

文章编号: 1003-0077(2021)03-0051-09

## 利用质量估计改进无监督神经机器翻译

徐 佳, 叶 娜, 张桂平, 黎天宇

(沈阳航空航天大学 人工智能研究中心, 辽宁 沈阳 110136)

**摘 要:** 传统上神经机器翻译依赖于大规模双语平行语料, 而无监督神经机器翻译的方法避免了神经机器翻译对大量双语平行语料的过度依赖, 更适合低资源语言或领域。无监督神经机器翻译训练时会产生伪平行数据, 这些伪平行数据质量对机器翻译最终质量起到了决定性的作用。因此, 该文提出利用质量估计的无监督神经机器翻译模型, 通过在反向翻译的过程中使用质量估计对生成的伪平行数据评分, 再选择评分 (HTER) 较高的平行数据训练神经网络。利用质量估计的方法可以控制反向翻译生成的伪平行数据的质量, 为对抗生成网络提供了更丰富的训练样本, 使对抗生成网络训练得更加充分。与基线模型相比, 该模型在 WMT 2019 德语—英语和捷克语—英语新闻单语语料上 BLEU 值分别提升了 0.79 和 0.55。

**关键词:** 无监督神经机器翻译; 反向翻译; 质量估计

**中图分类号:** TP391

**文献标识码:** A

## Improving Unsupervised Neural Machine Translation with Quality Estimation

XU Jia, YE Na, ZHANG Guiping, LI Tianyu

(Human-Computer Intelligence Research Center, Shenyang Aerospace  
University, Shenyang, Liaoning 110136, China)

**Abstract:** Traditionally, neural machine translation relies on large-scale bilingual parallel corpora. In contrast, unsupervised neural machine translation avoids the dependence on bilingual corpora by generating pseudo-parallel data, whose quality plays a decisive role in the model training. To ensure the final quality of machine translation, we propose an unsupervised neural machine translation model using quality estimation to control the quality of pseudo-parallel data generated. Specifically, in the process of back-translation, we use quality estimation to score the generated pseudo-parallel data, and then select parallel data with higher score (HTER) to train the neural network. Compared with the baseline system, the BLEU scores are increased by 0.79 and 0.55, respectively, on WMT 2019 German-English and Czech-English monolingual news corpora.

**Keywords:** unsupervised neural machine translation; back-translation; quality estimation

## 0 引言

最近, 神经机器翻译 (neural machine translation, NMT) 的方法在机器翻译领域中脱颖而出, 取得了极大的进展<sup>[1-3]</sup>。神经机器翻译通常由两个子神经网络组成, 编码器网络将源语言句子编码成上下文向量, 再由解码器网络将编码器网络编码的上下文向量迭代解码成相同意思的目标语言句子。通常在

有监督的环境下, 需要使用大量的双语平行句子对训练模型。然而绝大多数语言对几乎没有平行数据。无监督神经机器翻译 (unsupervised neural machine translation, UNMT) 成功打破了这种限制, 仅仅使用两种语言的单语语料进行训练便可以完成常规的机器翻译任务。无监督神经机器翻译是基于神经机器翻译的, 并结合去噪自动编码器和反向翻译<sup>[4-5]</sup>训练双重模型初始化与跨语言嵌入<sup>[6]</sup>, 达到机器翻译的目标。但是在无监督神经机器翻译模型训

收稿日期: 2020-02-07 定稿日期: 2020-05-12

基金项目: 教育部人文社会科学研究青年基金 (19YJC740107); 国家自然科学基金 (U1908216); 辽宁省重点研发计划 (2019JHZ/10100020)

练过程中,反向翻译会产生大量的伪平行数据,而这些伪平行数据的质量在反向翻译训练中就变得至关重要。因此,在训练反向翻译的过程中控制伪平行数据质量是提升无监督神经机器翻译质量的一个关键。

本文提出了一种利用质量估计(quality estimation, QE)技术的方法,在无监督神经机器翻译训练反向翻译时设置一个固定阈值筛选伪平行数据,有效地提升了无监督神经机器翻译的效果。本文研究基于生成对抗网络<sup>[7]</sup>,由反向翻译和语言建模分别作为生成器和鉴别器。通过使用质量估计评估并筛选反向翻译训练过程中生成的评分(HTER)比较高的伪平行数据。该方法在控制伪平行数据质量的同时,丰富了生成器的生成样本,让鉴别器收敛得更慢,使得生成对抗网络训练得更加充分,从而提升了无监督神经机器翻译的翻译效果。

本文在 WMT 2019 共享任务提供的德语—英语和捷克语—英语的单语语料上进行了无监督神经机器翻译的实验,其中质量估计模型采用预测器—估计器结构是在 WMT 2019 共享任务提供的德语—英语和捷克语—英语质量估计语料上训练得到的。首先需要训练预测器模型,然后再使用预测器模型来训练估计器模型,最后通过该模型的预测器对反向翻译训练过程中生成的所有伪平行数据进行评分和筛选。实验结果表明,利用质量估计的无监督神经机器翻译模型虽然训练速度有些下降,但是在 WMT 2019 德语—英语和捷克语—英语新闻单语语料上翻译性能分别提升了 0.79 和 0.55 个 BLEU 值。

本文组织结构安排如下:第 1 节论述了本研究的相关工作;第 2 节和第 3 节分别介绍了无监督神经机器翻译和质量估计原理;第 4 节说明了利用质量估计改进无监督 NMT 的方法;第 5 节给出了实验结果及其分析;最后部分为总结与展望。

## 1 相关工作

目前,学界已经提出了几种不直接使用双语平行语料训练 NMT 模型的方法。Leng 等人<sup>[8]</sup>使用的方法是训练从源语言翻译到枢轴语言,再将枢轴语言翻译到目标语言的独立翻译。这种方法虽然巧妙,但却对枢轴语言的依赖性很高,同时无形之中引入了第三种语言,甚至是第四种语言互译的误差,而且通过枢轴语言引入的误差无法消除。

由于跨语言嵌入的成功, Lample 等人和

Artetxe 等人同时提出了基于预训练跨语言嵌入训练无监督神经机器翻译的方法。该方法仅仅使用了两种语言的单语语料独立地训练两种语言的嵌入,并学习线性变换<sup>[9]</sup>和对抗训练<sup>[6]</sup>将它们映射到共享空间中。由此产生的跨语言嵌入用于初始化两种语言的共享编码器,整个系统使用去噪自动编码器、反向翻译和对抗训练<sup>[10]</sup>组合训练。Yang 等人<sup>[11]</sup>进一步改进了这种方法,他们使用两种语言特定的编码器,只共享其参数的一个子集,并结合本地和全局生成对抗网络训练无监督神经机器翻译。但这些方法训练模型时并未考虑到反向翻译生成伪平行数据的质量,导致模型训练不够充分。

现阶段,反向翻译技术经常被应用于无监督神经机器翻译中。Wang 等人<sup>[12]</sup>使用质量估计的方法,即使用 OpenKiw<sup>[13]</sup>筛选反向翻译生成的伪平行数据与真实的平行数据结合的方法训练 NMT,提升了机器翻译的性能。Li 等人<sup>[14]</sup>利用反向翻译使用单语语料扩充平行数据,采用新的数据增强方法在保持模型小的前提下进一步提高 NMT 对噪声的鲁棒性和翻译效果。Miguel 等人<sup>[15]</sup>采用合成数据生成的方案,在 NMT 模型的交叉熵优化范围内重新构造了反向翻译,阐明了其基本的数学假设和启发式用法之外的近似值。利用公式指出了基于采样的方法的基本问题,并提出通过禁用目标到源模型的标签平滑和从受限搜索空间采样来解决神经机器翻译数据增强问题。但 Wang 等人、Li 等人和 Miguel 等人的方法均使用了有监督的方法来解决数据不足和控制伪平行数据质量的问题,并没有解决不使用双语平行语料训练机器翻译任务的问题。

本文的方法基于 Lample 等人<sup>[5]</sup>的无监督 NMT 的方法,仅仅使用两种单语语料利用质量估计筛选伪平行数据训练反向翻译的同时,使用更少的数据训练并提升了无监督神经机器翻译的性能。

## 2 无监督神经机器翻译模型

本文基于 Lample 等人<sup>[5]</sup>的研究,提出的模型由编码器和解码器组成,其整体架构如图 1 所示。

**初始化** 虽然之前大部分的研究依赖于双语词典,但是使用 Conneau 等人<sup>①</sup>提出的预训练跨语言嵌入的方法更加简单、有效。首先,对于两种语言的

<sup>①</sup> 使用两种不相关语言的单语语料初始化一个跨语言嵌入文本。

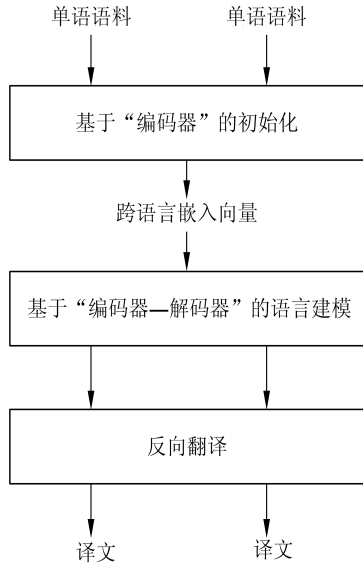


图1 无监督神经机器翻译模型整体架构

语料做字节对编码<sup>[16]</sup> (byte pair encoder, BPE) 处理。该处理减少了词表大小,并消除了输出译文中出现的未知单词。其次,通过联合处理两种语言的语料,应用联合 BPE 的一部分共享单词消除了双语词典的需要。最后,在相同的两种语言联合语料库上学习标记嵌入<sup>[17]</sup>,使用这些嵌入来初始化编码器和解码器中的查找表。

**语言建模** 语言建模通过去噪自动编码器实现,分别针对源语言和目标语言进行优化并更新神经网络参数,使得反向翻译生成的伪平行数据越来越好。最小化目标函数如式(1)所示。

$$L^{\text{lm}} = E_{x \sim S} [-\log P_{s \rightarrow s}(x | C(x))] + E_{y \sim T} [-\log P_{t \rightarrow t}(y | C(y))] \quad (1)$$

其中,  $C$  是一个部分单词被删除的噪声模型。 $P_{s \rightarrow s}$  和  $P_{t \rightarrow t}$  分别由源端和目标端工作的编码器和解码器组合而成。

**反向翻译** 在反向翻译过程中,通过使用普通的编码器和解码器实现,最小化目标函数如式(2)所示。

$$L^{\text{back}} = E_{y \sim T} [-\log P_{s \rightarrow t}(y | u * (y))] + E_{x \sim S} [-\log P_{t \rightarrow s}(x | v * (x))] \quad (2)$$

其中,  $u * (y)$  表示从  $y \in T$  推导出来的源语言句子,  $u * (y) = \arg\max P_{t \rightarrow s}(u | y)$ 。同样,  $v * (x)$  表示从  $x \in S$  推导出来的目标语言句子,  $v * (x) = \arg\max P_{s \rightarrow t}(v | x)$ 。将  $(u * (y), y)$  和  $(x, v * (x))$  组成自动生成的伪平行句子对。注意,最小化  $L^{\text{back}}$  这个目标函数时并没有产生数据,在随机梯度下降的每次迭代中,最小化目标函数式(1)中的  $L^{\text{lm}}$  和

式(2)中的  $L^{\text{back}}$  之和。

语言建模用于优化反向翻译生成的句子,反向翻译用于翻译模型。而 Lample 等人忽略了用于翻译模型训练的伪平行数据的质量以及生成对抗网络训练样本的丰富性,因此我们引入了质量估计模型。

### 3 质量估计模型

本文使用了质量估计模型——OpenKiwi<sup>[13]</sup>。OpenKiwi 是一个开源的质量估计框架,其实现了过去几年中比较流行的四个系统: QUETCH<sup>[18]</sup>, NUQE<sup>[19]</sup>, 预测器—估计器<sup>[20]</sup> 和线性堆叠系统<sup>[19]</sup>。本文的研究基于预测器—估计器模型,下面将介绍 OpenKiwi 的预测器—估计器模型。图 2 为 OpenKiwi 的预测器—估计器模型整体架构。

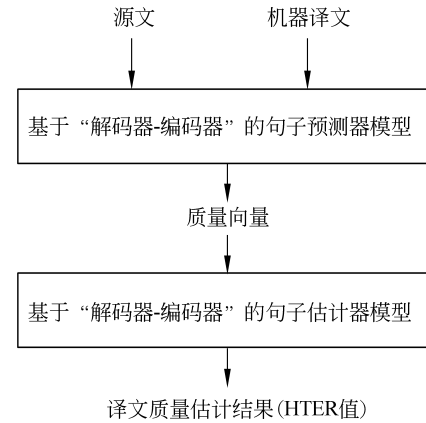


图2 OpenKiwi 的预测器—估计器模型整体架构

OpenKiwi 遵循 Kim 等人提出的架构,包含两个部分: 预测器和估计器。预测器用于在给定目标语言句子的源语言句子左、右上下文的情况下,预测出目标语言句子的每一个词。估计器用于接收由预测器产生的特征,并将每个词分类为 OK/BAD。预测器使用双向 LSTM 编码源语言,两个单向 LSTM 按从左到右 (LSTM-L2R) 和从右到左 (LSTM-R2L) 的顺序处理目标语言。对于每个目标单词  $t_i$ , 其左上下文和右上下文的表示被连接起来,并在最终的 softmax 层之前用作对注意模块的查询。它由 WMT 共享任务组织者提供,作为附加数据的大型并行语料库进行训练。估计器将特征序列作为输入: 对于每个目标单词  $t_i$ , 在 softmax 之前的最后一层(在处理  $t_i$  之前),以及 LSTM-L2R 和 LSTM-R2L 的第  $i$  隐藏状态的串联(在处理  $t_i$  之后)。此外,使用一个多任务架构来训练这个系统,以预测句子级的 HTER 值。总的来说,该系统能够预测句子

级分数和所有单词级标签(对于机器翻译单词、间隙和源单词),源单词标签是通过反向训练预测器产生的。

本文主要使用了 OpenKiwi 的神经质量估计模型(即预测器—估计器模型)对反向翻译生成的伪平行数据进行评估,再通过该模型给出的句子级 HTER 值对伪平行数据进行筛选。

#### 4 融合质量估计的无监督 NMT 模型

为解决神经机器翻译需要大量双语平行语料的问题,无监督神经机器翻译采用反向翻译的方法训练翻译模型。反向翻译的方法一直是在无监督神经机器翻译领域广泛使用的方法,通常在最小化损失函数时,不通过产生数据的模型进行反向传播,这样做是为了训练反向翻译的简单性<sup>[5]</sup>,但是却忽略了神经机器翻译训练使用数据质量的重要性。因此,本文提出了利用质量估计来控制反向翻译训练过程中产生的伪平行数据质量的方法,对应的无监督 NMT 单次训练流程如图 3 所示。

##### 4.1 质量估计模型训练

本文选择并训练了质量估计模型 OpenKiwi。为了训练 OpenKiwi 质量估计模型,我们选择了使用 WMT 2019 共享任务提供的质量估计测试集、验证集和训练集。为简化系统加载模型的复杂性,我们仅训练了英语—德语和英语—捷克语的质量估计模型,这样避免了反向翻译筛选伪平行数据时由于使用不同的模型而引入误差的问题。最终,本文分别使用皮尔逊系数为 58.51 和 54.91 的英语—德语和英语—捷克语的句子级质量估计模型融入到反向翻译中。

##### 4.2 融合质量估计的反向翻译

反向翻译的方法(又称作往返翻译),主要是利用编码器和解码器在两种语言单语语料上分别训练独立的翻译模型。因此,无监督神经机器翻译在训练过程中会通过两个方向轮流迭代训练同一个编码器和解码器,最后将该编码器和解码器应用到常规机器翻译中。

常规反向翻译的方法是一个将源语言翻译到目标语言再翻译回源语言的一个过程(图 4),假设源语言句子为 S,目标语言句子为 T,则常规反向翻译即  $S \rightarrow T \rightarrow S$ 。本文所使用的方法需要在反向翻译

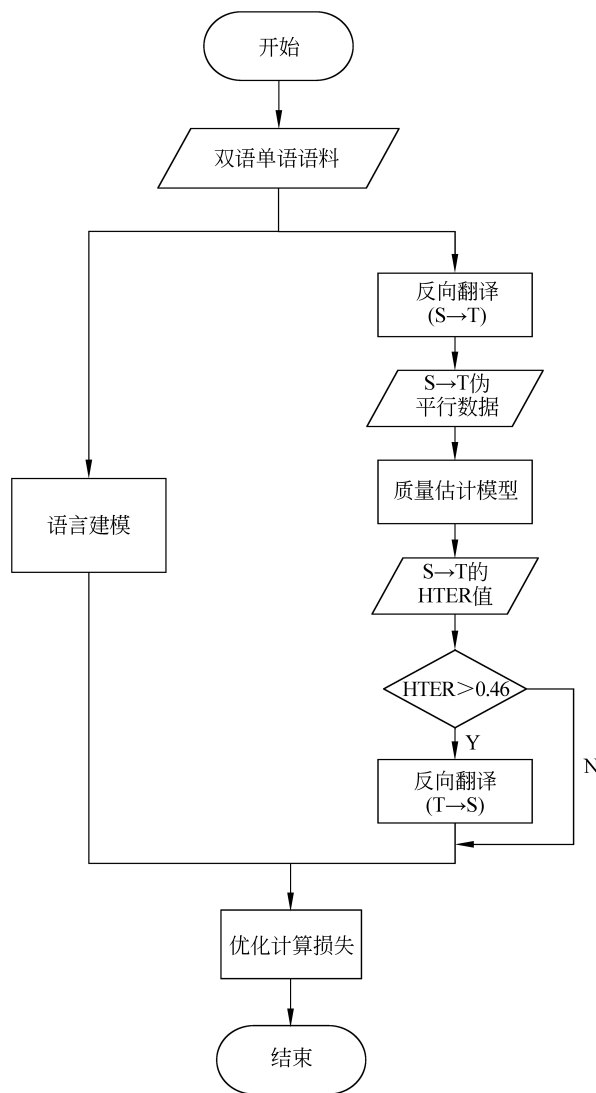


图 3 无监督 NMT 单次训练流程图

S 表示源语言句子, T 表示目标语言句子, 我们人为地将反向翻译分为两步

过程中使用 QE 模型, 首先, 将常规反向翻译过程分为两个阶段: ①将源语言句子翻译成目标语言句子, 即  $S \rightarrow T$ ; ②将①翻译出的目标语言句子翻译回源语言句子, 即  $T \rightarrow S$ 。然后, 在①阶段后加入 QE 模型对伪平行句子对进行质量评估, 评估结果为 QE 模型给出的 HTER 值, 再将 QE 模型给出的 HTER 值与阈值进行比较后筛选数据, 在第 4.3.2 节中会介绍阈值的选择方法。最后将满足 HTER 值大于阈值的伪平行数据继续执行②阶段并计算损失, 优化后更新神经网络参数, 结束当前反向翻译过程。

以德语—英语为例, 下文德语使用 De 表示, 英语使用 En 表示。在一次训练中, 反向翻译可能会先从德语语料中选取一个批次的德语训练  $De \rightarrow$



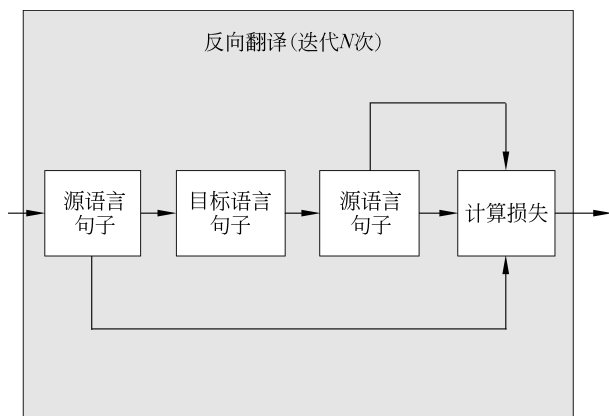


图4 常规反向翻译示意图

En→De,随后再从英语语料中选取一个批次的英语训练 En→De→En,最后反复迭代并不断优化模型。因此我们在一次训练中既要评估 De→En→De 伪平行数据也要评估 En→De→En 伪平行数据。我们选择在反向翻译的①阶段进行评估(即 De→En 和 En→De),而在反向翻译的②阶段并没有进行评估。其原因在于反向翻译也是神经网络训练机器翻译的过程,②阶段翻译的输入数据是由①阶段翻译得到的数据,若①阶段翻译所得数据 HTER 值不高,则②阶段翻译所得数据 HTER 值必然不高。并且,在一次反向翻译迭代中进行两次质量评估耗时更多,因此本文选择在反向翻译的①阶段进行评估。图5即为利用质量估计的反向翻译示意图。

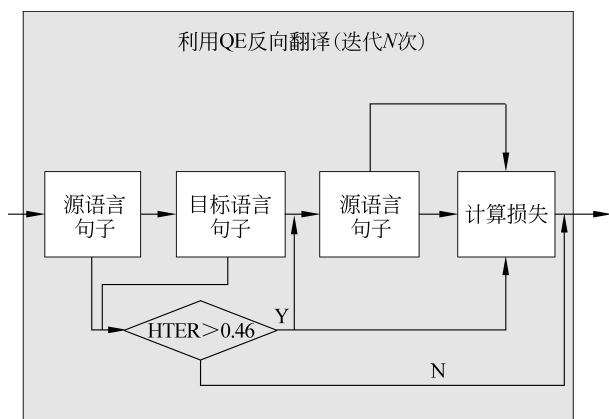


图5 利用质量估计的反向翻译示意图

### 4.3 人工翻译错误率与阈值选择

#### 4.3.1 人工翻译错误率

质量估计模型的估计器最终评估结果为人工翻译后编辑率(human translation error rate, HTER)。HTER<sup>[21]</sup>值是人工修正译文所需要编辑的百分比。

HTER 值的取值范围在 0~1 之间,越接近 0 代表翻译得越好,不需要人工编辑,越接近 1 代表翻译得越不好,需要多次人工编辑。

#### 4.3.2 阈值选择及使用

句子级的质量估计模型,将翻译出来的平行句子对放入到该质量估计模型,再由质量估计模型预测出 HTER 值供我们进行筛选。最后再将该批次评估得到的数据按照式(3)计算,作为该批次伪平行数据的评估结果。注意,我们为了提升评估效率,采用了将一个批次的伪平行数据放入到模型中预测的方式。

$$H_{batch} = (H_1 + \dots + H_i + \dots + H_n) / n \quad (3)$$

其中, $n$  为一个批次的句子的数量, $H_i$  为一个批次中其中一对伪平行数据的 HTER 值。筛选伪平行数据的阈值则是通过多次实验,取生成句子对的 HTER 最高值( $H_{max}$ )与最低值( $H_{min}$ )的平均值,如式(4)所示。

$$\beta = (H_{max} + H_{min}) / 2 \quad (4)$$

其中, $\beta$  为筛选平行数据的阈值(最终,阈值选择为 0.46 和 0.49)。最后,通过固定阈值筛选反向翻译所产生的伪平行数据,再使用满足该阈值的伪平行数据训练并更新神经网络参数。

对于阈值的设置,由于反向翻译训练过程中生成伪平行数据的质量是越来越好的,并且阈值和 HTER 值均不是固定值,若使用动态阈值的方法无法控制伪平行数据的留取比例,无法对实验结果进行分析,因此我们并没有选择使用设置动态阈值的方法,而是采取了设置固定阈值的方法来控制伪平行数据的质量。

## 5 实验与结果

本节首先描述我们使用的数据集和实验,然后将本文使用的方法与 Lample 等人的方法对比,最后针对本文的实验结果进行分析。

### 5.1 数据与预处理

OpenKiwi 使用了 WMT 2019 共享任务提供的英语—德语和捷克语—英语语料。其中包括随机抽取并打乱顺序的 500 万句子对作为训练集,以及验证集各 1 000 句。表 1 为 OpenKiwi 所使用的 De-En 和 Cs-En 数据信息。

表 1 OpenKiwi 所使用的数据集信息 (单位: 句)

信息	En-De		Cs-En	
	train	valid	train	valid
预测器	5 000 000	1 000	5 000 000	1 000
估计器	13 442	1 000	40 254	1 000

对于无监督 NMT,我们选取与 Lample 等人和 Artetxe 等人相同的数据集(WMT 2019 新闻语料)。首先分别将单语数据(WMT 2019 提供的各年份新闻单语语料)混合在一起,再分别将数据打乱顺序,并从中随机抽取 1 000 万句德语和英语的单语语料。测试集为 newstest2016 共 2 999 句,验证集为 newsdev2015 共 2 169 句。表 2 为无监督 NMT 所使用的 De-En 数据信息。

表 2 无监督 NMT 所使用的 De-En 数据信息 (单位: 句)

信息	train	test	valid
句子数量	10 000 000	2 999	2 169

为了验证方法的有效性,我们还选择了由 WMT 2019 共享任务提供的捷克语单语语料。与前文提到的方式相同,随机抽取 1 000 万句捷克语单语句子。捷克语测试集和验证集分别为 newstest2014 共 3 003 句和 newstest2015 共 2 656 句。英文使用的语料与上文所述相同。表 3 给出无监督 NMT 所使用的 Cs-En 数据信息。

表 3 无监督 NMT 所使用的 Cs-En 数据信息 (单位: 句)

信息	train	test	valid
句子数量	10 000 000	3 003	2 656

本文所述实验数据均使用 Moses<sup>[22]</sup> 提供的 tokenizer.perl 脚本处理。

## 5.2 初始化

无监督 NMT 的方法需要使用跨语言的 BPE<sup>[23]</sup> 嵌入<sup>①</sup>(来初始化共享查找表)。我们使用 fastText<sup>[24]</sup> 来生成嵌入,其嵌入维度为 512,上下文窗口大小为 5,负样本数为 10。将 fastText<sup>②</sup> 应用于源语言语料和目标语言语料的联合语料,实现了跨语言的 BPE 嵌入。

## 5.3 质量估计模型训练

本文首先使用第 5.1 节提到的 500 万训练数据

训练预测器,最终 En-De 和 En-Cs 预测器训练结果的 BLEU 值分别为 44.77 和 43.75。再使用质量估计语料训练估计器,最终 En-De 和 En-Cs 估计器训练结果的皮尔逊系数分别为 58.51 和 54.91。表 4 为本文所述实验最终使用的 OpenKiwi 质量估计模型的测试结果。

表 4 质量估计模型的测试结果

模型	OpenKiwi(预测器—估计器模型)				
	MT	gaps	source	$\gamma$	$\rho$
En-De	44.77	22.89	36.53	46.72	58.51
Cs-En	43.75	23.66	37.33	56.08	54.91

## 5.4 模型超参数和评估

对于无监督 NMT 模型,本文采用 Transformer 架构,在编码器和解码器中都使用了 4 层(其中 3 层编码器和解码器参数共享)、Multi-head Attention 参数为 8。在编码器和解码器之间、源语言和目标语言之间共享所有查找表。嵌入和隐藏层的维度设置为 512。使用 Adam 优化方法,学习率为  $10^{-4}$ , $\beta_1=0.5$ ,Batch Size 为 32,dropout 为 0.1。表 5 为本文采用的 Transformer 架构参数设置。

表 5 Transformer 架构参数设置

参数	值
网络层数	4
Multi-head Attention	8
词嵌入维度	512
隐藏层维度	512
学习率	0.0004
$\beta_1$	0.5
Batch Size	32
dropout	0.1

## 5.5 模型选择

对于模型选择,本文在模型连续 10 次没有改进的情况下停止对于模型的训练。BLEU<sup>[25]</sup> 被用作评估指标。对于德语—英语和捷克语—英语,我们使用由 Moses<sup>③</sup> 提供的 multi-bel.perl 脚本评估翻译性能。

① <https://github.com/glampl/fastBPE>

② <https://github.com/facebookresearch/fastText>

③ <http://www.statmt.org/moses/>

我们考虑了两个模型选择的过程：基于 Lample 等人的“反向”翻译(source→target→source 和 target→source→target)的 BLEU 值的无监督准则,以及使用包含 100 个平行语句的小验证集的交叉验证。在实验中发现,使用 Transformer 模型时,无监督准则与测试指标高度相关。因此,我们使用 100 个平行语句的小型验证集选择的无监督准则最佳的 Transformer 模型。

## 5.6 实验结果与分析

通过使用第 4.3.2 节所述的方法计算,在本文所述实验中所训练的质量估计模型对伪平行数据评估得出,其中英语—德语(英语—捷克语)HTER 最大值为 0.61(0.65),最小值为 0.31(0.33),因此将反向翻译训练时德语—英语(和捷克语—英语)所需阈值分别设置为 0.46(和 0.49)。与此同时,我们还针对该阈值进行放大和缩小,分别进行对比实验。放大后的阈值不能大于 0.61 且缩小后的阈值不能小于 0.31,于是我们选择 0.56 和 0.36 分别进行实验。表 6 为德语—英语 Baseline 模型与 UNMT+QE 模型的数据筛选比结果<sup>①</sup>以及其第 1 个 epoch 评测的 BLEU 值对比。由表 6 可见,本文所述方法成功地利用质量估计模型控制了无监督神经机器翻译训练反向翻译时所生成的伪平行数据质量。

表 6 德语—英语 Baseline 模型与 UNMT+QE 模型

的数据筛选比 (单位: word/s)

模型	阈值(HTER)	比率/%	BLEU(De-En)
Baseline	—	100	1.02
UNMT+QE	0.56	13.11	0.78
	0.46	53.21	2.43
	0.36	62.53	2.03

与 Baseline 模型相比,我们的模型丰富了由反向翻译的生成样本,使鉴别器收敛得更慢,使得生成对抗网络训练得更加充分,从而提升无监督神经机器的翻译效果。由第 4.3.2 节所述方法得到的阈值为 0.46(0.49)。由表 6 可见,QE 模型在筛选句子时减少了句子的数量,同时提高了模型训练的质量(BLEU 值)。当阈值为 0.36 时 UNMT+QE 模型的数据筛选比相对阈值为 0.46 时高 9.32%,但 BLEU 值却比阈值为 0.46 时要低;当阈值为 0.56 时,模型筛选去掉的数据数量比较多,不利于模型训练。实验结果表明,当筛选掉的数据与保留的数据

比接近 1:1 时模型取得的 BLEU 值比较高,实验中既不能筛选掉过多的数据,也不能保留过多的数据,因此本文在德语—英语和英语—捷克语实验时阈值选择了第 4.3.2 节所述方法,分别设置为 0.46 和 0.49。

表 7 为 Baseline 模型与本文模型结果数据对比。本文所提出的模型比基线系统在英语—德语和捷克语—英语翻译上 BLEU 值分别提升了 0.79 和 0.55。

表 7 WMT 2019 De-En 和 Cs-En 测试结果

模型	De-En (HTER>0.46)		Cs-En (HTER>0.49)	
	BLEU	PPL	BLEU	PPL
Baseline	17.31	46.91	11.28	55.87
UNMT+QE	<b>17.96</b>	<b>43.74</b>	<b>11.83</b>	<b>52.89</b>

图 6 和图 7 分别为德语—英语和捷克语—英语训练中 Baseline 模型与本文模型(阈值分别为 0.46 和 0.49)的每个 epoch 结果对比。由此可见,本文模型在训练绝大多数的 epoch 时均比基线模型高出 0.25~1.25 个 BLEU 值。我们认为,性能提高的原因是质量估计模型控制了产生的伪平行语料的质量(即保留了质量相对较差的译文),使鉴别器计算得到的损失值较大,因此收敛更慢,从而使得生成对抗网络训练得更加充分。

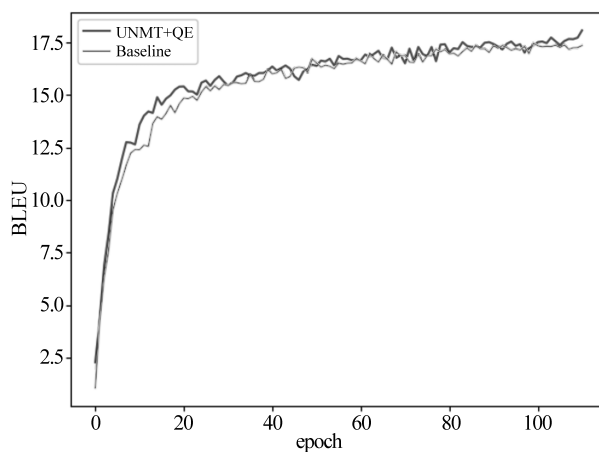


图 6 德语—英语训练中 Baseline 模型与本文模型的每个 epoch 结果对比

由于每次在训练反向翻译时,在①阶段需要将每个 Batch Size 生成的伪平行数据向量转换为伪平行句子对,因此在系统运行的临时存储空间上平均

① 由于训练过程中反向翻译的伪平行数据质量会改变,因此此处统计比率为第 1 个 epoch(留下数据/总数据)。

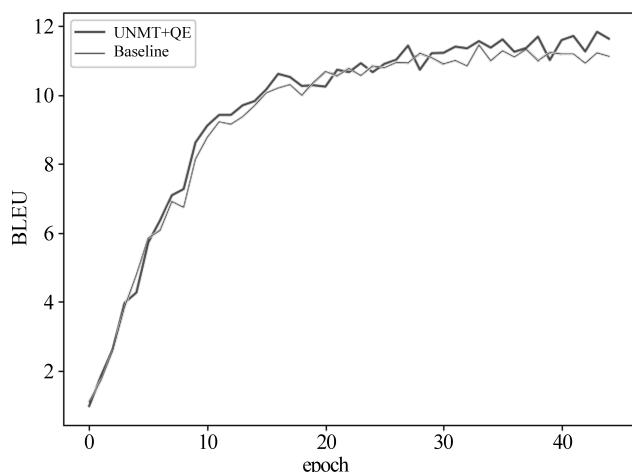


图7 捷克语—英语训练中 Baseline 模型  
与本分模型的每个 epoch 对比

增加 230MB 左右,增加了模型的空间复杂度。与此同时,需要对每个 Batch Size 生成的伪平行句子对评分,增加了模型对伪平行数据评估的时间复杂度。

利用质量估计的无监督神经机器翻译模型,可以有效控制伪平行数据的质量。虽然在训练时不可避免地增加了时间复杂度和空间复杂度,但成功地生成对抗网络提供了更多的训练样本,让生成对抗网络训练得更充分,从而使无监督神经机器的翻译效果得到有效提升。

## 6 总结与展望

无监督神经机器翻译旨在使用更少的单语语料解决机器翻译的双语平行语料不足和语种扩充等问题。本文提出的方法解决了现有无监督神经机器翻译在训练反向翻译时无法控制伪平行数据质量的问题。我们将常规的反向翻译分成两个阶段,在①阶段利用质量估计对其生成的伪平行数据进行评分(HTER),再通过该评分与我们设置的固定阈值比较,筛选出比阈值高的伪平行数据完成反向翻译的②阶段。通过实验发现,我们的方法基于生成对抗网络和 QE,在控制伪平行数据质量的同时丰富了由反向翻译作为生成器的生成样本,使由语言建模作为鉴别收敛得更慢,使得生成对抗网络训练更加充分,从而提升无监督神经机器翻译的翻译效果。本实验在 WMT 2019 共享任务提供的英语-德语和捷克语-英语单语语料上与基线相比,BLEU 值分别提升了 0.79 和 0.55。

本文提出的方法有一定的效果,但是由于使用

了另一个比较大的模型导致模型整体训练速度有所降低,尽管如此,本模型同时也在使用少量的数据的前提下提升了无监督神经机器翻译的效果。未来的研究将主要针对使用更少的单语数据,更加快速地训练完成无监督神经机器翻译,使得其训练效果和常规神经机器翻译同步,甚至取得更好的效果。

## 参考文献

- [1] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, et al. Learning phrase representations using RNN Encoder-Decoder for statistical machine translation [C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, PA: EMNLP, 2014: 1724-1734.
- [2] Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio. Neural machine translation by jointly learning to align and translate [J]. arXiv preprint arXiv: 1409.0473, 2014.
- [3] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is all you need [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Curran Associates Inc, 2017: 6000-6010.
- [4] Mikel Artetxe, Gorka Labaka, Eneko Agirre, et al. Unsupervised neural machine translation [C]//Proceedings of the 6th International Conference on Learning Representations, PA: ICLR, 2018.
- [5] Guillaume Lample, Myle Ott, Alexis Conneau, et al. Phrase-Based and neural unsupervised machine translation [C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, PA: EMNLP, 2018: 5039-5049.
- [6] Alexis Conneau, Guillaume Lample, Marc ' Aurelio Ranzato, et al. Word translation without parallel data [C]//Proceedings of the 6th International Conference on Learning Representations, PA: ICLR, 2018.
- [7] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, et al. Generative Adversarial Networks [J]. Communications of the ACM, 2020, 63(1): 139-144.
- [8] Leng Yichong, Tan Xu, Qin Tao, et al. Unsupervised pivot translation for distant languages [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, PA: ACL, 2019: 175-183.
- [9] Artetxe Mikel, Labaka Gorka, and Eneko Agirre. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings [C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, PA: ACL, 2018: 789-798.
- [10] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, et al. Unsupervised machine translation using



- monolingual corpora only[C]//Proceedings of the 6th International Conference on Learning Representations, PA: ICLR, 2018.
- [11] Zhen Yang, Wei Chen, Feng Wang, et al. Unsupervised neural machine translation with weight sharing [C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, PA: ACL, 2018: 46-55.
- [12] Wang Shuo, Liu Yang, Wang Chao, et al. Improving Back-Translation with Uncertainty-based confidence estimation[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 791-810.
- [13] Kepler, Fábio, Trénous, Jonay, Treviso M, et al. OpenKiwi: An open source framework for quality estimation[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, PA: ACL, 2019: 117-122.
- [14] Zhenhao Li, Lucia Specia. Improving neural machine translation robustness via data augmentation: Beyond back translation[C]//Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019), 2019: 328-336.
- [15] Miguel Graça, Yunsu Kim, Julian Schamper, et al. Generalizing Back-Translation in neural machine translation[C]//Proceedings of the Fourth Conference on Machine Translation, 2019: 45-32.
- [16] Rico Sennrich, Barry Haddow, Alexandra Birch. Neural machine translation of rare words with subword units [C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, PA: ACL, 2015: 1715-1725.
- [17] Tomas Mikolov, Ilya Sutskever, Kai Chen, et al. Distributed representations of words and phrases and their compositionality [C]//Proceedings of the 26th International Conference on Neural Information Processing Systems, Volume 2, 2013: 3111-3119.
- [18] Julia Kreutzer, Shigehiko Schamoni, Stefan Riezler. Quality estimation from scraTCH (QUETCH): Deep learning for word-level translation quality estimation [C]//Proceedings of the 10th Workshop on Statistical Machine Translation, 2015: 316-322.
- [19] André F. T. Martins, Marcin Junczys-Dowmunt, Fabio Kepler, et al. Pushing the limits of translation quality estimation[J]. Transactions of the Association for Computational Linguistics, 2017, 5: 205-218.
- [20] Jiayi Wang, Kai Fan, Bo Li, et al. Alibaba submission for WMT18 quality estimation task [C]//Proceedings of the 3rd Conference on Machine Translation: Shared Task Papers, 2018: 809-815.
- [21] Lucia Specia, Carolina Scarton, Gustavo Henrique Paetzold. Quality estimation for machine translation [J]. Synthesis Lectures on Human Language Technologies, 2018, 11(1): 1-162.
- [22] Philipp Koehn, Hieu Hoang, Alexandra Birch, et al. Moses: Open source toolkit for statistical machine translation [C]//Proceedings of the Association for Computational Linguistics ACL'07, 2007, 9(1):177-180.
- [23] Rico Sennrich, Barry Haddow, Alexandra Birch. Neural machine translation of rare words with subword units [C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, PA: ACL, 2015: 1715-1725.
- [24] Piotr Bojanowski, Edouard Grave, Armand Joulin, et al. Enriching word vectors with subword information [J]. Transactions of the Association for Computational Linguistics, 2017, 5:135-147.
- [25] Kishore Papineni, Salim Roukos, Todd Ward, et al. Bleu: a method for automatic evaluation of machine translation [C]//Proceedings of the 40th Annual Meeting on Association for Computational linguistics. Association for Computational Linguistics, Stroudsburg, PA: ACL, 2002: 311-318.



徐佳(1991—), 硕士研究生, 主要研究领域为自然语言处理、机器翻译。  
E-mail: jxu\_168@163.com



张桂平(1960—), 博士, 教授, 主要研究领域为自然语言处理与机器翻译、知识工程与知识管理。  
E-mail :zgp@ge-soft.com



叶娜(1981—), 博士, 讲师, 主要研究领域为自然语言处理、机器翻译。  
E-mail: yena\_1@126.com