面向低资源语言的一致性正则化半监督神经机器翻译

Viet H. Pham[1], **Thang M. Pham**[2*], **Giang Nguyen**[2*], **Long Nguyen**[1], **Dien Dinh**[1]1VNUHCM科学大学2奥本大学 {viethungpham0304, giangnguyenbkhn, long.hb.nguyen}@gmail.com thangpham@auburn.edu ddien@fit.hcmus.edu.vn

arXiv:2304.00557v1 [cs.CL] 2 Apr 2023

摘要

深度学习的出现为机器翻译带来了显著提升。然而，大多数研究需要规模庞大的平行数据集，而此类数据集稀缺、构建成本高，甚至对于某些语言而言根本不可获得。本文提出了一种简单而有效的方法，通过扩增高质量句对并以半监督方式训练神经机器翻译模型，以解决低资源语言中的这一问题。具体而言，我们的方法将用于有监督学习的交叉熵损失与用于无监督学习的KL散度相结合，后者基于从模型中获得的伪目标句和增强后的目标句子。我们还引入了一种基于SentenceBERT的过滤器，通过保留语义相似的句对来提升增强数据的质量。实验结果表明，我们的方法显著优于神经机器翻译基线方法，尤其是在低资源数据集上，BLEU 分数提升了0.46–2.03。我们还证明，对于增强数据，采用无监督训练比重复使用真实目标句进行有监督学习更为高效（Edunov 等人，2018年）。

1 引言

神经机器翻译（NMT）模型基于编码器-解码器架构（Sutskever等人，2014年；Gehring等人，2017年；Vaswani等人，2017b；Bahdanau等人，2015年；Luong等人，2015年），已被广泛用于有效提升翻译质量。

然而，训练神经机器翻译（NMT）模型需要大量平行数据和计算资源。只有英语、法国、德语和罗马尼亚语等部分高资源语言拥有大型平行语料库。对于低资源语言，则缺乏兼具高质量和大规模的现有语料。因此，为了让神经机器翻译模型在低资源语言上表现良好，许多研究正尝试利用数据合成和模型定制方法。

以回译（BT）方法为主的数据合成自2016年首次被提出用于神经机器翻译以来，已得到广泛应用（Sennrich等人，2016年）。此后，许多研究着重于将回译作为主要的数据增强方法来改进神经机器翻译模型。然而，神经机器翻译模型并不总能很好地应对数据增强引起的输入扰动，这可能导致模型的翻译质量大幅下降。因此，我们建议训练对数据增强具有不变性的神经机器翻译模型。

本文提出一种半监督方法，以解决神经机器翻译模型的数据短缺问题。我们利用平行语料和单语语料联合训练神经机器翻译模型。本文的核心思想是在传统有监督损失之外，为单语语料引入一个无监督目标。我们还利用SentenceBERT——一种多语言语义模型，过滤掉可能损害模型性能的不合格增强句子。我们的方法概览如图 1所示。我们的主要贡献概括如下：

- 我们提出一种半监督学习方法，利用增强数据提升翻译质量，效果优于性能强劲的基线模型。
- 我们引入语义过滤器，以提升通过增强方法生成的数据的质量。

2 相关工作

近年来，出现了若干利用单语语料提升神经机器翻译模型性能的方法。Cheng 等人，2016年提出了一种半监督

*共同贡献者

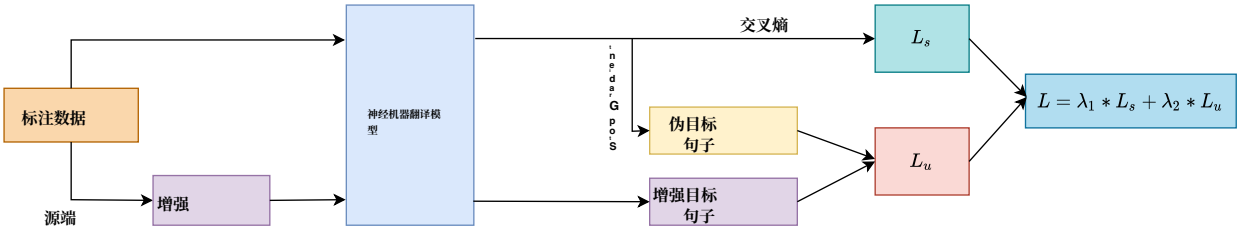


图1：我们提出的方法的工作流程。将标注数据输入神经机器翻译模型，以计算有监督损失（即交叉熵损失和软伪标签——上方分支）。另一方面，对标注数据的源端进行增强，并输入同一模型，以获得“增强后的目标句子”。然后，我们计算同一句子的“增强后的目标句子”与“伪目标句子”之间的接近程度（例如，如（Nguyen 等人，2019年，2021年）中的伪标签生成）。因此，最终目标是两项损失之和。

方法，利用编码器-解码器模型重构单语语料。Skorokhodov 等人，2018年 采用迁移学习，利用语言模型中的知识改进翻译系统。借助有监督源域数据，Jin 等人，2020年 提出了神经机器翻译的领域自适应方案，以提升翻译模型在缺少平行语料的目标领域中的性能。

回译（BT）（Edunov 等人，2018年；Sen- nrich 等人，2016年）是利用单语语料的另一种方法，它通过训练目标到源语言系统，从单语语料生成额外的合成平行数据。

与我们工作最接近的方法是 FixMatch（Sohn 等人，2020年），该方法将一致性正则化（Bachman 等人，2014年）应用于半有监督学习，并在图像分类任务上取得了最先进的结果。我们的方法在如何将这一概念用于翻译任务方面有所不同：使用 KL散度计算两个输出分布之间的语义距离（图1）。

3 一致性正则化机器翻译

如图所示1，所提出的方法非常简单。对于每个源-目标对，我们对源句应用回译方法以获得其增强版本，并用于计算伪目标与增强目标句之间的 KL散度。最终的损失函数由传统的交叉熵损失与 KL散度组合而成，如下所示：

$$L = \lambda_1 L_{CE} + \lambda_2 L_{KL} \quad (1)$$

其中， λ_1 和 λ_2 是用于平衡有监督损失与无监督损失的超参数。

数据集	训练集	Dev	Test
WMT14 英德	4.56M	3K	2.6K
IWSLT15 + 英越	133K	2K	2K
IWSLT17 + 英法	236K	888	1K
+ 英德	209K	888	1K

表1：数据集统计信息。英德：英语到德语；英越：IWSLT15英语到越南语；英法：IWSLT17英语到法国；英德：IWSLT17英语到德语。单位：句子数量。

4 实验

4.1 数据集

我们在两种设置下评估了该方法：使用 WMT14 英德的大规模语料库 [1]，以及使用 IWSLT 数据集的小规模语料库（IWSLT17 英语-法语、英语-德语和 IWSLT15 英语-越南语）。所有统计信息均显示在表1。

无标注数据构建：我们对每个语言对的英语端进行回译，以生成输入句子的增强版本。随后，过滤掉原始数据和增强数据中的所有重复示例，仅保留满足以下条件的示例：

$$\frac{\text{len}(x_i)}{\text{len}(u_i)} \in [\alpha_1, \alpha_2] \quad (2)$$

该标准确保两个句子的长度差异不会太大。在我们的实验中，我们将其设为[2] 和 $\alpha_1 = 0.7$ 。然后，我们使用 SBERT[3] 剔除相似度得分较低的句对（在我们的实验中为 < 0.5 ）。

[1]https://www.statmt.org/wmt16/translation-task.html[2]基于词语[3]https://www.sbert.net/

由于我们使用每个数据集中的英语部分进行扩充，因此 μ 设为 1。

在构建高质量无标注数据集后，对 IWSLT 数据集采用了 3000 次操作的字节对编码（BPE）（Sennrich 等人，2016c）；对于 WMT14 英语-德语，则采用了 30K 次操作。

4.2 模型与超参数

由于我们的方法不改变模型架构，因此可以使用任何神经机器翻译模型，如 Transformer、BART、长短期记忆网络等，作为骨干网络。受硬件限制，在我们的实验中使用了基于Transformer和基于 BART 的模型。超参数详见附录A。

4.3 结果

我们将所提出的方法与多种强大的基线方法进行比较，包括 Transformer（Vaswani等人，2017年）、BART（Lewis 等人，2020年）、Transformer 循环（Takase 和 Kiyono，2021年）以及回译（Edunov 等人，2018年）。表2展示了各种方法在测试集上的 BLEU 分数。如表中所示，我们的方法在所有数据集上显著提升了翻译质量，BLEU 分数从 +0.4（IWSLT17 英德）提升至 +2.0（IWSLT15 英越）。提升幅度最大的是英越，相较于原始 Transformer 高出 2.03 BLEU。值得注意的是，我们的方法还超过了回译；回译被认为是机器翻译中利用无标注数据集的最佳方法之一，至少提升了 +1.88 BLEU。对于英法和英德等其他低资源数据集，我们的模型也持续取得了更高的 BLEU 分数，相比 Transformer 基线方法，分别高出 1.18 和 0.46 。

在规模更大的数据集 WMT14 英德上，最佳结果以粗体标出，这证明一致性正则化框架在不同规模的数据集（即小规模和大规模数据集）上均能有效发挥作用。

4.4 消融实验

由于我们提出的方法包含两项损失，因此确定 λ_1 和 λ_2 的最优取值十分重要。此外，我们发现数据增强策略也显著提升了性能。在本节中，我们通过实验回答以下两个问题：

- 损失权重对最终结果有多重要?
- 增强策略如何影响翻译质量?

模型	IWSLT17		IWSLT15	WMT14
	英德	**英法**	**英越**	**英德**
基于Transformer	30.79	41.10	29.40	28.98
基于 BART	28.81	40.37	29.65	28.77
Transformer 循环				
+ 循环	29.76	40.87	30.47	28.32
+ 序列	29.84	40.80	30.72	29.01
+ 循环反向	29.29	40.52	30.07	28.24
回译	30.79	40.49	29.55	29.30
Ours				
+ Transformer	**31.25**	**42.28**	**31.43**	**29.39**
+ 基于 BART	30.39	41.30	29.96	25.90

表 2：各数据集上的 BLEU 分数。我们报告了使用不同骨干网络取得的结果。我们使用束宽为 5、长度惩罚为 0.6 生成翻译。

本节的所有消融研究均使用了三个 IWSLT 数据集（英越、英德和英法）。

4.4.1 损失权重

我们通过调整有监督和无监督部分的权重，研究了损失函数这两个组成部分之间的相互作用，同时仍施加约束 $\lambda_1 + \lambda_2 = 1.0$。如图 2所示，我们的模型在三个 IWSLT 数据集上均在 $\lambda_1 = \{0.4, 0.5, 0.6\}$时达到最佳性能，这与有标签和无标签样本的比例相近。此外，当 $\lambda_1 > 0.6$时，BLEU 分数容易大幅下降；而当 $\lambda_1 \leq 0.6$时，BLEU 分数则会显著上升。这表明 λ_1 与 λ_2 之间的比例应与有标签和无标签样本的比例相近。

4.4.2 增强策略

	IWSLT17		IWSLT15
	英德	英法	英越
回译	**31.25**	**42.28**	**31.43**
词语丢弃	30.68	41.27	26.93
同义词	30.13	40.90	28.11
词序	30.25	41.14	28.06

表 3：不同增强方法的实验。对于词语丢弃、同义词和词序，我们分别随机丢弃句子中的 $p = 0.2$ 个词元，使用预训练的 RoBERTa 替换同义词，并随机选择句子中相邻的两个词交换其位置。

如表 3所示，我们发现模型性能对数据增强方法相对敏感。具体而言，与回译和同义词替换相比，词语丢弃（Iyyer 等人，2015年）和词序调整的 BLEU 分数较低。一项

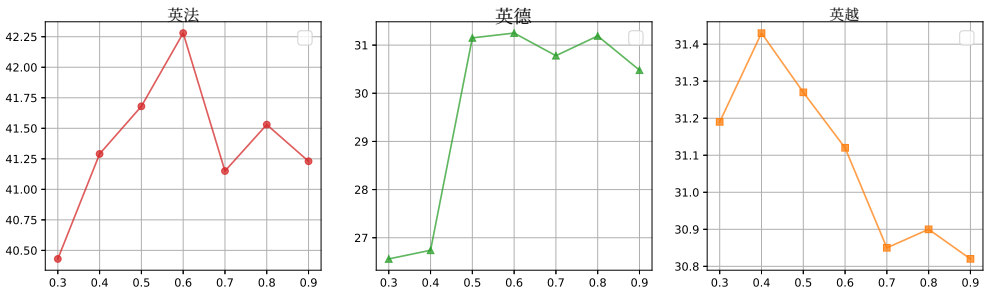


图2：改变 λ_1 时，模型在三个 IWSLT 数据集上的表现。纵轴表示 BLEU4 分数，横轴表示 λ_1 的取值，即交叉熵的权重。在本实验中，我们仍对 $\lambda_1 + \lambda_2 = 1.0$ 施加约束。

可能的解释是，句法信息遭到破坏，导致含义发生变化，使两个句子之间的 KL散度增大，进而造成翻译错误。与此同时，回译和同义词替换只会替换含义相同的词语，因此句子的句法和语义层面均保持不变。

5 结论

我们提出了一种简单的神经机器翻译半监督方法，同时利用平行语料和单语语料。无论是在小规模还是大规模数据集上，我们的方法都始终优于基线方法。我们还发现，增强策略的选择十分重要，并直接影响 BLEU 分数。未来的工作将从模型架构和数据增强方法方面改进我们的方法，并进一步探索其在不同低资源语料库设置下的有效性。

致谢

我们谨向 VinAIResearch 的人工智能工程师 Viet Phan 致谢，感谢您的宝贵讨论和意见。

参考文献

Philip Bachman、OuaisAlsharif 和 Doina Precup。2014年。伪集成学习。*arXiv*，abs/1412.4864。

Dzmitry Bahdanau、KyunghyunCho 和 Yoshua Ben- gio。2015年。通过联合学习对齐与翻译的神经机器翻译。载于第三届国际学习表征会议，*ICLR 2015*，圣迭戈，加利福尼亚州，美国，*2015*年*5*月*7*日至*9*日，会议论文集。

Yong Cheng、Wei Xu、Zhongjun He、Wei He、Hua Wu、Maosong Sun 和 Yang Liu。2016年。神经机器翻译的半有监督学习。载于计算语言学协会第*54*届年会论文集（第*1*卷：长篇论文），第1965–1974页，柏林，德国。计算语言学协会。

Sergey Edunov、Myle Ott、Michael Auli 和 David Grangier。2018年。深入理解大规模回译。载于*2018*年自然语言处理实证方法会议论文集，第489–500页，比利时布鲁塞尔。计算语言学协会。

Jonas Gehring、Michael Auli、David Grangier、Denis Yarats 和 Yann N. Dauphin。2017年。卷积序列到序列学习。国际机器学习大会'17。机器学习研究期刊。

Mohit Iyyer、Varun Manjunatha、Jordan Boyd-Graber 和 Hal Daumé III。2015年。深度无序组合在文本分类中的表现优于句法方法。载于计算语言学协会第*53*届年会暨第*7*届国际自然语言处理联合会议论文集（第*1*卷：长篇论文），第1681–1691页，中国北京。计算语言学协会。

Di Jin、Zhijing Jin、Joey Tianyi Zhou 和 Peter Szolovits。2020年。利用迭代回译进行神经机器翻译的无监督领域适应。*arXiv*，abs/2001.08140。

Diederik P. Kingma 和 Jimmy Ba。2015年。Adam：一种随机优化方法。计算研究资料库，abs/1412.6980。

Mike Lewis、Yinhan Liu、Naman Goyal、Marjan Ghazvininejad、Abdelrahman Mohamed、Omer Levy、Veselin Stoyanov 和 Luke Zettlemoyer。2020年。BART：用于自然语言生成、翻译和理解的序列到序列去噪预训练。载于计算语言学协会。

Thang Luong、Hieu Pham 和 Christopher D. Man- ning。2015年。基于注意力的神经机器翻译的有效方法。载于自然语言处理实证方法会议。

Giang Nguyen、Shuan Chen、Tae Joon Jun 和 Daey- oung Kim。2021年。通过深度可视化解释深度神经网络如何遗忘。载于模式识别。*ICPR* 国际研讨会与挑战赛：线上活动，*2021*年*1*月*10*日至*15*日，论文集，第三部分，第162–173页。施普林格。

Giang Nguyen、Tae Joon Jun、Trung Tran、Tolcha Yalew 和 Daeyoung Kim。2019年。Contcap：一个可扩展的持续图像描述框架。*arXiv* 预印本 *arXiv:1909.08745*。

Rico Sennrich、Barry Haddow 和 Alexandra Birch。2016年。利用单语语料改进神经机器翻译模型。载于计算语言学协会第*54*届年会论文集（第*1*卷：长篇论文），第86–96页，柏林，德国。计算语言学协会。

Rico Sennrich、Barry Haddow 和 Alexandra Birch。2016年。利用单语语料改进神经机器翻译模型。*arXiv*，abs/1511.06709。

Rico Sennrich、Barry Haddow 和 Alexandra Birch。2016c。使用子词单元进行低频词的神经机器翻译。载于计算语言学协会第*54*届年会论文集（第*1*卷：长篇论文），第1715–1725页，柏林，德国。计算语言学协会。

Ivan Skorokhodov、Anton Rykachevskiy、Dmitry Emelyanenko、Sergey Slotin 和 Anton Ponkratov。2018年。基于语言模型的半监督神经机器翻译。载于《*AMTA 2018* 低资源语言机器翻译技术研讨会论文集》（*LoResMT 2018*），第 37–44 页，波士顿，马萨诸塞州。美洲机器翻译协会。

Kihyuk Sohn、David Berthelot、Chun-Liang Li、Zizhao Zhang、Nicholas Carlini、Ekin Dogus Cubuk、Alexey Kurakin、Han Zhang 和 Colin Raffel。2020年。FixMatch：利用一致性与置信度简化半监督学习。*arXiv*，abs/2001.07685。

Ilya Sutskever、Oriol Vinyals 和 Quoc V. Le。2014年。神经网络的序列到序列学习。NIPS’14，第 3104–3112 页，剑桥，马萨诸塞州，美国。麻省理工学院出版社。

Sho Takase 和 Shun Kiyono。2021年。关于 Transformer模型 跨层参数共享的经验。*arXiv*，abs/2104.06022。

Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Ł ukasz Kaiser 和 Illia Polosukhin。2017年Attention Is All You Need。收录于神经信息处理系统进展，第30卷。Curran Associates, Inc.

Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N. Gomez、undefine- dukasz Kaiser 和 Illia Polosukhin。2017b。《注意力就是你所需要的一切》。载于第*31*届神经信息处理系统国际会议论文集，NIPS’17，第6000–6010页，美国纽约州雷德胡克。Curran Associates Inc.

附录：面向低资源语言的半监督神经机器翻译与一致性正则化

A 训练细节

我们使用 PyTorch 和 fairseq 工具包实现了 Transformer 模型[4]。在模型架构方面，所有实验采用相同的超参数，即4个 Transformer 模块、8个注意力头，词嵌入和前馈层的维度分别为1024和2048。Dropout 设为0.2，并对最后十个训练轮次的检查点取平均。采用 Adam 优化器（Kingma和Ba，2015年），并使用 $\alpha = 0.9$，$\beta = 0.98$ 学习率为0.0005，逆平方根调度器的初始值为 10^{-7}。所有模型均采用在词汇表上均匀分布的先验标签平滑$\epsilon = 0.1$。我们采用动态批次大小，每个批次共包含3000个词元。为避免过拟合，所有模型均使用0.0001的权重衰减，并训练30个训练轮次。

[4]https://github.com/facebookresearch/fairseq