基于模型不确定性约束的半监督汉缅神经机器翻译

王琳钦[1,2]，余正涛[1,2]，毛存礼[*1,2]，高盛祥[1,2]，满志博[1,2]，王振晗[1,2]
1.昆明理工大学，信息工程与自动化学院，昆明，650500
2.昆明理工大学，云南省人工智能重点实验室，昆明，650500
2424172505@qq.com, ztyu@hotmail.com, maocunli@163.com
gaoshengxiang.yn@foxmail.com, 270004294@qq.com, 243836042@qq.com

摘要

基于回译的半监督神经机器翻译方法在低资源神经机器翻译取得了明显的效果，然而，由于汉缅双语资源稀缺、结构差异较大，传统基于Transformer的回译方法中编码端的Self-attention机制不能有效区别回译中产生的伪平行数据的噪声对句子编码的影响，致使译文出现漏译，多译，错译等问题。为此，该文提出基于模型不确定性为约束的半监督汉缅神经机器翻译方法，在Transformer网络中利用基于变分推断的蒙特卡洛Dropout构建模型不确定性注意力机制，获取到能够区分噪声数据的句子向量表征，在此基础上与Self-attention机制得到的句子编码向量进行融合，以此得到句子有效编码表征。实验证明，本文方法相比传统基于Transformer的回译方法在汉语-缅甸语和缅甸语-汉语两个翻译方向BLEU值分别提升了4.01和1.88个点，充分验证了该方法在汉缅神经翻译任务的有效性。

关键词： 回译 ； 模型不确定性 ； 模型不确定性注意力机制 ； 半监督神经机器翻译 ； 汉缅神经机器翻译

Semi-Supervised Chinese-Myanmar Neural Machine Translation based Model-Uncertainty

Linqin Wang[1,2], Zhengtao Yu[1,2], Cunli Mao[*1,2], Shengxiang Gao[1,2], Zhibo Man[1,2], Zhenhan Wang[1,2]
1. Faculty of Information Engineering and Automation, Kunming University of Science and Technology Kunming 650500, China
2. Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology Kunming 650500, China
2424172505@qq.com, ztyu@hotmail.com, maocunli@163.com
gaoshengxiang.yn@foxmail.com, 270004294@qq.com, 243836042@qq.com

Abstract

The semi-supervised neural machine translation based on back-translation has achieved remarkable results in low-resource neural machine translation. However, due to the scarcity of Chinese-Myanmar bilingual corpus and large structural differences, the self-attention mechanism at the encoder of traditional Transformer based back-translation cannot effectively distinguish the influence of the noise of pseudo parallel data in translation on sentence encoding, resulting in problems such as omission, multiple translation and mistranslation of the translated text. Therefore, this article proposes a semi-supervised Chinese-Myanmar neural machine translation based model uncertainty. In Transformer network, Monte Carlo dropout based on Variational Inference is used to construct model uncertainty attention mechanism to obtain sentence vector

*毛存礼(通信作者):maocunli@163.com
国家自然科学基金重点项目(61732005)；国家自然科学基金(61866019，61761026，61972186)；云南省应用基础研究计划重点项目(2019FA023)；云南省中青年学术和技术带头人后备人才项目(2019HB006)

representations that can distinguish noise data. On this basis, it is fused with the sentence encoding vector obtained by the self-attention mechanism, so as to obtain the effective sentence encoding representation. Experimental results show that compared with traditional Transformer based back translation , the BLEU score of the proposed method improves by 4.01 and 1.88 points respectively in Chinese-Myanmar and Myanmar-Chinese translation directions, which fully verify the effectiveness of the proposed method in Chinese-Burmese neural machine translation tasks.

Keywords: back-translation , model-uncertainty attention , Semi-Supervised Neural Machine Translation , Chinese-Myanmar Neural Machine Translation

1 引言

基于回译的方法在结构差异性较小的低资源语言对，如英法，英德，取得了较为显著的效果。然而，对于汉缅机器翻译，由于语料资源稀缺而且结构差异较大，回译生成的伪语料存在漏译，多译，错译的问题，传统基于Transformer编码端的Self-attention机制不能有效区别回译中产生的伪平行数据的噪声对句子编码的影响，致使回译方法应用于结构差异较大的低资源的语对上效果欠佳。如图1所示，在汉语→缅甸语翻译方向上，my是目标语言缅甸语的真实单语语料，zh-base是传统回译方法用真实平行双语语料训练得到的缅甸语→汉语翻译模型翻译生成的人造伪语料。传统基于Transformer的回译方法将缅甸语单语my回译成伪语料zh-base“天@@ 花@@ 板@@ 和@@ 白@@ 色，白@@ 色。”，(@@ 是BPE分词以后的结果）），与词对齐的数字则表示通过蒙特卡洛Dropout建模的回译过程中词级别的模型不确定性大小，模型不确定性值越小，则表示该词在模型训练时越可信 (Wang et al., 2019)。“白色”这个词在这个句子中重复出现了两次，与图1上半部分翻译正确的语料中名词（如“他”、“她”、“比赛”）相比，观察到在译文中重复的名词拥有较大的模型不确定性值（图1中红色标识的“0.150885”），同时对比观察到最后一个句号的模型不确定性值很高（“0.116445”），表明zh-base在模型预测时有很大概率存在漏译，错译，多译的问题。

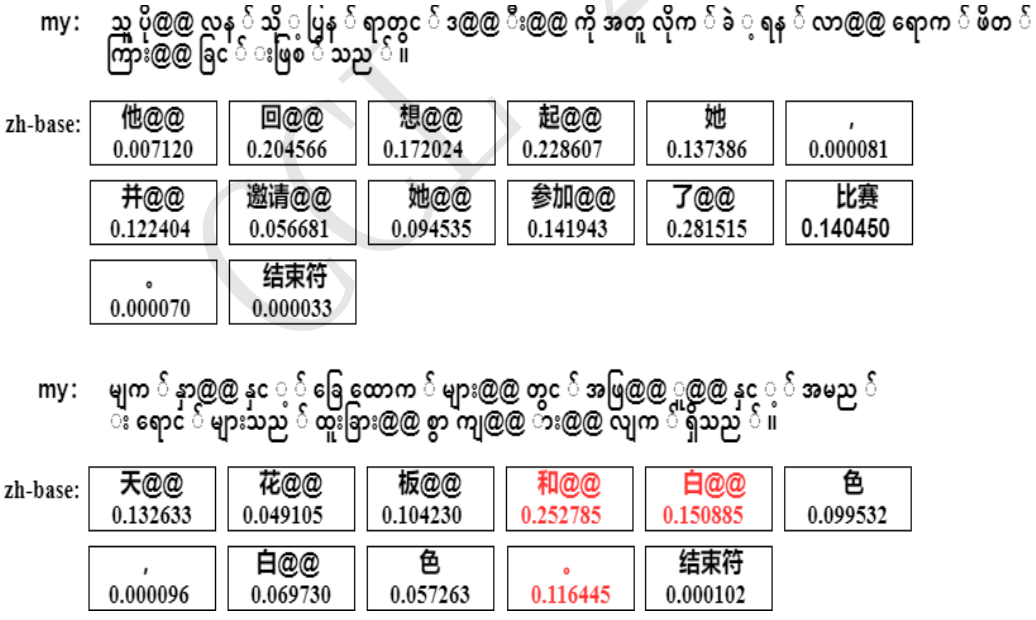

Figure 1: 汉语→缅甸语翻译任务中回译过程及model-uncertainty示例

my为缅甸语原文，zh-base是my经传统回译方法训练的翻译模型得到的汉语译文。本文在传统transformer架构体系下结合模型不确定性建模思想 (Wang et al., 2019),提出一种新的算法，基于模型不确定性约束的半监督汉缅神经机器翻译算法，探讨了将模型不确定性融进Transformer编码端每一层表征，提出模型不确定性注意力机制，用自注意力机制将模型不确

定性与词嵌入层Embedding进行深度融合，在汉缅神经机器翻译的任务上实验证明模型不确定性约束的半监督汉缅神经机器翻译算法能有效提高翻译质量，提升回译性能，在16万的汉缅双语数据集上汉语-缅甸语翻译任务上我们达到了24.72的BLEU值，缅甸语-汉语翻译任务上我们达到15.95的BLEU值。本文的贡献如下：

（1）提出了模型不确定性约束的半监督汉缅神经机器翻译算法，提出以实验为支撑的融合机制策略，实现了将模型不确定性融进Transformer编码端每一层表征，解决了回译方法应用在汉缅神经机器翻译任务中性能较差的问题。

（2）提出模型不确定性注意力机制，用自注意力机制将模型不确定性与Embedding进行深度融合，使编码端能更好的得到伪语料的句子向量表征。

（3）在16万的汉缅双语数据集上汉语-缅甸语翻译任务我们达到24.72的BLEU值，缅甸语-汉语翻译任务我们达到15.95的BLEU值。

第2节概述了缅甸语和低资源神经机器翻译的相关工作，第3节描述了模型不确定性约束的半监督汉缅神经机器翻译算法的算法思想；第4节通过在传统迁移学习，传统回译方法以及Wang等人 (2019)方法下进行实验对比证明本文方法的优势；第5节对全文进行总结并指出进一步的研究工作。

2 相关工作

本文将相关工作的分为两类，分别是缅甸语的神经机器翻译以及低资源语言的神经机器翻译方法研究。针对于缅甸语的机器翻译的研究工作：目前，由于缅甸语的双语资源较少，针对于缅甸语的机器翻译的研究工作较少，Nwet等人 (2011)提出一种通过缅甸语-英语词对齐的英缅统计机器翻译方法，由于这种方法在一定程度上受到词表大小的限制，Nwet等人 (2011)进一步提出通过扩展英缅双语的平行语料的机器翻译方法。以上针对于缅甸语的机器翻译研究都是基于统计的方式，基于统计的方式需要大规模的双语词典或者是双语平行语料，缅甸语是一种典型的资源稀缺型语言，利用统计的方式不能完全适用于缅甸语。缅甸语是一种资源稀缺型语言，解决低资源语言的神经机器翻译的方法主要包括：(1)基于迁移学习的神经机器翻译方法 (Lakew et al., 2018; Sachan and Neubig, 2018; Dabre et al., 2019; Firat et al., 2016)：借助预训练的思想，将资源丰富语言对训练模型或参数迁移到低资源语言对的训练过程中。缅甸语和其他语言之间的语法差异性极大，直接利用迁移学习的思想将模型迁移到缅甸语上的效果性能不佳。(2)半监督神经机器翻译中单语语料的应用是改善低资源神经机器翻译的有效方法 (Sennrich et al., 2015a; Gulcehre et al., 2015; He et al., 2016),近些年来，回译（Back-translation）在半监督机器翻译利用单语语料方法中逐渐成为主流 (Sennrich et al., 2015a)，回译的基本思想是用有限的真实语料语料库训练得出神经机器翻译模型，用训练得到的神经机器翻译模型将单语语料翻译生成人造的伪平行的双语语料库。将真实的双语平行语料和生成的伪平行混合后参与模型再次训练。因为这种做法的简单性和有效性，回译被广泛应用于低资源神经机器翻译中。然而，因为神经机器翻译模型生成的人造语料无可避免的带有噪声，导致翻译错误会传播到后续步骤，极其容易阻碍翻译性能，特别在零资源（zero-shot）语言上，由于源语言与目标语言的语言差异性较大，加上训练目标语言到源语言方向翻译模型训练平行语料稀缺，模型训练不充分，无可避免的导致训练得出的目标语言到源语言的翻译模型质量差，bleu值低，进一步导致回译得出的源语言端的人造语料存在语义错误，句子不通畅，漏译，多译，错译的问题。

3 方法

本文工作延续Wang等人 (2019)工作中量化模型预测置信度的方法，但探究一种将模型不确定性更深层次融入神经网络翻译模型的算法: 基于模型不确定性为约束的汉缅机器翻译算法，使模型训练时在接受伪语料句子输入时能知道当前句子中的词对齐模型不确定性值，将模型不确定性融进transformer每一层表征，用自注意力机制将词级别模型不确定性与词嵌入层Embedding进行深度融合。3.1部分描述模型不确定性为约束的汉缅机器翻译算法的基本思想，3.2部分探讨了模型不确定性注意力机制与Self-Attention融合机制策略，3.3部分研究了模型不确定性注意力机制原理

3.1 模型不确定性为约束的汉缅神经机器翻译算法

由于贝叶斯深度学习的最新进展，不确定性在量化特定映射对不同输入的置信度方面取得了重大进展，在本文工作中，我们延续 (Wang et al., 2019)工作中量化模型预测不确定性的方法，用贝叶斯神经网络中广泛使用的变分推断近似方法对神经机器翻译模型进行分析，借助被广泛应用的Monte Carlo Dropout (Gal and Ghahramani, 2016)来获得回译过程中翻译概率的样本。通过计算概率样本的期望和方差进一步表征模型不确定性。D_b来表示真实的双语平行语料库，给定一个在目标语言单语语料库D_m中的目标语言单语句子y和y对应的译文$\hat{x}$，贝叶斯神经网络旨在求出回译过程中神经网络模型参数上的后验分布$P(\theta_{y\to\hat{x}}|D_b)$，如（1）所示，回译过程中词级别目标语言到源语言翻译的概率表示为：

$$P(\hat{x}\,|y^{(n)},\hat{x}^{(n)}_{<i},\hat{\theta}_{y\to\hat{x}},D_b)=\int P(\hat{x}\,|y^{(n)},\hat{x}^{(n)}_{<i},\theta_{y\to x}|D_b)d\theta_{y\to x},y^{(n)}\in\{D_m\} \tag{1}$$

图2阐明了模型不确定性为约束的汉缅机器翻译算法的中心思想，图的下半部分借鉴了 (Wang et al., 2019)工作中量化模型预测不确定性的方法，在回译过程中，给定一个真实的目标语言句子y，$y\in\{D_m\}$，通过标准的transformer模型解码预测其对应的$\hat{x}$，为了量化模型在预测时的模型不确定性，首先将词级别的翻译概率视作随机变量。通过随机停用NMT模型的部分神经元（dropout）并重新计算翻译概率（同时保持y和$\hat{x}$固定）来进行翻译概率的采样。当模型预测K次时，为单词翻译概率生成K个样本。用$\hat{\theta}^{(k)}_{y\to x}$来表示在第k次dropout后模型的参数，给定K次采样$\{P(\hat{x}|y,\theta^{(\hat{k})}_{y\to x})\}^K_{k=1}$，词级别的翻译概率的期望可表示为：

$$E[P(\hat{x}\,|y,\hat{x}^{(n)}_{<i},\hat{\theta}_{y\to\hat{x}})]\approx\frac{1}{K}\sum_{k=1}^{K}(\hat{x}\,|y,\hat{x}^{(n)}_{<i},\hat{\theta}^{(n)}_{y\to\hat{x}}) \tag{2}$$

词级别翻译概率的方差可大约为：

$$Var[P(\hat{x}\,|y,\hat{x}^{(n)}_{<i},\hat{\theta}_{y\to\hat{x}})]\approx\frac{1}{K}\sum_{k=1}^{K}(\hat{x}\,|y,\hat{x}^{(n)}_{<i},\hat{\theta}^{(k)}_{y\to\hat{x}})^2-E[P(\hat{x}\,|y,\hat{x}^{(n)}_{<i},\hat{\theta}_{y\to\hat{x}})]^2 \tag{3}$$

结合式(2)和(3)，回译过程中词对齐级别的模型不确定性表征为：

$$uncertainty_{bt}=(1-\frac{Var[P(\hat{x}\,|y,\hat{x}^{(n)}_{<i},\hat{\theta}_{y\to\hat{x}})]}{E[P(\hat{x}\,|y,\hat{x}^{(n)}_{<i},\hat{\theta}_{y\to\hat{x}})]})^{\beta} \tag{4}$$

(4)式中β是模型超参数，默认值为2，$uncertainty_{bt}$取值范围为0到1。为了使解决传统基于Transformer编码端的Self-attention机制不能有效区别回译中产生的伪平行数据的噪声对句子编码的影响，本文方法在回译得到伪平行双语语料库$\hat{D}_b$和伪平行双语语料库对应的$uncertainty_{bt}$后，任何一个输入都要进入结合模型不确定性的编码端。

3.2 编码端模型不确定性注意力机制与Self-Attention融合机制

用H^l_E表示编码端第l层结合词嵌入embedding层$H^0_E=(x^0_1,x^0_2,...,x^0_{l_x})$的隐藏层状态，其中$x^0_i$表示源语言句子x第i个词的embedding单元。标记h^0_i为H^l_E第i个元素，其中$i\in[l_x]$。用attn(q,K,V)示注意力机制，其中q, K, V分布表示query，key，value。这里q是一个d维的向量，K和V设置为$|K|=|V|$。每一个$k_i\in K$和$v_i\in V$同样也是d维的向量，$i\in[|K|]$。注意力机制原理如下：

$$attn(q,K,V)=\sum_{i=1}^{|V|}\alpha_iW_Vv_i,\alpha_i=\frac{exp((w_qq)^T(W_KK_i))}{K},Z=\sum_{i=1}^{|K|}exp((w_qq)^T(W_KK_i)) \tag{5}$$

其中W_q，W_K和W_V是可以学习的参数。在Vaswani等人 (2017)工作中注意力机制attn(q,K,V)用的是多头注意力机制，本文也采取了这种做法，延续 (Vaswani et al., 2017)工作，把非线性变换层定义为：

$$FFN(x)=W_2max(W_1x+b_1,0)+b_2 \tag{6}$$

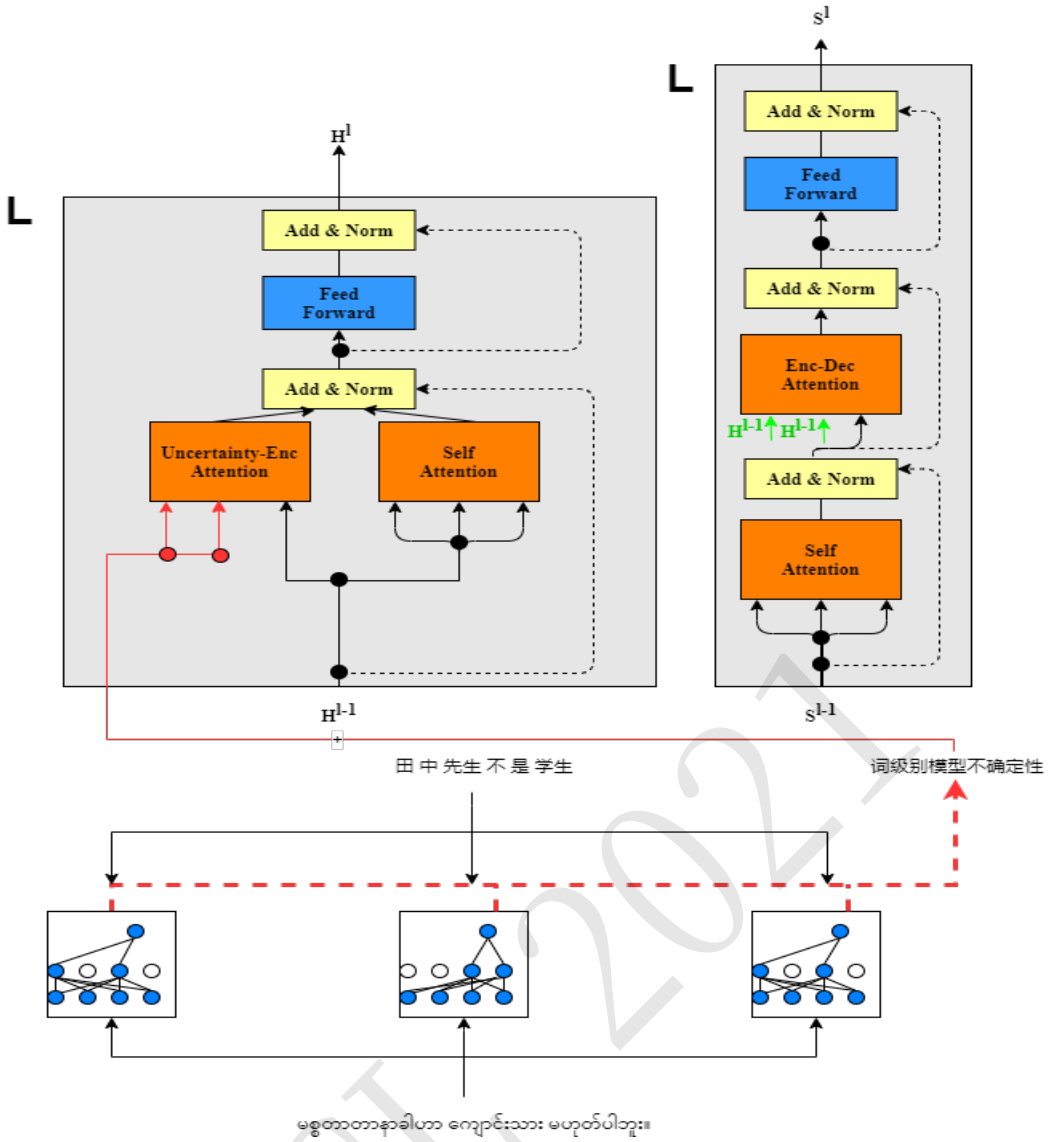

Figure 2: 模型不确定性为约束的汉缅机器翻译算法

其中x是输入，W_1，W_2，b_1，b_2是学习得到的参数。为使编码端既能学习回译质量较高的词的编码表征，又能有效降低质量较低的词（漏译，错译，多译的词）对句子有效向量表征的影响，本文提出（7）：在第l层时，$l \in [L]$，

$$h_i^l = \alpha(attn_s(h_i^l, H_E^{l-1}, H_E^{l-1})) + \gamma(attn_{un}(h_i^{l-1}, uncertainty_{bt}, uncertainty_{bt})) \quad (7)$$

其中$attn_s$和$attn_{un}$是不同参数的注意力机制模型（见公式(5)），α和γ是模型中的超参数（在第四部分我们会结合实验进一步探讨α和γ的取值和模型不确定性注意力机制与Self-Attention融合机制）。随后进一步送入公式(6)定义的非线性变换层FFN()，在此得到了能够有效处理回译语料中噪声的编码向量：$H_E^l = (FFN(h_i^l), ..., FFN(h_{l_x^l}))$。最后编码端会输出最后一层的隐藏层状态$H_E^L$。解码端是常规的transformer解码端，解码过程持续进行直到遇到结束的特殊字符为止。其中$attn_s$，$attn_{un}$分别表示self-attention和模型不确定性注意力机制。

3.3 模型不确定性注意力机制

Transformer的核心是运用多头的自注意力Self-attention机制，每一个注意力机制头都是对n个元素的输入序列$x = (x_1, ..., x_n)$，其中$x \in d$，接着计算得出同样长度的序列$c = (c_1, ..., c_n)$，其中$c \in d$。在此文中，我们用$x^{text} \in n \times d$标记文本句子向量特征，

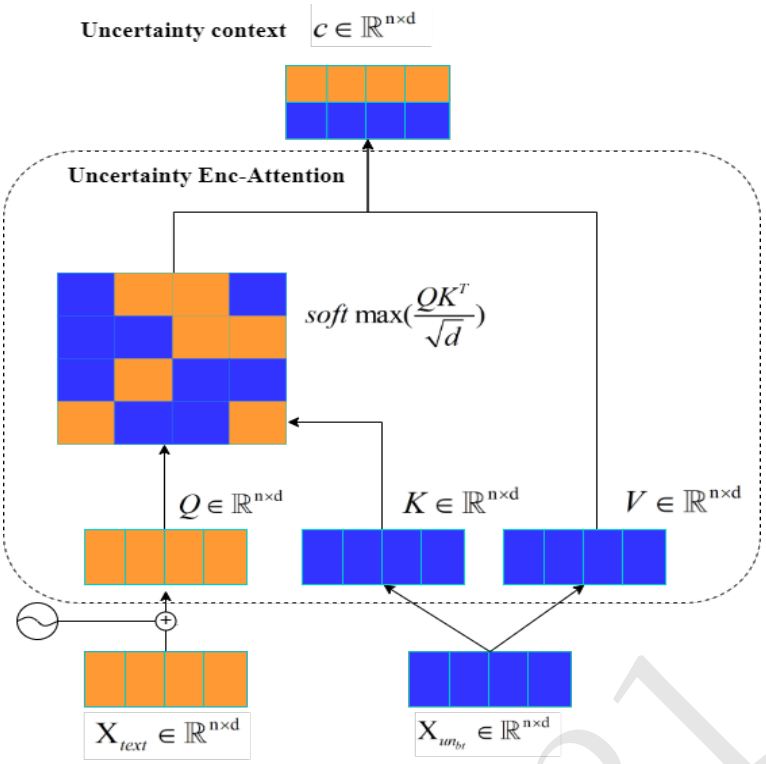

Figure 3: 模型不确定性注意力机制算法原理图

用$x^{un_{bt}} \in n \times d$标记与文本句子词对齐的模型不确定性。如图3所示，模型不确定性注意力机制可表示为：

$$c_i = \sum_{j=1}^{n} \hat{\alpha}_{ij}(x_j^{un_{bt}} W^V) \tag{8}$$

其中$\hat{\alpha}_{ij}$是softmax函数计算的权重系数：

$$\hat{\alpha}_{ij} = softmax(\frac{(x_i^{text} W^Q)(x_j^{un_{bt}} W^K)^T}{\sqrt{d}}) \tag{9}$$

模型不确定性注意力机制最后计算得出结合句子文本向量表征和模型不确定性表征的隐藏层向量c，其中$c \in n \times d$。在Transformer编码端最后一层，c被送入解码端去产生目标语言序列。能区分伪语料中噪声的句子编码向量是用文本句子向量$x^{text} \in n \times d$与该句子词对齐的模型不确定性值向量$x^{un_{bt}} \in n \times d$作注意力机制得到，这样句子编码表征可以捕捉到上下文的模型不确定性值，在编码时可以对模型不确定性值较大的词给予更多的关注，以此实现模型在编码过程中能更好的区分来自伪语料中的噪声数据，即漏译，错译，多译的词。在编码端每一层都采用了残差网络和层归一化。

4 实验

4.1 实验设置

我们在汉缅翻译任务验证了我们的方法，评价的方法是multi-bleu.perl脚本提供的BLEU (Papineni et al., 2002)计算方法。训练集是160k的汉缅双语平行语料，其中20k的汉缅双语平行语料来自于Asian Language Treebank(ALT) (Riza et al., 2016),其余的来自于多语言圣经语料库 (Christodouloupoulos and Steedman, 2015)以及人工收集，语料覆盖旅游，文学等领域。缅甸语的单语语料是维基百科上爬取的单语缅甸语文本段落，在分句，移除少于5个词，大于30个词的语句以后，缅甸语单语语料库规模为200k的缅甸语单语句子。测试集是训练集中截取查重后的5k双语平行语句。汉语和缅甸语的句子用分词粒度参数为16k的byte pair encoding (Sennrich et al., 2015b)进行预处理，汉语词典大小为10k，缅甸语词典大小为5k。本文将模

型不确定性约束的汉缅机器翻译算法应用在Transformer (Vaswani et al., 2017).的基础上。使用参数设置$\beta_1 = 0.9$fi$\beta_2 = 0.98$和$\epsilon = 10^{-9}$的Adam等人 (Kingma and Ba, 2014)优化器优化模型。我们参照Vaswani等人 (2017)同样使用参数设置$warmsteps = 4000$的warm-up策略来调整学习率。在模型训练期间，lable smoothing的超参数$\epsilon_{ls} = 0.1$ (Szegedy et al., 2016)。在训练和Monte Carlo Dropout的过程中，dropout的超参数设置为0.1，K设置成20。在实验中，我们的超参数β设置为2，α，γ取值将在4.2节详细讨论。所有试验在1 NVIDIA GTX 2080Ti GPU上进行。

4.2 编码端模型不确定性注意力机制与Self-Attention融合机制策略

4.2.1 模型不确定性注意力机制与Self-Attention融合比例

表1表明了公式(7)中α ： γ不同取值对实验结果的影响，由表可以得出α ： γ=0.6∶0.4的取值是一个粗糙的分界点，在α ： γ=0.6∶0.4取值的基础上减小模型不确定性注意力机制的融合比例会促进模型性能提升，初步探讨最大提升是当α ： γ=0.8∶0.2时最大提升值是24.72，较baseline提升了4.01个bleu点。由此可以看出α ： γ的不同取值对模型性能有不同的影响。

Algorithm	Bleu	Δ
Baseline	20.71	-
α ： γ=0.9∶0.1	22.52	+1.81
α ： γ=0.8∶0.2	24.72	+4.01
α ： γ=0.7∶0.3	16.47	-4.24
α ： γ=0.5∶0.5	16.73	-3.98

Table 1: α ： γ不同取值下模型不确定性注意力机制与Self-Attention融合比例实验结果，Encoder融合层数全为6层

4.2.2 模型不确定性注意力机制与Self-Attention融合Encoder层数

Algorithm	Bleu	Δ
Baseline	20.71	-
第一层	22.46	+1.75
前三层	23.86	+3.15
后三层	23.90	+3.19
六层	24.72	+4.01

Table 2: α ： γ=0.8∶0.2取值下模型不确定性注意力机制与Self-Attention在编码端融合层数探讨

实验证明在模型编码端融入句子中词对齐的$uncertainty_{bt}$对模型学习伪语料的句子表征是有实际意义的，在适当的模型不确定性注意力机制与Self-Attention融合比例下可协助编码端更好的处理伪语料中的噪声，但是当α ： γ=0.6∶0.4时模型性能有较大幅度的下降，表明编码端融入过多的表征会妨碍模型收敛，影响模型性能。进一步证明了编码端的模型不确定性注意力机制与Self-Attention融合比例和融合方式具有深远的可探讨意义。表2表明了在α ： γ=0.8∶0.2取值下模型不确定性注意力机制融合在编码端不同层数对实验结果影响，由表可以得出，模型不确定性注意力机制融合在Transformer的编码端每一层时模型的效果是最好的。将模型不确定性注意力机制融合在编码端前三层和后三层效果差别不大，当模型不确定性注意力机制只融合在第一层时此方法带给模型提升较小。

4.3 实验结果及对比分析

本文设置对比实验如下：

(1)Sennrich 等人 (2015a)利用有限的真实平行语料训练的神经机器翻译模型去生成的伪平行人造语料，再用得到的伪平行人造语料和真实平行语料一起训练模型的回译方法。

(2)Zoph等人 (2016)利用学习好的父模型参数迁移到低资源子模型方法改善低资源语言翻译性能。

(3)Wang等人 (2019)利用基于模型不确定的词级别置信度和句子级别置信度改善回译性能。

(4)Transformer (Vaswani et al., 2017)：我们将比较仅在Transformer模型上，不使用回译方法和模型不确定性方法的实验结果。

(5)结合4.2节实验结果，此节实验本文方法设置 α ： γ =0.8：0.2，模型不确定性注意力机制与Self-Attention融合编码端层数为6层。

Algorithm	Chinese→Myanmar	Myanmar→Chinese
Transformer	20.59	13.18
Zoph等	20.63	14.34
Sennrich等(回译)	20.71	14.07
Wang等	21.44	15.17
本文方法	24.72	15.95

Table 3: 主要实验结果

如表3所示，本文提出的方法模型不确定性约束的汉缅机器翻译算法在汉语-缅甸语翻译方向上BLEU值达到了24.72。相较于不加回译方法和的基础Transformer (2017)有明显的提升，提升了4.13个BLEU值，而Sennrich等人 (2015a)提出的传统回译方法只在基础Transformer上提升0.12个点，表明引入回译过程中的$uncertainty_{bt}$能更好的让模型处理回译过程中生成源语言单语伪语料多译，错译，漏译的问题，使模型在训练的过程中更好的处理伪平行双语语料库的噪声。同时，相较于Wang等人 (2019)利用基于模型不确定建模的词级别置信度和句子级别置信度（CEV）改善回译性能方法，本文方法也有明显提升，提升了3.28个BLEU值，证明本文的方法在处理低资源语言结构差异较大的语言对上，比Wang等人 (2019)提出$uncertainty_{bt}$的利用改变模型结构，仅修改Transformer中对数似然函数值和Self-attention权重值更为激进，也更有效。另外，本文提出的方法相较于Zoph等人 (2016)传统迁移学习依靠高资源语言预训练父模型改善低资源子模型的方法有明显的提升，提升了4.09个BLEU值，证明缅甸语和其他语言之间的语法差异性极大，直接利用迁移学习的思想将模型迁移到缅甸语上的效果性能不佳。在缅甸语-汉语翻译方向上，本文提出的方法BLEU值达到了15.95。相较于不加回译方法和模型不确定性的基础Transformer提升了2.77个BLEU值，而与汉语-缅甸语方向情况不同的点在于Sennrich等人 (2015a)的传统回译方法在基础Transformer提升幅度比汉语-缅甸语翻译方向大，提升了0.89个BLEU值点，此结果进一步证明了回译中的噪声数据会影响回译方法的性能，汉语-缅甸语方向翻译模型比缅甸语-汉语方向翻译模型训练的更充分，BLEU值更高，故传统回译方法运用在缅甸语-汉语翻译方向上（该回译模型为汉语-缅甸语翻译模型）效果更好，本文方法的核心点也是在于当回译模型受语料稀缺导致训练不充分时，能够借助词对齐的$uncertainty_{bt}$使模型能够更好的学习到伪语料的句子表征。同时，在缅甸语-汉语翻译方向上本文方法相较于Wang等人 (2019)提升了0.78个BLEU值点，证明本文方法具有较好的泛化性。

4.4 缅甸语-汉语翻译示例

在汉缅机器翻译中由于语料的稀缺和汉语，缅甸语语言结构差异性较大导致机器翻译的模型不能得到很好的训练，表现为对词编码过程中，词的不确定性很高，这种不确定性会导致译文存在漏译，多译，错译的问题如图4所示，缅甸语原文翻译成“因此，我们@@ 国家@@ 有很多@@ 园@@ 丁。”，对比汉语参考原文zh“因此，我国有很多花。”zh-base将“花”翻译成了“园丁”，这是典型翻译时错译的问题，同时我们在评估生成译文句子$uncertainty_{bt}$时，可以发现错译的词$uncertainty_{bt}$有较大的值（0.139943）；而当用本文的方法结合$uncertainty_{bt}$训练模型时，由my翻译得到的zh-un大体来说是基本正确的，值的一提的是本文方法模型不确定性注意力机制很好的处理错译的问题，在训练时将更多注意力集中在$uncertainty_{bt}$值较大的词上，本文方法不光将错译的词“园丁”翻译正确“花”，同时也可以看到“花”一词的$uncertainty_{bt}$也有所下降（值为0.130557）。如图4所示，缅甸语原文my翻译成了zh-base“小说，短@@ 篇@@ 小说，小说，短@@ 篇@@ 小说，小说，小说，小说@@ 和@@ 杂@@ 志。”，对比汉语参

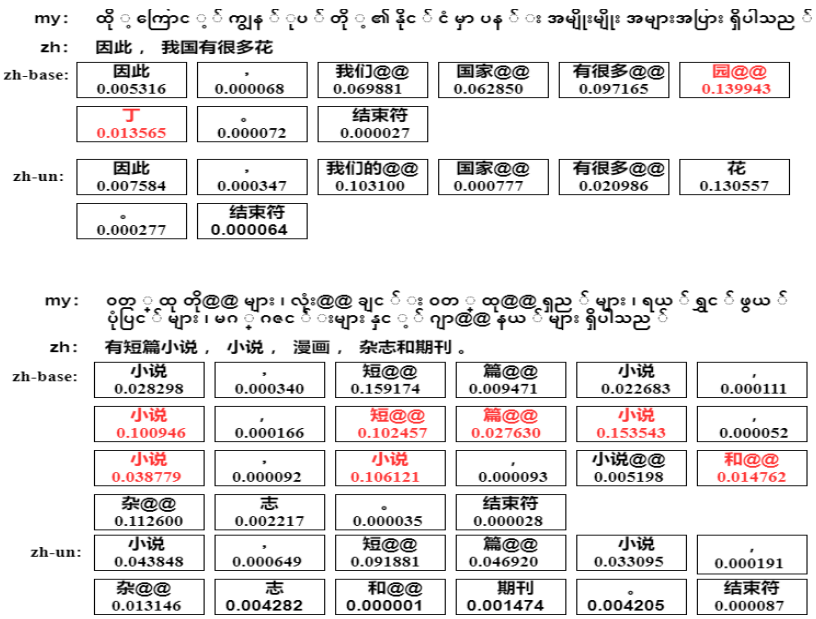

Figure 4: 缅甸语汉语方向结合翻译示例，my为缅甸语原文，zh是与缅甸语my互译的汉语参考原文，zh-base是my经传统回译方法训练的翻译模型得到的汉语译文，zh-un是my经uncertainty-fused Chinese-Myanmar NMT训练的翻译模型得到的汉语译文

考原文zh“有短篇小说，小说，漫画，杂志和期刊。”，zh-base存在严重的多译问题，“小说”和“短篇小说”在zh-base重复出现，同时我们在评估生成译文句子$uncertainty_{bt}$时，可以发现多译重复出现的词有较大的$uncertainty_{bt}$值（图4中红色标识的词）；而当用本文方法结合$uncertainty_{bt}$训练模型时，zh-un“小说，短@@ 篇@@ 小说，杂@@ 志和@@ 期刊。”中重复的词得到有效减少，同时观察到翻译得到的词的$uncertainty_{bt}$有明显下降。证明本文方法确实能让模型更好的学习好伪语料结合$uncertainty_{bt}$的句子表征，最终实现以基于模型不确定性约束的半监督汉缅神经机器翻译。

5 结论

针对汉语和缅甸语的语言差异性较大导致汉缅翻译回译方法利用单语语料过程中借助有限的真实汉缅平行语料训练的神经机器翻译模型去生成的源语言端人造语料富有噪声，存在语义错误，句子不通畅，漏译，多译，错译的问题。提出模型不确定性约束的汉缅机器翻译算法，探讨了将$uncertainty_{bt}$融进Transformer编码端每一层表征，用自注意力机制将$uncertainty_{bt}$与词嵌入embedding进行深度融合，使伪语料句子在编码端能得到更好的表征，在汉缅神经机器翻译和缅中神经机器翻译的任务上实验证明模型不确定性约束的汉缅机器翻译算法能有效提高翻译质量，提升回译性能，在16万的汉缅双语数据集上汉语-缅甸语翻译方向和缅甸语-汉语翻译方向我们分别达到了24.72的BLEU值和15.95的BLEU值，相比较于基线模型均有明显的提升。未来将探讨更完善的模型不确定性建模机理和把模型不确定性表征融入到transformer解码阶段。

参考文献

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Raj Dabre, Atsushi Fujita, and Chenhui Chu. 2019. Exploiting multilingualism through multistage fine-tuning for low-resource neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1410–1416.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. *arXiv preprint arXiv:1705.00440*.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. *arXiv preprint arXiv:1601.01073*.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, and Yoshua Bengio. 2017. On integrating a language model into neural machine translation. *Computer Speech & Language*, 45:137–148.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. *Advances in neural information processing systems*, 29:820–828.

Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *arXiv preprint arXiv:1703.04977*.

Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. 2015. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Philipp Koehn, Franz J Och, and Daniel Marcu. 2003. Statistical phrase-based translation. Technical report, UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY INFORMATION SCIENCES INST.

Surafel M Lakew, Aliia Erofeeva, Matteo Negri, Marcello Federico, and Marco Turchi. 2018. Transfer learning in multilingual neural machine translation with dynamic vocabulary. *arXiv preprint arXiv:1811.01137*.

Khin Thandar Nwet, Khin Mar Soe, and Ni Lar Thein. 2011. Developing word-aligned myanmar-english parallel corpus based on the ibm models. *International Journal of Computer Applications*, 975:8887.

Khin Thandar Nwet. 2011. Building bilingual corpus based on hybrid approach for myanmar-english machine translation. *International Journal of Scientific & Engineering Research*, 2(9).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Mārcis Pinnis, Rihards Krišlauks, Daiga Deksne, and Toms Miks. 2017. Neural machine translation for morphologically rich languages with improved sub-word units and synthetic data. In *International Conference on Text, Speech, and Dialogue*, pages 237–245. Springer.

Devendra Singh Sachan and Graham Neubig. 2018. Parameter sharing methods for multilingual self-attentional translation models. *arXiv preprint arXiv:1809.00252*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015a. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015b. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Shuo Wang, Yang Liu, Chao Wang, Huanbo Luan, and Maosong Sun. 2019. Improving back-translation with uncertainty-based confidence estimation. *arXiv preprint arXiv:1909.00157*.

Yijun Xiao and William Yang Wang. 2019. Quantifying uncertainties in natural language processing tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7322–7329.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.

Hammam Riza, Michael Purwoadi, Teduh Uliniansyah, Aw Ai Ti, Sharifah Mahani Aljunied, Luong Chi Mai, Vu Tat Thang, Nguyen Phuong Thai, Vichet Chea, Sethserey Sam, et al. 2016. Introduction of the asian language treebank. In *2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6. IEEE.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Language resources and evaluation*, 49(2):375–395.