系统。

1.1 总体框架

以 Transformer 为主体框架,首先在编码器端训练藏语单语语言模型,将其作为输入;然后将藏语单语语言信息与编码器端的预输出进行加权处理,使源语言与目标语言产生映射关系,最终输出目标语言,如图 1 所示。

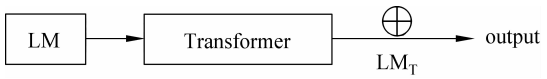

图 1 总体框架

图 1 中,LM 表示编码器端训练的藏语单语语言模型,⊕表示归一化处理,LM_T 表示在解码器端加入的藏语语言信息。

1.2 Transformer 框架

2014 年 Ilya Sutskever 等为了解决神经网络对序列任务不适用的问题,提出了一种端到端的神经网络机器翻译构架[5]。这种构架用一个多层的 LSTM 网络将输入序列映射(编码)为一个固定大小维度的向量,再用另外一个多层的 LSTM 网络来解码该向量作为输出序列[5]。同年,Bahdanau D 等使用固定长度向量提高编码器—解码器架构性能,并且为了打破这种架构的瓶颈,使用词表的自动对齐来扩展模型的性能[6]。直到 2017 年,Google 的 Ashish Vaswani 等提出了一种基于自注意力机制(self-attention)的模型构架,这种构架可以建模各种自然语言处理问题,并在多项任务中取得了最好成绩。相较于利用 RNN 或者 CNN 作为编码器—解码器(encoder-decoder)的传统的神经机器翻译,谷歌提出的基于 attention 的 Transformer 模型抛弃了传统的构架,并没有用任何 CNN 或者 RNN 的结构。该模型可以完全地进行并行运算,在提升翻译性能的同时训练速度非常快。Transformer 模型构架如图 2 所示。

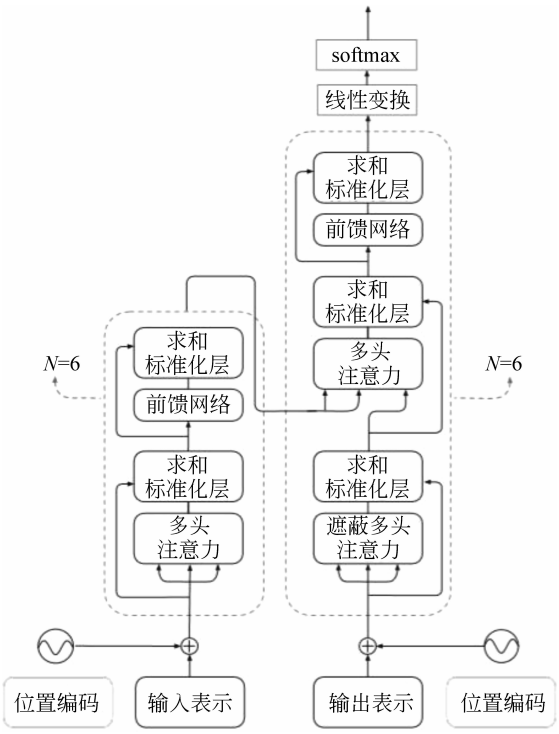

图 2 Transformer 模型框架

从图 2 可以看出,编码器由多个相同的层堆叠在一起,每一层又有两个支层,第一个支层是一个多头的自注意力机制,第二个支层是一个简单的全连

接前馈网络，解码器和编码器的结构相似，但多了一个多头注意力机制，如式(1)所示。

$$\text{sub\_layer\_output} = \text{LayerNorm}(x + (\text{SubLayer}(x))) \quad (1)$$

因为在编码器和解码器中都没有递归和卷积运算，Transformer 无法自然地利用序列中的位置信息，但是对于机器翻译任务，序列中的各个元素的位置是非常重要的。为解决这个问题，Transformer 使用了一个称为位置编码(positional encoding)的方法将每个元素的位置信息显式地嵌入到网络中，即编码器的输入为位置编码向量加上(⊕)输入序列的嵌入式表示。位置编码的学习通过三角函数完成，如式(2)所示。

$$p_{(\text{pos},2i)} = \sin(\text{pos}/10000^{2i/d_{\text{model}}}) \quad (2)$$

其中，pos 代表序列中元素的位置，$2i$ 表示位置向量的维度；d_{model} 表示自注意力网络的输出维度。因为三角函数具有周期性，对于固定长度的偏差 k，$P(\text{pos}+k)$可以表示为 P)的线性函数，使模型能够很容易地学习序列中各个元素的相对位置关系信息[6-7]。

图 3 与式(3)表示 Transformer 模型中矩阵的相关计算，$\boldsymbol{Q}$ 表示查询矩阵，$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 表示键值对矩阵；$\sqrt{d_k}$ 为 softmax 的缩放系数。

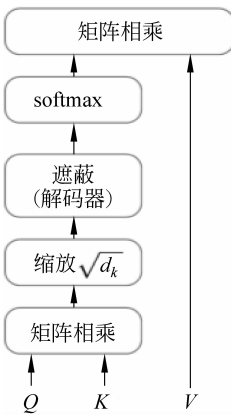

图 3　缩放点积注意力的计算示意图

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)V \quad (3)$$

考虑到在 d_k 较大时，q_i 和 k_i 的积会将 softmax 函数推向梯度极小的区域，从而影响有效的反向传播，作为 softmax 输入的点积被缩小 $\sqrt{d_k}$ 倍。在机器翻译中，Transformer 将编码器的隐状态视为一组键(*Key*) 值(*V*alue) 对的集合，而在解码器中 t 时刻之前生成的输出序列被压缩为查询 (Query)矩阵，当前 t 时刻解码器的输出通过查询与键值集合的映射完成。多头注意力对 $\boldsymbol{Q}$、$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 进行 h 次不同的投射，每次投射的维度都是 d_k 和 d_v，然后经过缩放点积运算，再将 h 次计算的结果通过线性映射，获得最终的多头注意力网络输出[7]，如图 4 所示。

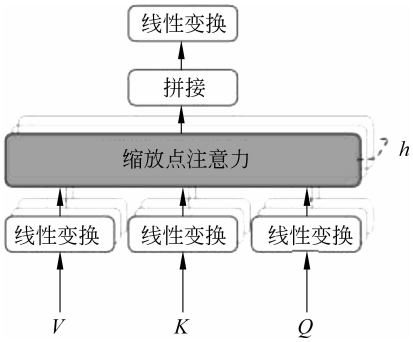

图 4　多头注意力网络示意图

1.3　单语语言模型融合策略

本文使用 Transformer 构架来对系统进行实现，在神经网络机器翻译中，总共有三个参数影响其翻译性能，如式(4)所示。

$$P(y_i \mid x, y_{<i}, \theta) = \sigma(y_{i-1}, C_i, S_i) \quad (4)$$

式(4)中，y_i 表示 i 时刻生成的目标语言，x 表示源语言的输入，$y_{<i}$ 表示 i 时刻之间已生成的所有目标语言，θ 表示模型的参数，C_i 表示 i 时刻上下文的信息，S_i 表示隐藏层 i 时刻的状态。在机器翻译过程中，翻译质量由以下两个因素决定：一是编码器所训练的语言模型；二是解码器所学习到的对应关系，这种对应关系需要大规模的语料作为支撑才能完全学到源语言与目标语言之间的语言关系，而在藏汉(汉藏)机器翻译中，平行语源较少，无法完全学习到藏语—汉语之间的映射关系，那么只能提高编码器所训练的语言模型的质量来提高翻译性能。在神经网络机器翻译中[8]：

$$p(x \mid y) = p(\text{LM}) \cdot p(\text{decoder}) \quad (5)$$

如式(5)所示，翻译的性能是由编码器和解码器共同决定的(乘积的关系)，在低资源的藏汉神经网络机器翻译中，解码器的性能无法再次得到提升(因为需要大规模的平行语料)，那么只能通过提高编码器的性能来提升机器翻译的性能，而在神经网络模型架构中，整个训练过程是一个完整体，很难被打断或者是分割，嵌入语言模型的难度也很大，在编码端把源语言置空，只训练单语的语言模型，从而达到与嵌入单语语言模型相同的效果。

本文将藏语单语训练的语言模型作为编码器一端，本质上是删除编码端上下文向量 $\boldsymbol{C}_i$ 的信息，神

经网络必须完全依赖于前一个网络的输出来预测下一个网络的输出，这就相当于上下文信息被删除。本文将这种设置看作是多任务学习，当源语言已知时，这个任务就是藏汉机器翻译，当源语言未知时，神经网络进行藏语单语语言建模。在训练过程中，本文以 1∶1 的比例使用对齐语料和藏语单语语料进行训练，并随机打乱。在解码器端，本文将上一时刻 y_{t-1} 作为当前时刻的输入。同时，藏语单语语言模型也在影响整个网络的输出，训练的翻译模型生成的词和语言模型所生成的词重新加权排序，得到一个最优的输出，如图 5 所示。

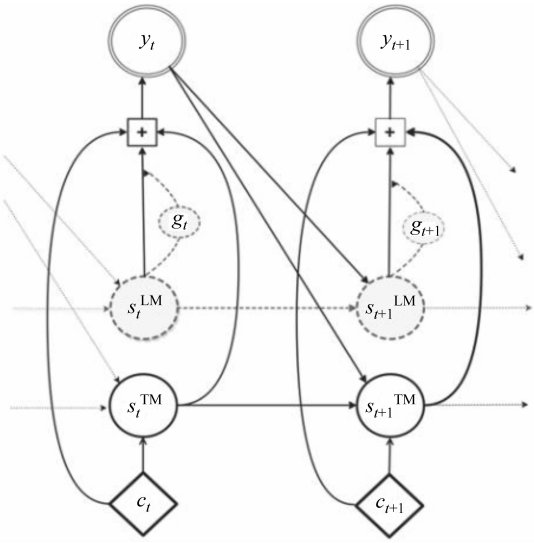

图 5 融合单语语言模型信息

在每步预测每个词之前，将神经网络的解码器的隐藏状态 s_t^{TM} 与藏语单语语言模型 s_t^{LM} 进行合并，控制器 g_t 用于重新计算语言模型的权重。如式(6)所示。

$$p(y_t \mid y_{<t}, x) \propto \exp(y_t^T (W_o f_o(s_t^{LM}, s_t^{TM}, y_{t-1}, c_t) + b_o)) \quad (6)$$

本文首先将藏语单语语言模型与神经网络模型的解码器进行融合，使隐藏状态串联起来(图 5)。然后，在计算下一个单词的输出概率时，对模型进行微调，使用这两个模型的隐藏状态(式(6))。与一般的神经网络机器翻译模型不同，每个网络输出的隐藏层除了神经网络本来拥有的解码器、前一时刻的单词之外，还将藏语单语语言模型的隐藏状态作为输入。其中，本文使用 s_t^{TM} 和 s_t^{LM} 分别表示神经网络解码端和单语语言模型的隐藏状态。在训练过程中，只更新用于参数化输出的参数，以确保藏语单语语言模型所学到的特性不会被覆盖[9]。

2 实验分析

2.1 数据的来源(准备)

本文总共收集 400 万句藏语单语语料，其中单语语料中 310 万为新闻领域的语料，40 万为法律领域语料，50 万为其他领域的语料；收集 160 万句对为藏汉双语平行语料，其中 90 万为新闻领域语料，40 万为法律领域语料，30 万为其他语料。语料的整体结构如表 1 所示。

表 1 语料领域分布表

语料类别	单语语料	双语语料
新闻语料/万	310	90
法律语料/万	40	40
其他/万	50	30
共计/万	400	160

2.2 实验

深层融合方法(deep fusion)见式(6)与图 5，在训练过程中，只更新用于参数化输出的参数，以确保藏语单语语言模型所学到的语言特性不会被覆盖。在融合过程中，本文将权值和标准差进行了设置，在训练速率上，每 10K 训练数据对模型进行一次模型 BLEU 值的计算，直到模型性能不再提升为止。本文 Transformer 的参数设置如表 2 所示。

表 2 模型参数设定

参数名称	参数选择
标签平滑率	0.1
优化器	Lazy Adam Optimizer
学习率	2.0
学习率衰减类型	noam
学习率热启动轮数	10 000
长度惩罚率	0.6
批处理类型	token
批处理大小	4 200
编码器输入的维度	512
自注意力层数	4
自注意力层的隐层单元个数	512

续表

参数名称	参数选择
自注意力层多头注意力网络头数	8
前馈网络的隐层单元个数	2 048
前馈网络的 dropout	0.1
自注意力网络的 dropout	0.1
ReLu 层的 dropout	0.2

各个模型的 BLEU 值如表 3 所示：

表 3　各个模型 BLEU 值

模型	藏—>汉	汉—>藏
Transformer	21.1	18.6
Transformer+deep fusion(融合)	24.5	23.3

2.3　分析

实验结果显示，基线系统藏汉机器翻译的 BLEU 值为 21.1，汉藏机器翻译的 BLEU 值为 18.6，而融合藏语单语语言模型的机器翻译系统，藏汉机器翻译的 BLEU 值为 24.5，汉藏机器翻译的 BULE 值为 23.3，比原有的基线系统 BLEU 值分别提高了 3.4 和 4.7 。BLEU 实验结果表明，基于单语语言模型融合的藏汉(汉藏)神经网络机器翻译系统比原有的基线系统更加有效。

3　总结与展望

本文以目前效率最高的 Transformer 为基线系统，对藏汉(汉藏)神经网络机器翻译系统进行了实现，首先对单语语言模型融合的机器翻译系统进行了实现，将藏语单语训练的语言模型作为编码器一端，以 1∶1 的比例使用对齐语料和藏语单语语料进行训练，并把藏语单语语言模型与神经网络模型的解码器进行融合，将它们的隐藏状态串联起来，再计算下一个输出的概率，对模型进行微调，实现了一个融合单语语言模型的藏汉(汉藏)机器翻译系统，最终的实验结果表明，单语语言模型融合策略可以有效地提高原有藏汉(汉藏)神经网络机器翻译系统的性能。

参考文献

[1] Zoph B, Deniz Y, Jonathan M, et al. Transfer learning for low-resource neural machine translation[C]// CoRR abs/1604.02201. 2016.

[2] Robert Ostling, Jorg Tiedemann. Neural machine translation for low-resource languages[C]//Proceedings of the EMNLP 2017.

[3] Ebtesam H. Almansor, Ahmed Al-Ani. A hybrid neural machine translation technique for translating low resource languages[C]//Proceedings of the 14th International Conference, MLDM, 2018.

[4] Tao Feng, Miao Li, Xiaojun Liu, et al. Improving low-resource neural machine translation with weight sharing[C]//Proceedings of the CCL, 2018.

[5] Ilya Sutskever, Oriol Vinyals, Quoc V Le. Sequence to sequence learning with neural networks[C]//Proceedings of the NIPS, 2014.

[6] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv:1409.0473. 2014.

[7] Ba J L, Kiros J R, Hinton G E, Layer normalization [J]. arXiv:1607.06450. 2016.

[8] Gu J, Hassan H, Devlin J, et al. Universal neural machine translation for extremely low resource languages [C]//Proceedings of the NAACL-HLT, 2018.

[9] Yoshua Bengio. On integrating a language model into neural machine translation[J]. Science Direct, 2016, 15(1):137-148.

慈祯嘉措(1989—)，博士研究生，主要研究领域为计算语言学、藏文信息处理、机器翻译。
E-mail：543819011@qq.com

桑杰端珠(1986—)，博士研究生，主要研究领域为计算语言学、藏文信息处理、机器翻译。
E-mail：sangjeedondrub@live.com

孙茂松(1962—),通信作者,教授,博士生导师,
主要研究领域为自然语言理解、中文信息处理、
机器翻译等。
E-mail:sms@mail.tsinghua.edu.cn

CCKS 2019 全国知识图谱与语义计算大会在杭州隆重召开

2019 年全国知识图谱与语义计算大会(CCKS 2019)于 8 月 24 日至 27 日在杭州召开,由中国中文信息学会语言与知识计算专业委员会主办,浙江大学承办。本次会议主题是“知识智能”。大会吸引了来自海内外的 800 多名科研学者、工业界专家和知名企业代表参加。会议回顾了知识图谱与语义计算的进展情况,探讨了领域内的新发现、新技术和新应用,旨在让社会各界了解知识图谱与与语义计算的新方向和新趋势,以推动我国语言与知识计算领域的进一步发展。

CCKS 2019 会议分为学科前沿讲习班和大会主会两个阶段。8 月 24 日至 25 日,中国中文信息学会《前沿技术讲习班》(ATT)第十六期在杭州宝盛水博园大酒店举行。前沿技术讲习班邀请了国内外优秀青年学者及工业界专家,内容涵盖了知识图谱的推理、构建,自然语言的推理、关系抽取及知识图谱应用等方面,分别从知识图谱的构建及在实际场景中的应用等角度介绍了知识图谱的最新进展和实战经验。

8 月 26 日,CCKS 会议主会开幕式也在杭州宝盛水博园大酒店举行,中国中文信息学会理事长方滨兴院士致欢迎辞,语言与知识计算专业委员会主任、清华大学计算机科学与技术系李涓子教授介绍了语言与知识计算专委会以及 CCKS 大会历史,大会主席清华大学朱小燕教授和大会程序委员会主席哈尔滨工业大学秦兵教授分别介绍了大会的组织情况。浙江大学陈华钧教授主持了开幕式。

主会包括特邀报告、优秀学术论文报告、知识图谱相关顶级会议回顾、知识图谱评测与竞赛及知识图谱工业界论坛等环节。特邀报告环节邀请了海内外知名学者和工业界代表介绍了学科前沿信息及重要成果,英国南安普顿大学(University of Southampton)的 Dame Wendy Hall 教授作了题为“Web Science, AI and Future of the Internet”的特邀报告;美国伊利诺伊大学香槟分校(University of Illinois at Urbana-Champaign)的 Heng Ji 教授作了题为“ PaperRobot: Automated Scientific Knowledge Graph”的特邀报告;加拿大滑铁卢大学(University of Waterloo)的李明教授作了题为“第三代聊天机器人”的特邀报告,介绍了第三代聊天机器人架构和可行的实现方法;百度 CTO 王海峰博士作了题为“知识图谱与语义理解”的特邀报告,介绍了百度知识图谱与语义理解技术及应用,并探讨了未来发展方向。

CCKS 是中国中文信息学会语言与知识计算专委会定期举办的全国年度学术会议,致力于促进我国语言与知识计算领域的学术研究和产业发展,为从事相关领域理论和应用研究的学者、机构和企业提供广泛交流的平台。CCKS 2019 聚集了知识表示及获取、知识推理、自然语言理解、智能问答等相关技术领域的重要学者和研究人员,为所有与会者带来了一场学术与技术的饕餮盛宴。