doi:10.6043/j.issn.0438-0479.202011027

一种简单的神经机器翻译的动态数据扩充方法

刘志东,李军辉*,贡正仙
(苏州大学计算机科学与技术学院,江苏 苏州 215006)

摘要:反向翻译作为一种用于神经机器翻译的数据扩充方法,被广泛应用于单语数据的训练.然而,这些方法通常需要大规模源端或目标端单语数据、双语词典等.基于此,提出了一种在不引入外部资源情况下的简单数据扩充方法.该方法在每次加载目标端句子时,按照一定策略对句子中单词进行随机噪声化,以实现原始平行数据目标端的动态数据扩充,从而提高目标端语言模型对句子的表达能力.不同于需要大量单语数据的反向翻译,该方法只使用平行语料.这一策略意味着不需要训练额外的逆向模型.在英德和中英翻译任务上的实验结果表明,该方法使标准 Transformer 系统的双语互译评估(BLEU)值分别提高了 0.69 和 0.66 个百分点.

关键词:神经机器翻译;数据扩充;单词覆盖
中图分类号:TP 391.2 **文献标志码**:A **文章编号**:0438-0479(2021)04-0680-07

和统计机器翻译(statistical machine translation, SMT)[1]相比,神经机器翻译(neural machine translation, NMT)[2-4]仅用一个神经网络就可以实现源语言到目标语言的翻译,省去了搭建特征工程的困扰,显著提高了机器翻译的质量.NMT 模型通常由一个编码器和一个解码器构成,其中编码器将源端句子中的每个单词根据其上下文编码成含上下文信息的隐藏状态;基于其隐藏状态,解码器按从左到右的顺序生成目标端单词.

神经网络本质是一种数据驱动的方法,大量的数据有利于神经网络学习到更合理的参数.特别是对于数据规模受限的小语种来说,通过增加训练数据带来的性能提升往往效果更加明显.因此,如何更多、更好地生成大量平行数据成为许多研究者日益关注的问题.

作为一种增加训练数据的常用方法,数据扩充技术已经被广泛应用于计算机视觉[5]和自然语言处理[6-8]领域.在计算机视觉领域,主要通过对图片进行翻转和随机剪裁操作实现图像数据的扩充.在自然语言处理领域,数据扩充的思路总体上主要分为两大类:1) 句子级别数据扩充,从句子级别生成更多高质量的训练样本,提高模型的泛化能力.2) 单词级别数据扩充,对句子中的单词进行随机交换、丢弃和替换等操作,得到更多带有噪声的数据,提高模型的鲁棒性.

作为一种句子级数据扩充的方法,反向翻译被应用在很多无监督机器翻译模型上,取得了不错的效果.Sennrich 等[7]提出用反向翻译技术构造伪平行句对.该方法首先在已有平行语料的基础上训练一个反向翻译的模型,然后利用这个反向翻译模型来翻译提前收集到的大规模目标端单语语料,获得伪平行句对,最后将伪平行句对和人工标注平行句对合在一起进行模型训练.然而,反向翻译技术需要额外训练一个反向的翻译模型,这无疑会增大运算开销.此外,收集到的单语语料往往存在噪声,对带有噪声的语句进行反向翻译会进一步降低伪平行数据的质量,从而影响翻译模型的性能.He 等[9]发现任何机器翻译任务都有一个对偶任务,能够使得翻译系统自动地从无标注数据中进行学习.原任务和对偶任务能够形成一个闭环,即使没有人类标注者的参与,也能够生成含信息量的反馈信号用以训练翻译模型.

在单词级别数据扩充方面,Iyyer 等[6]在求解一句话的平均词向量前,随机去除文本中的某些单词.

收稿日期:2020-11-16 **录用日期**:2021-04-17
基金项目:国家自然科学基金(61876120,61976148)
* **通信作者**:lijunhui@suda.edu.cn
引文格式:刘志东,李军辉,贡正仙.一种简单的神经机器翻译的动态数据扩充方法[J].厦门大学学报(自然科学版),2021,60(4):680-686.
Citation:LIU Z D,LI J H,GONG Z X. A simple dynamic data expansion method for neural machine translation[J]. J Xiamen Univ Nat Sci,2021,60(4):680-686. (in Chinese)

Artetxe 等[8]设置一个固定长度的窗口,在窗口内随机和相邻的单词进行替换. Fadaee 等[10]利用在大规模单语语料上训练得到语言模型,寻找可以被低频词汇替换的高频词汇,通过这种方法大大提高低频词的出现频率,缓解数据相对稀疏的问题. 相较于直接替换为某个确定的单词,Gao 等[11]提出一种融合多个单词信息的方法. 该方法首先训练一个语言模型,把语言模型预测下一个单词的概率分布作为每个候选单词嵌入表示的权重,然后将线性组合词表中每个单词的嵌入表示作为要替换的单词.

为了解决数据缺乏导致的 NMT 泛化能力不足的问题,同时避免反向翻译技术中单独训练反向模型的开销,受预训练模型 BERT(bidirectional encoder representations from Transformer)[12]启发,本研究提出了一种简单有效且可以对原始平行数据的目标端进行动态扩充的方法. 该方法在每次加载目标端句子时按照一定策略对句子中单词进行随机噪声化,从而提高目标端语言模型对句子的表达能力. 具体地,在加载一批数据时,随机选择目标端句子中的一些单词,并将其进行噪声化,然后约束编码器预测出被覆盖的单词. 如果在整个训练过程中同样的一批数据被加载了 n 次,就等效于将训练数据扩充了 n 倍. 通过约束编码器还原原始语句,可以使自身学到更深层的语言表征能力.

1　背景知识

1.1　NMT

NMT 由编码器和解码器构成,训练的目标是使模型参数在平行语料 $S=\{(x^{(s)},y^{(s)})\}_{s=1}^{|S|}$($|S|$表示平行语料的句子数)上取得最大似然. 近年来,国内外很多研究者提出的模型[3,13-14]均是基于编码器-解码器结构. 由于 Transformer[14]序列到序列模型在很多任务上都能取得较好的性能,本研究选择 Transformer 作为基准模型. 值得注意的是,本方法与模型的内部结构无关,同样适用于其他序列到序列模型.

NMT 中的编码器首先将源句子集合 $x=\{x_1,x_2,\cdots,x_N\}$ 映射成词向量 $\boldsymbol{e}(x)=[\boldsymbol{e}(x_1),\boldsymbol{e}(x_2),\cdots,\boldsymbol{e}(x_N)]$,然后把这 N 个词向量编码成隐藏状态 $\boldsymbol{h}$. 根据隐藏状态 $\boldsymbol{h}$ 和目标端句子 T 个词的集合 $y=\{y_1,y_2,\cdots,y_T\}$,解码器从左到右逐个生成目标端单词的概率,得到 y 的概率:

$$P(y\mid x;\theta_{\mathrm{mt}})=\prod_{i=1}^{T}P(y_i\mid y_{<i},x;\theta_{\mathrm{mt}}). \tag{1}$$

其中:$\theta_{\mathrm{mt}}=\{\theta_{\mathrm{enc}},\theta_{\mathrm{dec}}\}$,为整个模型的参数;$\theta_{\mathrm{enc}}$ 和 θ_{dec} 分别为编码器的解码器的参数;$y_{<i}$ 表示在预测第 i 个目标端单词时已经翻译得到的目标端单词. 模型在训练集 S 上定义的损失函数为

$$L(\theta_{\mathrm{mt}})=\frac{1}{|S|}\sum_{(x,y)\in S}-\log P(y\mid x;\theta_{\mathrm{mt}}). \tag{2}$$

1.2　降噪自编码器

和自编码器相比,降噪自编码器[15]可以学习叠加噪声的原始数据,而其学习到的特征和从未叠加噪声的数据学习到的特征几乎一致,因此降噪自编码器具有更强的鲁棒性;同时降噪自编码器可以避免自编码器简单地保留原始输入数据的信息.

降噪自编码器的训练过程如图 1 所示,给定一个单词序列 $x=\{x_i\}_{i=1}^{n}$,首先引入一个损坏过程 $f(x)$ 得到带有噪声的单词序列 $x'=\{x'_i\}_{i=1}^{n}$. 通过最小化损失 $L(x\mid x')$ 使得模型从带有噪声的 x' 重构干净数据点 x,其中损失函数为

$$L(x\mid x')=-\log P_{\mathrm{dec}}(x\mid f^{\mathrm{en}}(x')). \tag{3}$$

其中,$f^{\mathrm{en}}(x')$表示 x'输入编码器后的输出,$P_{\mathrm{dec}}(x\mid f^{\mathrm{en}}(x'))$表示编码器输入为 x'时,解码器输出 x 的概率.

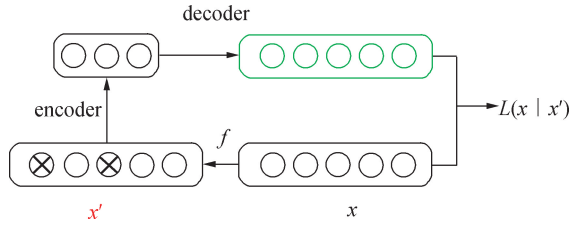

图 1　降噪自编码器的训练过程

Fig. 1　The training process of denoising auto-encoder

2　目标端动态数据扩充方法

对于 NMT,扩充训练数据的方法除了需要大规模的单语语料外,往往还需要训练一个辅助的模型. 而对于资源缺乏的语言来说,引入质量较低的单语语料往往会损害翻译模型的质量. 针对上述问题,本研究提出一种在不引入外部语料的情况下实现数据动态扩充的方法. 该方法首先对输入的目标端语句按照一定策略随机进行噪声化,然后利用编码器将受损的句子还原,以提高编码器对目标单词的预测能力,实现翻译性能的整体提升. 如图 2 所示,和基础的 NMT 系统相比,本方法仅增加了一个随机添加噪声的模块,对于模型的其余部分并没有改动,可以方便应用于其他序列到序列模型.

2.1 构建噪声输入

和降噪自编码器类似,本研究首先构造带有噪声的目标输入.假设给定目标序列输入 $y=\{y_i\}_{i=1}^n$,对每个序列 15% 的单词进行随机覆盖得到 $y'=\{y'_i\}_{i=1}^n$,并保证每句话覆盖的最大单词数不超过 20,对于同一个句子可以同时使用以下 3 种策略得到噪声序列:1) 以 80%的概率用[MASK]替换随机选中的单词;2) 以 10%的概率用词表中的任意一个单词替换选中的单词;3) 以 10%的概率保持选中的单词不变.

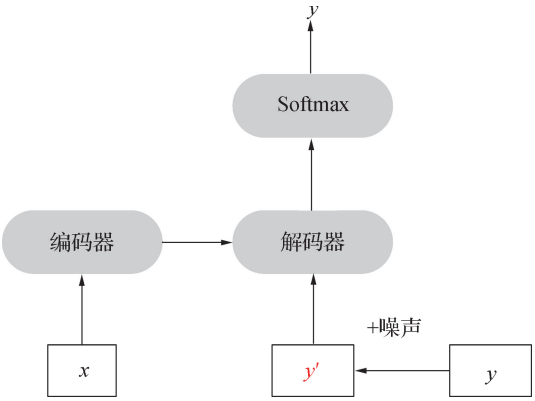

图 2 数据动态扩充的 NMT 模型的整体框架

Fig. 2 The architecture of NMT model with dynamic data augmentation

假设目标端的输入序列为:中国 消费者 信心 支持 中国 经济 增长.在构造带有噪声的输入序列时分别选择第二、第五和最后一个单词(消费者、中国、增长)进行以上 3 种策略的替换,示例如表 1 所示.

表 1 噪声替换策略示例

Tab. 1 Examples of noise replacement strategy

替换方式	示例
80%的概率替换为[MASK]	消费者→[MASK]
10%的概率替换为其他单词	中国→世界
10%的概率保持不变	增长→增长

采用以上 3 种策略后,得到的最终噪声输入为:中国[MASK] 信心 支持 世界 经济 增长.

2.2 重构目标句子

在得到含有噪声的目标序列 $y'=\{y'_i\}_{i=1}^n$ 之后,解码器需要结合编码器的输出将 y' 还原为 y.与 BERT 相比,不同之处在于:本方法不仅预测出被覆盖的单词,而是重构整个目标端序列.

解码端重构目标序列的过程可以认为是最大化条件概率 $P(y\mid \boldsymbol{h},y';\theta_{\text{dec}})$,如式(4)所示.

$$P(y\mid \boldsymbol{h},y';\theta_{\text{dec}})=\prod_{i=1}^{n}P(y_i\mid \boldsymbol{h},y'_{<i};\theta_{\text{dec}}). \quad (4)$$

通过解码器递归地从左至右逐一生成目标词,最终得到完整的译文 $y=\{y_i\}_{i=1}^n$.因此,模型在每个伪平行句对 (x,y') 定义的损失函数为

$$L(\theta_{\text{mt}})=\sum_{i=1}^{n}-\log P(y_i\mid x,y'_{<i};\theta_{\text{mt}}). \quad (5)$$

3 实验结果与分析

本研究对训练数据的源语句和目标语句分别进行静态和动态扩充,使用 multi-bleu.perl(https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl)脚本评测翻译性能.

3.1 语料说明

为了验证本研究提出的动态数据扩充技术,分别在 WMT14 英德(http://www.statmt.org/wmt14/translation-task.html)和 NIST 中英(https://www.nist.gov/srd)双语平行语料上实验.

1) WMT14 英德翻译:训练集共包含 450 万英语到德语平行语料,由 Europarl v7、Common Crawl Corpus 和 News Commentary 数据集构成.此外,实验使用 newstest2013 和 newstest2014 分别作为开发集和测试集.

2) NIST 中英翻译:训练语料使用的是语言数据联盟(Linguistic Data Consortium,LDC)提供的 125 万对中英双语平行语料.实验使用 NIST06 作为开发集,NIST02、NIST03、NIST04、NIST05 和 NIST08 作为测试集.

实验去除两个语言对中训练集长度大于 90 的平行句对,并使用字节对编码(byte pair encoding,BPE)[16]将单词切分成更小的单元.其中,对英德翻译,在英德语料上联合 BPE 处理并设置操作次数为 3;对中英翻译,分别在中文和英文端使用 BPE 处理并设置操作数为 3 和 2.处理后的各数据集样本数如表 2 所示.

3.2 实验设置

本实验使用开源 OpenNMT[17]实现的 Transformer(https://github.com/OpenNMT/OpenNMT-py)和 Bahdanau 等[3]提出的 RNNSearch 模型作为基准模型.在预处理时,共享英德的源端与目标端词表,词表大小为 33 663;中英语料不进行词表共享,得到的中英文词表大小分别为 30 587 和 19 877.

表 2　数据集统计

Tab. 2　Dataset statistics

语料	训练集/10^6	开发集	测试集					
			NIST02	NIST03	NIST04	NIST05	NIST08	newstest2014
WMT14 英德	4.0	3 000						3 003
NIST 中英	1.1	1 664	878	919	1 788	1 082	1 357	

1) Transformer 模型设置. 训练时,英德和中英模型设置相同的参数主要有:编码器与解码器的层数均为 6 层,多头注意力机制均为 8 个头,批处理大小为 4 096,词向量、编码器和解码器的隐藏层维度均为 512,前馈神经网络的维度为 2 048,失活率[18]为0.1. 使用 Glorot 方法初始化模型参数,其他参数均使用默认配置. 表 3 给出了英德和中英实验不同的参数设置.

表 3　参数设置

Tab. 3　Parameter setting

参数	英德	中英
梯度累积数	2	1
学习率	2.0	0.5
热启动训练步数/10^4	0.8	1.6
训练步数/10^6	3.0	2.5

实验模型分别在一块 GTX 1080Ti 显卡上训练. 在网络训练过程中,采用 Adam 算法进行参数更新,其参数 β_1 为 0.9,β_2 为 0.998,ϵ 为 10^{-9},训练过程中每隔 5×10^3 步保存一次模型. 在测试过程中,使用束搜索算法生成最终译文,束搜索的大小设置为 5,长度惩罚因子 α 为 0.6,选择开发集性能最高的模型作为实验最终模型.

2) RNNSearch 模型设置. 英德和中英模型采用相同的实验设置,具体为:编码器和解码器的维度为 1 000,批处理大小为 80,设置源端目标端最长单词序列为 50,失活率[18]为 0.3,训练过程中学习率为 0.000 5,梯度裁剪的大小为 1. 实验模型分别在一块 GTX 1080Ti 显卡上训练 6 轮. 在测试过程中,使用束搜索算法生成最终译文,设置束搜索的大小为 10,在开发集上选择性能最高的模型作为实验的测试模型.

3.3　实验结果及分析

为了验证本研究提出的动态数据扩充技术的有效性,分别在 Transformer 和 RNNSearch 基准模型上进行以下几组实验的对比分析:在 Transformer 模型上对目标端序列静态扩充(tgt-SA),即对同样一批数据即使加载多次也采取同样的覆盖方式;在加载一批数据时对源端句子(src-DA)和目标端句子进行动态扩充(tgt-DA),即对同样一批数据每次加载都采用不同的覆盖方式. 由于本研究主要为验证目标端动态数据扩充方法技术的有效性,所以在 RNNSearch 模型上仅对比 tgt-DA 和 RNNSearch 基准模型的性能.

3.3.1　Transformer 中英翻译

对所提出的方法,本研究在中英数据集上分别进行 3 组实验:静态扩充的方法仅用于目标端(tgt-SA)、动态扩充的方法分别作用于源端和目标端句子(src-DA,tgt-DA). 表 4 给出了中英翻译的实验结果,可以看出:相较于基本的 Transformer 系统,单纯对目标端输入序列静态扩充会带来双语互译评估(BLEU)值的微弱提升(0.25 个百分点),而对目标序列动态扩充的方法可以在 NIST02～NIST08 数据集上取得持续的提升,BLEU 值平均提高 0.66 个百分点. 这验证了动态数据扩充技术的有效性. 然而将动态扩充的方法作用于源端语句时,BLEU 值反而降低了 0.11 个百分点.

由表中数据可以得出以下结论:

1) 在中英翻译实验上:对于目标单词序列,静态扩充方法和动态扩充方法都会提高编码器预测单词的能力;并且动态扩充技术增加了目标句子的多样性,比静态扩充可以带来更高质量的翻译译文.

2) 对源语言动态扩充时,编码器得到的隐藏层状态会丢失部分语义信息,因此不仅不会提升模型的翻译性能反而会降低译文质量.

3.3.2　Transformer 英德翻译

表 5 给出了英德翻译实验结果,可以看出:Transformer 基准系统在测试集上的 BLEU 值为 27.05%,对目标端语句进行静态扩充时,BLEU 值为 26.96%,BLEU 值不仅没有提升反而降低了 0.09 个百分点;然而对于目标端语句进行动态扩充可以获得显著的性能提升,BLEU 值为 27.74%,提高了 0.69 个百分点.

表 4 NIST 数据集上静态扩充和动态扩充的 BLEU 值对比

Tab. 4 Comparison of BLEU values between static and dynamic data augmentation on NIST datasets %

系统	开发集 NIST06	测试集					
		NIST02	NIST03	NIST04	NIST05	NIST08	平均值
Transformer	45.66	46.89	46.04	47.29	46.42	36.21	44.57
Transformer+src-DA	45.47	47.51	45.72	47.43	45.72	35.92	44.46
Transformer+tgt-SA	45.51	47.40	46.05	47.51	46.85	36.31	44.82
Transformer+tgt-DA	46.17	47.71	46.38	48.05	47.19	36.84	45.23

表 5 WMT14 数据集上静态扩充和动态扩充的 BLEU 值

Tab. 5 BLEU values of static and dynamic data augmentation on WMT14 datasets %

系统	开发集	测试集
Transformer	26.12	27.05
Transformer+src-DA	25.96	27.22
Transformer+tgt-SA	26.11	26.96
Transformer+tgt-DA	26.19	27.74

根据表 5 的实验结果,在英德翻译系统上可以得到如下结论:

1) 对目标端语句进行静态数据扩充可能会损害模型的翻译性能. 然而在中英翻译实验上,静态数据扩充能够获得有限提升. 由此可见,静态数据扩充方法带来的翻译性能可能会受到语系的影响.

2) 本研究提出的动态扩充的方法应用于源端语句和目标端语句时都会提升模型的翻译性能,并且应用于目标端时提升的效果更为明显.

3.3.3 RNNSearch 动态数据扩充

为了进一步论证本研究提出方法的有效性,将目标端动态数据扩充技术应用在 RNNSearch[3] 机器翻译模型上. 表 6 给出了 RNNSearch 模型上的中英和英德实验结果,可以看出:动态数据扩充方法在英德翻译任务上提高了 0.51 个百分点,在中英翻译任务上平均提高了 0.41 个百分点. 由此可以得出无论是在当前的主流翻译模型 Transformer 上,还是在 RNNSearch 上,本研究提出的动态数据扩充方法虽然简单,但是都能够带来翻译性能的提高.

表 6 动态数据扩充技术在 RNNSearch 上的 BLEU 值

Tab. 6 BLEU values of dynamic data augmentation on RNNSearch %

翻译任务	系统	开发集	测试集					
			NIST02	NIST03	NIST04	NIST05	NIST08	平均值
英德	RNNSearch	20.74						21.31*
	RNNSearch+tgt-DA	21.19						21.82*
中英	RNNSearch	37.20	39.60	37.88	40.39	37.90	28.60	36.87
	RNNSearch+tgt-DA	37.82	39.79	37.99	41.08	38.73	28.81	37.28

注:* 英德翻译仅有一个测试集 newstest2014,平均值即为测试集的结果,下同.

3.3.4 计算开销对比

本研究提出的目标端语句动态扩充方法不需要改变模型的基本结构,因此并没有引入额外的模型参数,和基线系统相比训练产生的额外开销仅花费在构造目标端噪声输入上;当使用反向翻译技术时,在模型参数和训练数据不变的情况下需要额外训练一个反向的模型,因此参数量和训练时间开销均为基线系统的 2.0 倍,如表 7 所示.

3.3.5 添加噪声分析

由于本研究提出的动态数据扩充方法是对目标端序列进行修改,所以可以视为一种添加噪声的方法. 为了探究动态数据扩充方法和对单词进行噪声化方法的关系,本研究使用 Transformer 翻译模型在英德和中英数据集上做如下对比实验:对目标端句子进

表 7　模型参数及训练速度对比

Tab. 7　Comparison of model parameters and training speed

翻译任务	系统	参数量/10^6	训练时间开销倍数
英德	Transformer	61.4	1.0
	Transformer+反向翻译	122.8	2.0
	Transformer+tgt-DA	61.4	1.4
中英	Transformer	70	1.0
	Transformer+反向翻译	140	2.0
	Transformer+tgt-DA	70	1.4

行动态扩充(tgt-DA)和对目标端句子中每个单词的词嵌入表示添加均值为 0、方差为 0.01 的高斯噪声(tgt-GN).

表 8 给出了在 Transformer 模型上不同添加噪声方法的实验结果,可以看出:对目标端单词的词嵌入表示添加噪声时相较于基准系统可以带来微弱的性能提升,英德和中英翻译任务上 BLEU 值都提高了 0.02 个百分点. 虽然本研究提出的动态数据扩充方法也可以看作是一种动态添加噪声的方法,但是在英德和中英翻译任务上能够带来更多提升,BLEU 值分别提高了 0.69 和 0.66 个百分点.

表 8　tgt-DA 和 tgt-GN 的 BLEU 值对比

Tab. 8　Comparison of BLEU values between tgt-DA and tgt-GN　　%

翻译任务	系统	开发集	测试集					
			NIST02	NIST03	NIST04	NIST05	NIST08	平均值
英德	Transformer	26.12						27.05*
	Transformer+tgt-DA	26.19						27.74*
	Transformer+tgt-GN	26.17						27.07*
中英	Transformer	45.66	46.89	46.04	47.29	46.42	36.21	44.57
	Transformer+tgt-DA	46.17	47.71	46.38	48.05	47.19	36.84	45.23
	Transformer+tgt-GN	45.42	47.43	46.30	47.00	46.15	36.09	44.59

4　结　论

本研究针对 NMT 面临训练语料不足的问题,提出了一种新的数据扩充方法. 该方法在每次加载一批训练数据时,通过不同的覆盖、替换等操作随机修改句子中的单词,得到新的目标句子,然后和源端语句构成新的平行句对,对翻译模型进行训练;通过约束解码器重构原始目标语句,提高模型对抗噪声的能力.

在英德和中英翻译的实验结果表明,本研究提出的动态数据扩充技术可以有效提高 NMT 模型的鲁棒性,相对于基准系统 BLEU 值分别提高了 0.69 和 0.66 个百分点.

然而,该方法也存在一个缺点,即随机将一些单词替换为其他单词可能会损坏句子的语义信息,甚至会完全颠倒句子的语义信息. 因此,在未来的工作中,将考虑加入句子的句法信息,在扩充数据的同时尽可能保持句子的本来信息,进一步提升机器翻译的质量.

参考文献:

[1] KOEHN P, OCH F J, MARCU D. Statistical phrase-based translation[C]// Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Edmonton: ACL, 2003: 48-54.

[2] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[EB/OL]. [2020-11-13]. https://arxiv.org/pdf/1409.3215.pdf.

[3] BAHDANAU D, CHO K H, BENGIO Y. Neural machine translation by jointly learning to align and translate[EB/OL]. [2020-11-13]. https://arxiv.org/pdf/1409.0473.pdf.

[4] KALCHBRENNER N, BLUNSOM P. Recurrent continuous translation models [C] // Conference on Empirical Methods in Natural Language Processing. Seattle: ACL, 2013: 1700-1709.

[5] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[6] IYYER M, MANJUNATHA V, BOYD-GRABER J, et al. Deep unordered composition rivals syntactic methods for text classification[C] // Annual Meeting of the Association for Computational Linguistics and International Joint Conference on Natural Language Processing. Beijing: ACL, 2015: 1681-1691.

[7] SENNRICH R, HADDOW B, BIRCH A. Improving neural machine translation models with monolingual data[C] // Annual Meeting of the Associa-tion for Computational Linguistics. Berlin: ACL, 2016: 86-96.

[8] ARTETXE M, LABAKA G, AGIRRE E, et al. Unsupervised neural machine translation[EB/OL]. [2020-11-13]. https: // arxiv. org/pdf/1710. 11041. pdf.

[9] HE D, XIA Y C, QIN T, et al. Dual learning for machine translation[EB/OL]. [2020-11-13]. https: // arxiv. org/pdf/1611. 00179. pdf.

[10] FADAEE M, BISAZZA A, MONZ C. Data augmentation for low-resource neural machine translation[C] // Annual Meeting of the Association for Computational Linguistics. Vancouver: ACL, 2017: 567-573.

[11] GAO F, ZHU J H, WU L J, et al. Soft contextual data augmentation for neural machine translation [C] // Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019: 5539-5544.

[12] GEHRING J, AULI M, GRANGIER D, et al. Convolutional sequence to sequence learning[EB/OL]. [2020-11-13]. https: // arxiv. org/pdf/1705. 03122v3. pdf.

[13] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL]. [2020-11-13]. https: // arxiv. org/pdf/1810. 04805v2. pdf.

[14] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. [2020-11-13]. https: // arxiv. org/pdf/1706. 03762v5. pdf.

[15] VINCENT P, LAROCHELLE H, BENGIO Y, et al. Extracting and composing robust features with denoising autoencoders[C] // International Conference on Machine Learning. Helsinki: ACM, 2008: 1096-1103.

[16] SENNRICH R, HADDOW B, BIRCH A. Neural machine translation of rare words with subword units [C] // Annual Meeting of the Association for Computational Linguistics. Berlin: ACL, 2016: 1715-1725.

[17] KLEIN G, KIM Y, DENG Y T, et al. OpenNMT: Open-source toolkit for neural machine translation[EB/OL]. [2020-11-13]. https: // arxiv. org/pdf/1701. 02810. pdf.

[18] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

A simple dynamic data expansion method for neural machine translation

LIU Zhidong, LI Junhui*, GONG Zhengxian

(School of Computer Science and Technology, Soochow University, Suzhou 215006, China)

Abstract: As a type of data expansion method for neural machine translation, back-translation has been widely used to train with monolingual data. However, these methods often require large-scale source side or target side monolingual datasets, bilingual dictionaries and so on. This paper proposes a simple data expansion method without introducing external resources. Each time the target sentence is loaded, the words in the sentence are randomly noised according to a certain strategy to realize the target data dynamic expansion of the original parallel data, so as to improve the expression ability of the target language model to the sentence. Specifically, different from back-translation which requires huge amount of monolingual data, this method only use parallel corpuses. This strategy means that we do not need to train an additional reverse model. Experimental results regarding English-German and Chinese-English translation tasks show that our approach significantly improves the bilingual evaluation understudy (BLEU) values of a standard Transformer system by 0. 69 and 0. 66 percentage points respectively.

Keywords: neural machine translation; data expansion; word masking