doi:10.6043/j.issn.0438-0479.201811015

基于半监督学习的小语种机器翻译算法

陆雯洁[1],谭儒昕[1],刘功申[1,2*],孙环荣[2]

(1.上海交通大学电子信息与电气工程学院,上海 200240;
2.上海交通大学-上海嵩恒信息内容分析技术联合实验室,上海 200240)

摘要:近年来,基于神经网络的机器翻译取得了快速发展,然而由于它需要大规模的平行语料库,所以对于资源稀缺的小语种的翻译往往显得效果不佳.在分析编码-解码框架和注意力机制的基础上,基于对偶学习的思想,提出了一种面向小语种翻译的半监督神经网络模型.该模型利用较大的单语语料库与少量平行语料库来实现小语种翻译.实验结果表明,当平行语料资源不足以训练一个普通神经网络模型时,使用半监督网络模型能够取得较好的结果,但所采用的半监督学习模型对单语语料库的数量要求非常高,要达到一定数量级才能达到良好效果.

关键词:半监督学习;小语种;机器翻译

中图分类号:TP 391.2 **文献标志码**:A **文章编号**:0438-0479(2019)02-0200-09

随着“一带一路”倡议的深化,亟需通过网络公开信息自动感知“一带一路”沿线国家和地区的各类情报信息,为国家和地区间合作提供决策支持的必要信息来源.当前的最大困难是沿线国家和地区的语言多属于小语种(例如乌尔都语、越南语、哈萨克语、藏语、维吾尔语等),完全靠人工掌握这些语言是不可能的.因此,需要借助计算机来实现便捷而高效的翻译,即机器翻译.

最早的机器翻译需要借助人工知识进行单词转换,出现了基于规则和基于例子的翻译方法.20 世纪 90 年代末,Brown 等[1]提出了统计机器翻译(statistical machine translation,SMT),即根据机器对文本的统计结果进行翻译.随着谷歌提出神经机器翻译(neural machine translation,NMT)后,SMT 取得了重大突破[2].NMT 的核心结构是“编码器-解码器”,最早由 Forcada 等[3]提出.之后,Bahdanau 等[4]引入注意力机制,弥补了基础 NMT 的缺点.不同于传统的 SMT 使用 N 元语法(N-gram)模型每次几个字地评估目标语言的句子流畅程度,NMT 将编码器-解码器结构作为一个整体进行训练,每次评估整个句子的流畅度,且可以更多地捕获构成语言的复杂特征之间的相互依赖关系.然而 NMT 需要大量数据才可以获得流畅的翻译效果,因此缺乏语法、词性标注、语料库等[5]语言资源的低资源语言面临着巨大的挑战,如小语种语言.

目前为止,可以说 NMT 超越传统的 SMT 成为了主流的翻译方法,尤其是在有大量语料库资源时,NMT 可以获得十分不错的结果.然而 NMT 仍然有许多需要提升的地方,例如谷歌翻译可以说达到了新的里程碑,但是在中英翻译的测试中仍然存在着一些差错.可想而知,对于那些资源稀缺的小语种,翻译的难度更大.因而如何针对资源稀缺的小语种进行翻译引起了学界的关注[6].

近年来有许多针对低资源小语种的翻译方法被提出,主要可以分为 4 个研究方向:对偶学习、枢轴语言、迁移学习以及多任务学习.其中仅对偶学习可利用少量平行语料与大量单语语料实现无监督的学习,其他 3 种方法均需要大量的平行语料.故本研究在分析编码-解码框架和注意力机制的基础上,基于对偶学习的思想,提出了一种面向小语种翻译的半监督神经网络模型.

收稿日期:2018-11-10 **录用日期**:2019-01-08

基金项目:国家自然科学基金(61772337,61472248)

* **通信作者**:lgshen@sjtu.edu.cn

引文格式:陆雯洁,谭儒昕,刘功申,等.基于半监督学习的小语种机器翻译算法[J].厦门大学学报(自然科学版),2019,58(2):200-208.

Citation:LU W J,TAN R X,LIU G S,et al. Machine translation algorithm of low-resource languages based on semi-supervised learning[J]. J Xiamen Univ Nat Sci,2019,58(2):200-208.(in Chinese)

1　通用的 NMT 模型

1.1　编码-解码框架

NMT 的基本架构是编码-解码框架,它用于解决序列到序列(seq2seq)问题,即根据一个输入序列 $X=(x_1,x_2,\cdots,x_m)$ 生成一个输出序列 $Y=(y_1,y_2,\cdots,y_n)$,其框架如图 1 所示.

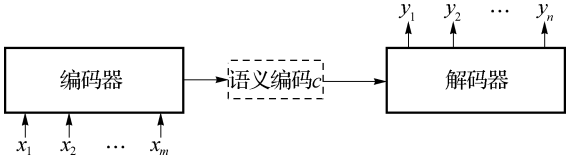


图 1　自然语言处理领域的编码-解码框架示意图

Fig. 1　Schematic diagram of encoder-decoder framework of natural language processing domain

编码器接受输入序列 X 并进行编码,通过非线性变化 f 将输入序列转换为一个固定长度的语义编码 c:

$$c=f(x_1,x_2,\cdots,x_m). \tag{1}$$

解码器根据语义编码 c 和历史输出 $y_1,y_2,\cdots,y_{i-1}$,通过非线性变化 g 来产生 i 时刻的输出 y_i:

$$y_i=g(c,y_1,y_2,\cdots,y_{i-1}). \tag{2}$$

该框架应用广泛,但存在一些弊端,其中最主要的问题是当输入序列变长时,所有信息都压缩到一个固定长度的向量中,必然会丢失许多重要的信息[4].尤其最基础的编码器和解码器由循环神经网络(recurrent neural network,RNN)组成,结构简单直观,但是存在着梯度消失的问题.为了解决这个问题,长短时记忆(long-short term memory,LSTM)方法被提出,其在 RNN 隐藏层的结构上增加了神经元结构,用于确定哪些信息被记住以及哪些信息被遗忘.在 LSTM 的基础上,Cho 等[7]提出了门控循环单元(gated recurrent unit,GRU),简化了 LSTM 的复杂结构,同时保留了长距离时的记忆能力.

1.2　注意力机制

为了解决长距离依赖的问题,注意力机制被提出.加入注意力机制的模型在产生输出序列时会参考输入序列中的一些部分,具体如图 2 所示.此时,编码器接受输入序列 X,通过非线性变化 f_i 将其转换为一个固定长度的语义编码 c_i,体现不同输入序列中词的注意力分配情况.

$$c_i=f_i(x_1,x_2,\cdots,x_m). \tag{3}$$

解码器根据语义编码 c_i 和历史输出 $y_1,y_2,\cdots,y_{i-1}$,通过非线性变化 g 来产生 i 时刻的输出 y_i:

$$y_i=g(c_i,y_1,y_2,\cdots,y_{i-1}). \tag{4}$$

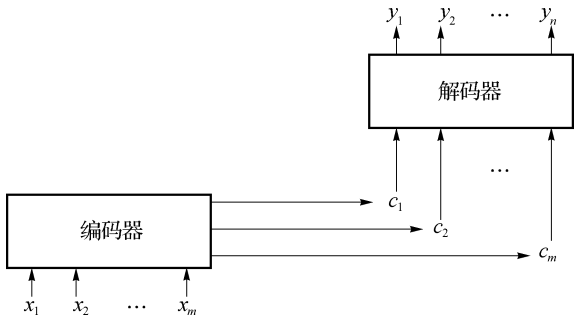


图 2　加入注意力机制的编码-解码框架示意图

Fig. 2　Schematic diagram of encoder-decoder framework with attention mechanism

2　半监督的 NMT 模型

2.1　对偶学习

对偶学习包含原始任务和对偶任务,用一个代理来表示原始任务模型,用另一个代理来表示对偶任务模型,两者通过一个加强学习过程给予对方反馈信号,进行相互学习[8],如图 3 所示.

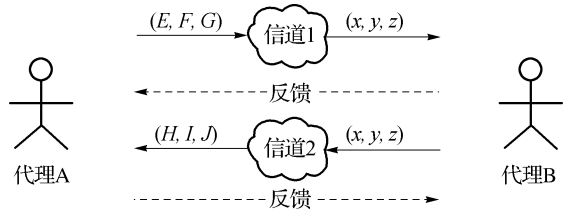


图 3　机器翻译中对偶学习的过程模拟

Fig. 3　Simulation of dual learning in machine translation

假设代理 A 只会语言 L1,代理 B 只会语言 L2,整个交流过程可以分为 4 步:

1) 代理 A 发送一则消息 (E,F,G),通过有噪声的信道 1 给代理 B,信道 1 可使用翻译模型将语言 L1 转换成语言 L2,经过信道 1 后的消息变为 (x,y,z).

2) 代理 B 收到消息 (x,y,z),虽然 B 不能确定信道 1 的翻译结果是否正确,但是 B 可以给予 A 反馈,告知 A 这个消息在语言 L2 中是否是个自然的句子.

3) 代理 B 将消息 (x,y,z) 通过有噪声的信道 2 将语言 L2 转换成 L1,发给代理 A 消息 (H,I,J).

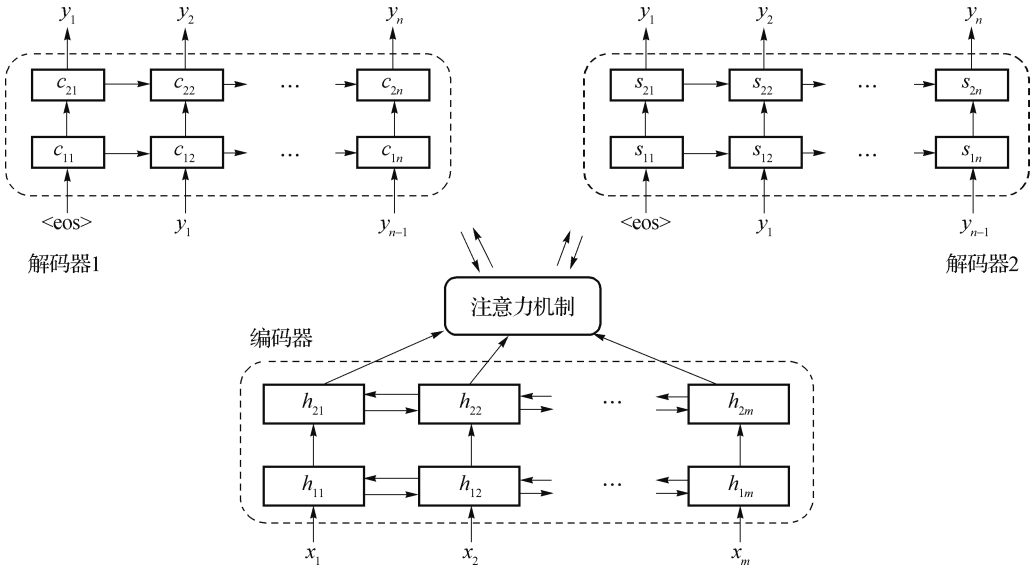
4) 代理 A 收到消息 (H,I,J),给予 B 反馈,告知 B 这个消息在语言 L1 中是否是个自然的句子.

通过两个反馈,代理 A 和 B 可以得知两个信道和两个翻译模型的效果,继而进行改善.整个交流过程

可以重复多次.

2.2 基于对偶学习的半监督 NMT 模型

为了利用对偶学习的思想实现半监督机器翻译,采用图 4 所示的模型,由一个共享编码器和两个解码器组成,两个解码器分别为源语言和目标语言的解码器.编码器端为双层双向循环神经网络(BiGRU),解码器端为双层循环神经网络(UniGRU).在编码器端,使用预训练的词向量,接受输入序列并生成与语言无关的表征.而解码器端的词向量会随着训练不断更新,通过两个解码器进行训练和翻译.



h_{ij} 表示第 i 层第 j 个隐藏层状态;c_{ij} 和 s_{ij} 分别表示相应解码器的第 i 层第 j 个隐藏层状态,<eos>表示句尾标识.

图 4 半监督的 NMT 模型

Fig. 4 Semi-supervised neural machine translation model

由于机器翻译的对称性,使用两个解码器可以应用回译[9]的方法进行训练,实现数据增广,形成更多的数据帮助训练.即,通过将目标语言端的单语语料库用已有的翻译模型翻译成源语言端的句子,形成伪平行语句对;再将这些新的语料资源与已有的平行语料库一起,形成一个新的训练集.该方法基于这样的事实[9]:对于翻译而言,目标语言端的句子流畅性对于翻译有很大的影响,而源语言端句子有少量用词不准确对翻译结果的影响不大.

由于有两个解码器,输入序列 X 为语言 L1 时,序列 X 经过共享编码器成为与语言无关的表征后有两种解码情况:1) 通过解码器 1 产生输出序列 Y,Y 应该是与 X 完全相同的句子;2) 通过解码器 2 产生输出序列 Y',Y' 是 X 的译文,应该是语法正确且与 X 语义相同的句子.

为了模型在情况 1)下更好地学习到语言的内部结构,而不是简单地进行词的复制操作,在输入序列 X 中引入随机噪声,即对于包含 N 个元素的序列进行 $N/2$ 次随机互换操作.这样预训练的词嵌入中的词语顺序是错误的,故无法根据语序信息来区分词汇,需要通过学习来恢复输入序列的正确词序,从而避免在翻译时模型进行简单的逐词翻译.

假设语言 L1 的输入序列为 X_{L1},语言 L2 的输入序列为 X_{L2},译文为语言 L1 的输出序列为 Y_{L1},译文为语言 L2 的输出序列为 Y_{L2},平行语料库为序列对 (P_{L1},P_{L2}).模型的整个训练过程可以分为 5 种情况如图 5 所示.

因此,共有 5 组输入输出序列对,分别为 (X_{L1},Y_{L1})、(X_{L2},Y_{L2})、(X_{L1},Y_{L1}'')、(X_{L2},Y_{L2}'') 以及 (P_{L1},P_{L2}'),将其统一表示为 $D=(X,Y)$,采用负对数似然损失函数来构造目标函数:

$$L(\theta;D)=-\sum_{i=1}^{N}\log P(Y\mid X,\theta). \tag{5}$$

此外,使用困惑度(perplexity)对训练过程、效率以及收敛性进行评估.困惑度被定义为

$$\mathrm{ppl}(X,Y)=\exp\left(\frac{-\sum_{i=1}^{|Y|}\log P(y_i\mid y_{i-1},\cdots,y_1,X)}{|Y|}\right). \tag{6}$$

当困惑度下降并且数值变得较小时,意味着模型对于训练数据来说训练得很好.对于验证集,可以通过困惑度来确定模型对于未知数据的表现.

为了避免神经网络过拟合,同时为了提高泛化能力,可以使用一些正则化方法.本研究在训练时用丢

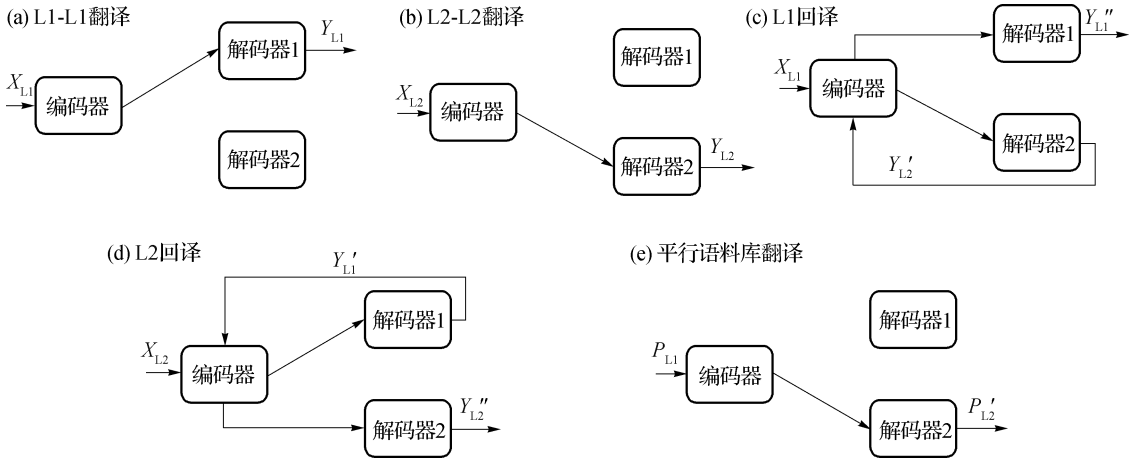


图 5　半监督模型的 5 个训练过程

Fig. 5　Five training processes of semi-supervised model

包的方法以概率 p 来丢弃一个批次中的一些神经元. 丢包率(dropout)值设置为 0 时表示不丢弃,在每一层输出处以及注意力层的输出处都应用了dropout.

值得注意的是,由于模型中使用的是共享编码器,需要对两个输入序列预训练的词向量进行映射. 通常情况下,词向量的映射依赖于一个规模为几千对单词的双语词典,通过线性变化使两个词向量被映射到另一个空间,映射后得到的结果能够使得两个词向量在词典中的距离最小[10]. 考虑到小语种语料资源稀缺的情况,可以采用数字对齐的方式进行词向量映射[11],如图 6 所示. 具体步骤如下:

1) 词向量映射

假设语言 L1 和 L2 的词嵌入矩阵分别为 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$, $\boldsymbol{X}_{i*}$ 为源语言的第 i 个词的词向量, $\boldsymbol{Y}_{j*}$ 为目标语言的第 j 个词的词向量;词典 $\boldsymbol{D}$ 为一个二进制的矩阵,当源语言第 i 个词与目标语言的第 j 个词对齐时, $D_{ij}=1$. 词映射的目标是找到一个映射矩阵 $\boldsymbol{W}^*$,使映射后的 $\boldsymbol{X}_{i*}$ 和 $\boldsymbol{Y}_{j*}$ 的欧几里得距离最近,即:

$$\boldsymbol{W}^* = \underset{\boldsymbol{W}}{\operatorname{argmin}} \sum_i \sum_j D_{ij} \parallel \boldsymbol{X}_{i*}\boldsymbol{W} - \boldsymbol{Y}_{j*} \parallel^2. \quad (7)$$

对矩阵 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 进行标准化和中心化,并将 $\boldsymbol{W}$ 设置为正交矩阵后,上述求解欧几里得距离的问题相当于最大化点积:

$$\boldsymbol{W}^* = \underset{\boldsymbol{W}}{\operatorname{argmax}} \operatorname{Tr}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{D}^{\mathrm{T}}), \quad (8)$$

其中,Tr 表示矩阵的迹运算. 可以求解得到最优解为 $\boldsymbol{W}^* = \boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}$ ($\boldsymbol{U}$、$\boldsymbol{V}$ 表示两个正交矩阵),经过奇异值分解, $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{Y} = \boldsymbol{U}\sum\boldsymbol{V}^{\mathrm{T}}$. 鉴于矩阵 $\boldsymbol{D}$ 是稀疏的,可以在线性时间内得到解.

2) 词典自学习

映射后的源语言词的词向量与目标语言词的词向量在同一个空间. 根据最近邻检索的方法,为每个源语言词分配一个距离最近的目标语言词,将对齐的词对添加到词典中,再次进行迭代,直到收敛.

以图 6 为例,一开始词典中对齐的词对为("1-a","2-b"),根据词典 L1 进行了一次映射,使得映射后的"1"与"a"以及"2"与"b"之间的欧几里得距离最近. 然后在映射后的空间里,为其他词寻找距离最近的对应词,可以发现"3"与"c"的距离较近,因此把它也加入词典中. 此时,尽管词典中包含了所有的词对,但并不是

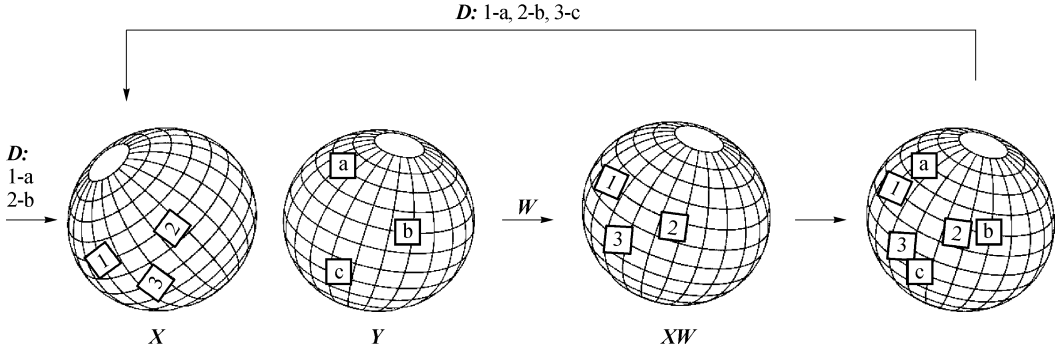


图 6　使用数字对齐进行词映射过程的示意图(修改自文献[11])

Fig. 6　Schematic diagram of mapping process using numeral alignment (modified from reference[11])

最佳的结果.将更新后的词典("1-a","2-b","3-c")作为新的参考词典,重新进行欧几里得距离的计算,将会得到新的映射矩阵 $\boldsymbol{W}^*$,从而获得新的对齐结果.

训练完后用集束搜索(beam search)进行翻译,束的大小需权衡翻译的时间以及搜索的准确性来确定.

3 实验与分析

3.1 实验设置

本研究使用的语料库包括约 7 万句藏语-汉语平行语料库、约 150 万句藏语单语语料库以及约 2 380 万句汉语单语语料库,其中单语语料库的来源是各类新闻网站,例如人民网、西藏日报、西藏法制报等,平行语料库涵盖了新闻、日常对话等不同领域,使用机器双语互译评估(BLEU)[12]指标进行评估.

首先,需要对语料库进行基本的预处理,使用 Moses 内置工具对语料库进行标准化,随后分别使用结巴(Jieba) (https://github.com/fxsjy/jieba)和 tiplas[13]对汉语和藏语进行分词.对分词后的文本,使用字节对编码(byte pair encoding,BPE)进行编码.BPE 对子词进行操作,因而模型可以基于所有子词单元来翻译和产生在训练过程中没有遇到过的新词,有效地缓解了 NMT 常见的集外词(out of vocabulary,OOV)和罕见词问题[10,14].对于藏语与汉语的单语语料库,都使用 4 万个操作符.为了缩短训练时间,所有大于 50 个词的句子都被删除.由于模型的共享编码器部分使用的是固定的跨语言词嵌入,使用 word2vec 对预处理完成后的单语语料库进行词嵌入操作,词向量的维度设置为 150,训练窗口大小设置为 10,negative 值设置为 10,hs 设置为 1,采样阈值设置为 1×10^{-5},迭代次数设置为 1.

完成词嵌入后,将藏语和汉语的词向量映射到同一个空间,使用上文提到的数字对齐的方法,经过 150 次自学习的迭代,获得映射后的藏语与汉语的词向量,此时两种语言间已经建立了一些相关的表征.

完成词向量的映射后,模型按照 5 种情况进行训练.值得注意的是,训练中的情况(c)虽然也是源语言到目标语言的训练过程,但是与使用平行语料库的情况有所不同.在使用单语语料库的情况下,源语言句子 X_{L1} 通过翻译模型得到翻译结果 Y_{L2}',但是没有正确的译文供 Y_{L2}' 参考,无法得知翻译结果的正确性,因而也就不能评判模型参数的好坏.因此,通过回译的方法,将翻译结果 Y_{L2}' 重新输入共享编码器,并通过源语言的解码器翻译成为源语言的句子 Y_{L1}''.理论上 X_{L1} 与 Y_{L1}'' 应该是相同的句子,然而由于翻译模型的不足之处,X_{L1} 可能会得到不同于 Y_{L1}'' 的翻译结果,通过这样的方式就可以对翻译模型进行调整.然而,当使用小规模的平行语料库时,将源语言句子 P_{L1} 通过翻译模型得到翻译结果 P_{L2}',此时拥有正确的译文供 P_{L2}' 参考,因此可以直接对模型进行调整,并且此时只有一个源语言-目标语言方向的模型参与了翻译,可以更加准确地对这个模型进行调整.

考虑到实验条件以及训练的成本,将模型的 dropout 值设置为 0.3,批处理大小设置为 50,并设置 30 万次迭代,迭代所使用的训练数据规模涵盖了汉语单语语料库.只训练了一轮,对于汉语来说,如果条件允许有更多次的迭代与训练,可以获得更好的效果.

图 7 展示了不同训练集情况下困惑度与迭代次数的关系.从图 7(a)的源语言回译情况可以看到,刚开始模型的困惑度很大,经过 1 万次迭代后模型困惑度下降至 184.06,经过 30 万次迭代后模型的困惑度为 89.06.图 7(b)的目标语言回译情况中,模型一开始的困惑度同样很大,经过 1 万次迭代后下降至 101.51,经过 30 万次迭代后模型的困惑度为 23.58.

类似于源语言和目标语言的回译训练,源语言-源语言以及目标语言-目标语言的翻译过程中,随着迭代的进行,困惑度急剧下降.为了更直观地观察它们的变化情况,图 7(c)为从第 1 万次迭代后源语言-源语言以及目标语言-目标语言模型的困惑度与迭代次数的关系.可以看到,在第 1 万次迭代后,二者的困惑度分别为 4.76 和 13.89,比起前面两种回译情况的结果要低得多.这是由于此时训练的是由共享编码器接受输入序列 X,通过源语言解码器输出序列 Y,模型需要考察的是这两个序列的相似性并做出相应的调整.在经过 30 万次迭代后,源语言-源语言以及目标语言-目标语言模型的困惑度分别为 1.41 和 2.16.

图 7(d)为使用平行语料库进行源语言-目标语言翻译时模型的困惑度与迭代次数的关系,可以看出,困惑度随着迭代次数的增加而下降,在经过 30 万次迭代后,困惑度为 3.51,可见模型在平行语料库上获得了较好的训练结果.训练完成后,模型在测试集上得到的 BLEU 值为 12.69%.

3.2 实验结果的分析和讨论

为了进一步比较该半监督模型的翻译结果,使用相同的规模为 7 万的平行语料库在 SMT 和 NMT 的基线系统上进行实验,分别选择基于统计的翻译系统 Moses 以及基于神经网络的翻译系统 OpenNMT,实验结果见表 1.

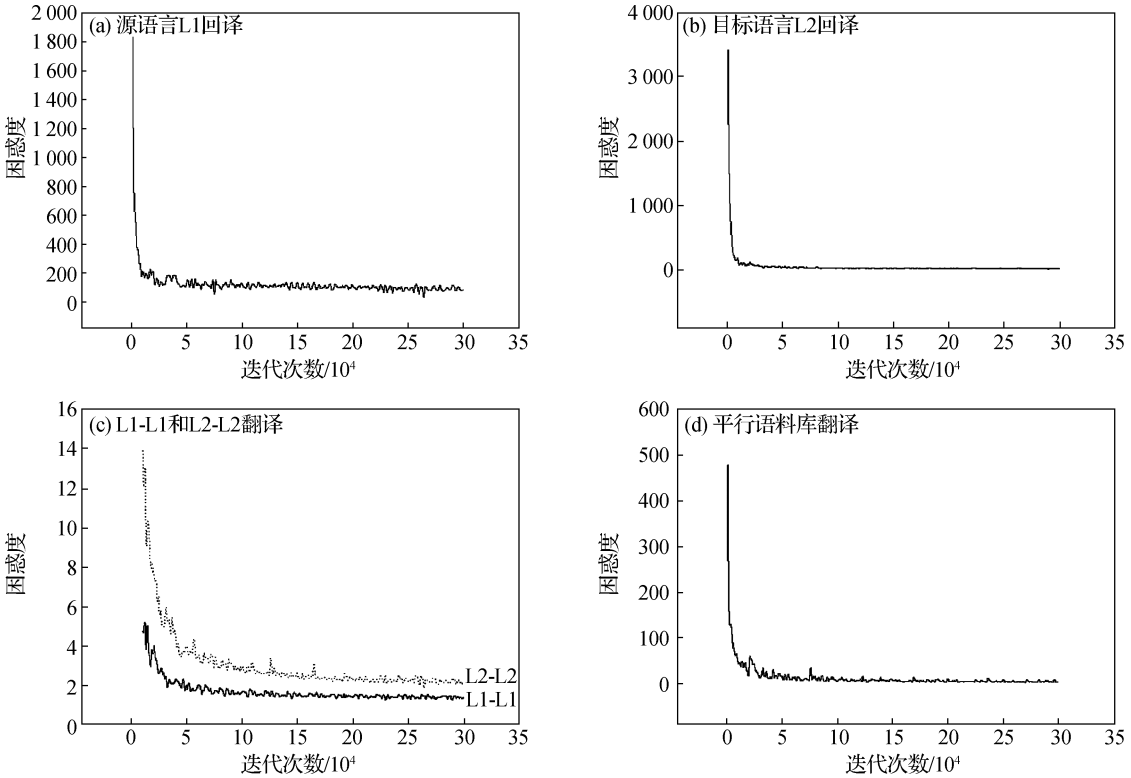


图 7　不同训练集情况下的困惑度-迭代次数关系

Fig. 7　Perplexity-iteration relationship of different training sets

表 1　3 种模型的实验结果

Tab. 1　Experimental results of three models

模型	BLEU/%
基线 SMT	12.13
基线 NMT	11.62
半监督	12.69

从实验结果看,基线 NMT 模型的 BLEU 值低于 SMT 模型,这是因为训练集的规模较小,NMT 模型难以捕获高质量的语法、语义等信息,这也印证了小规模的平行语料资源是不足以训练一个良好的神经网络模型的[15].另一方面,小语种平行语料资源匮乏,如何在仅有少量资源的情况下获得较好的效果是一个有重要意义的问题,这也是本文仅使用较小的平行语料库规模的出发点.可以看到,本文的半监督模型相比于基线 NMT 在 BLEU 值上提升了 1.07 个百分点,比基线 SMT 在 BLEU 值上提升了 0.56 个百分点,说明借助于单语语料库获得了更好的翻译效果.

在已有的基于对偶学习思想的机器翻译研究中,效果较好的是 He 等[16]的研究,将对偶学习方法与基线 NMT[4]和基于 NMT 的伪平行语料的方法[9]进行比较,使用的语料库均为近 1 200 万句英语-法语双语语料,结果如表 2 所示.实验结果表明,对偶学习使用 10%的平行数据就能比拟 NMT 使用全部数据所达到的翻译准确度,这对于只有少量平行数据的翻译场景有借鉴意义.然而对于小语种翻译而言,由于语料资源匮乏,无法获取与英语-法语双语语料库规模相当的语料库,这也是本文实验结果与其存在较大差异的原因.

表 2　英语-法语对偶学习翻译的实验结果[16]

Tab. 2　Experimental results of dual learning English-French translation results[16]

模型	BLEU/%			
	英→法(L)	法→英(L)	英→法(S)	法→英(S)
NMT	29.92	27.49	25.32	22.27
pseudo-NMT	30.40	27.66	25.63	23.24
dual-NMT	32.06	29.78	28.73	27.50

注:L 表示使用了全部平行数据,S 表示使用了 10%平行数据.

分析近几年关于小语种机器翻译的研究,一些效果较好的藏语-汉语翻译研究如表 3 所示,其中李亚超

等[17]使用迁移学习的方法,利用规模为 10 万的藏语-汉语平行语料库以及规模为 125 万的英语-汉语平行语料库,首先使用英语-汉语的语料训练一个基线 NMT 模型,再使用训练得到的模型参数直接对藏语-汉语的基线 NMT 模型进行初始化,实验结果显示该方法相比 SMT 方法的 BLEU 值提高了 3 个百分点.该方法的特点是简单并且具有语言无关性,但是问题在于直接迁移预训练得到的参数对于捕获特定小语种的语言学信息是存在瓶颈的,而本文中提出的方法可以从两种语言的单语语料库中学习到额外的语言学信息,从而有利于获得更好的翻译结果.

此外,位素东等[18]采用人工统计的方法扩充统计词翻译概率表,以此缓解由于平行语料库规模较小导致的数据稀疏问题,实验结果显示该方法相比基线 SMT 方法的 BLEU 值提高了 0.8 个百分点.该方法虽然简单直观,但是人工扩展词典十分困难,并且依赖于人为参与,似乎与机器翻译初衷背道而驰.

表 3 藏语-汉语机器翻译研究的实验结果

Tab.3 Experimental results of Tibetan-Chinese machine translation

模型	训练集大小	BLEU/%	BLEU5-SBP/%	来源
基线 NMT	7 万	11.62	21.1	本文
半监督	7 万+单语语料	12.69	22.1	
基线 NMT	10 万	30.39		文献[17]
迁移学习	10 万+125 万英汉	36.52		
基线 SMT	10 万		28.7	文献[18]
扩展词典	10 万		29.5	

一些效果较好的维吾尔语-汉语翻译研究如表 4 所示,罗延根等[19]使用了无监督的归一化方法,将一些非正规词替换为能正确翻译的词,在口语文本的机器翻译中相比基线系统提高了 0.73 个百分点.潘一荣等[20]使用汉维调序表重构模型学习维吾尔语语义内容以及调序方向,提高解码时调序信息的准确度,相比基线系统提高了 0.18～0.42 个百分点.哈里旦木·阿布都克里木等[21]使用 5 万的维吾尔语-汉语平行语料库在 6 种 NMT 系统上进行了实验,结果显示 6 种 NMT 系统效果均不如 SMT 模型,说明对低资源、形态丰富语言的神经翻译技术还有待提高.由此可见,通过加入语义特性可以提高翻译性能,但是一般来说训练集的规模对于翻译的效果影响较大.

表 4 维吾尔语-汉语机器翻译研究的实验结果

Tab.4 Experimental results of Uyghur-Chinese machine translation

模型	训练集大小	BLEU/%	来源
无监督	20 万	18.27～18.64	文献[19]
调序表重构	11 万	29.49～29.91	文献[20]
NMT	5 万	13.10～18.00	文献[21]

此外,表 5 是一些无监督机器翻译研究的实验结果,其中 Aretxe 等[22]使用规模为 3 600 万的来源于机器翻译 WMT14 数据集的英语-法语平行语料库,通过回译等方法相比基线系统提高了 5.6～8.1 个百分点.Lample 等[23]使用同样规模的 WMT14 数据集的英语-法语平行语料库、WMT16 数据集的英语-德语平行语料库以及包含 3 万图像的 Multi30k-Task1 数据集,通过加入去噪自编码器相比基线系统提高了 2.58～8.77 个百分点.Yang 等[24]提出权重共享的无监督机器翻译模型,相比基线系统提高了 5.01～13.37个百分点.相比半监督机器翻译研究,无监督机器翻译优势在于大规模单语语料库比较容易获得,因此,在小语种翻译平行语料库极其匮乏的情况下,它可以取得较为合理的翻译结果.但是,仅仅通过单语语料库学习两种语言的映射关系有很大的难度,难以达到有监督学习时的效果.

表 5 无监督机器翻译研究的实验结果

Tab.5 Experimental results of unsupervised machine translation

BLEU/%				来源
英→法	法→英	英→德	德→英	
15.1	15.6			文献[22]
15.0	14.3	9.6	13.3	文献[23]
17.0	15.6	10.9	14.6	文献[24]

因此,可以考虑借助于大量的单语语料库,使用小规模平行语料作为辅助训练.由经验表明,语料库的规模以及质量对于神经网络模型有重大的影响.对于机器而言,每个领域都需要不同的语料进行训练,例如一个由体育类语料训练出的模型,在翻译医学类内容时,效果必然是比较糟糕的.同时,不同领域的语料库在用词与表达上也存在着差异,例如新闻领域的

语料表达方式较为正式,而日常用语较为随性.这些不确定性对于半监督神经网络模型也同样存在.在本研究中,参考了单语语料库的来源,选取了一些与新闻、政治、法律等领域相关的平行语料,由于人为筛选语料,其中仍然夹杂着一些令人不满意的数据.然而可以看到,借助于这些小规模的平行语料库实现半监督学习,可以取得较好的结果,这对于平行语料资源很少的情况是一个值得探究的方向.

由于在小语种机器翻译研究中很少有公开且免费的语料资源,如何公正地评测那些在低资源语言对上进行实验的方法是一个巨大的挑战[25],近期 Guzmán 等[25]从维基百科收集并公开了尼泊尔语-英语以及僧伽罗语-英语的语料资源,并在此基础上进行了一些评测,实验结果显示小语种机器翻译结果仍然不容乐观.本文的后续工作也将基于该语料库展开实验,并进一步探究如何提高小语种机器翻译结果.

4　结　论

随着深度学习技术的不断发展,机器翻译取得了巨大的突破,由最初的需要人为干预翻译逐渐演变成为使用计算机来学习语言与翻译模型,掌握语言的内部结构,并进行自动翻译.伴随着越来越多的语料库资源,NMT 在翻译性能上超过了传统的 SMT,并且取得了出色的效果.NMT 不仅能够产生更加流畅的译文,而且对于文本的上下文关系以及语言的结构有更好的掌握.然而值得关注的是,NMT 的翻译性能基本上取决于训练集的平行语料库规模,因此,对于资源匮乏的小语种而言,如果语料资源不足以训练一个效果可接受的模型,那么要如何改进或寻找新的方法进行翻译是个重要的问题.

针对这个问题,本文中根据现有的语料库规模以及实验条件,探究了基于对偶学习的半监督 NMT 模型,并在藏语-汉语的语料库上进行了实验.由 He 等[16]提出的对偶学习方法使用两个独立的编码器与解码器,使用少量平行语料进行模型初始化,然后用对偶学习的迭代过程不断更新.本文中提出的模型除去了初始化的过程,通过源语言与目标语言间的词向量映射建立两种语言间的连接,使用共享编码器对两种语言映射后的词向量进行编码,再分别使用两个语言相关的解码器进行解码,实现半监督翻译.

考虑到获取一个语言的单语语料库比获取平行语料库会简单得多,半监督神经网络模型是一个值得研究的方向.从实验结果来看,使用半监督模型得到了较好的结果,但是它存在着两个缺陷:一个是对单语语料库的内容要求较高,如果语料涉及到各种各样的领域,学习效果会非常糟糕,这个问题对于神经网络模型也同样存在;另一个是它对于单语语料库的规模要求也较高,本研究中使用了约 2 000 万的汉语和 100 万的藏语数据,但是这对于半监督模型来说还是远远不够的,这也是本研究不足的地方.

参考文献:

[1] BROWN P F, PIETRA V J D, PIETRA S A D, et al. The mathematics of statistical machine translation: parameter estimation[J]. Computational Linguistics, 1993, 19(2): 263-311.

[2] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[C]// Advances in Neural Information Processing Systems. [S. l.]: NIPS, 2014: 3104-3112.

[3] FORCADA M L, ÑECO R P. Recursive hetero-associative memories for translation [C] // International Work-Conference on Artificial Neural Networks. Berlin: Springer, 1997: 453-462.

[4] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[EB/OL]. [2018-11-08]. https://arxiv.org/pdf/1409.0473.

[5] KARAKANTA A, DEHDARI J, VAN GENABITH J. Neural machine translation for low-resource languages without parallel corpora[J]. Machine Translation, 2018, 32(1/2): 167-189.

[6] 杜金华,张萌,宗成庆,等.中国机器翻译研究的机遇与挑战:第八届全国机器翻译研讨会总结与展望[J].中文信息学报,2013,27(4):1-8.

[7] CHO K, VAN MERRIËNBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [EB/OL]. [2018-11-08]. https://arxiv.org/pdf/1406.1078.

[8] HE D, XIA Y, QIN T, et al. Dual learning for machine translation [C] // Advances in Neural Information Processing Systems. [S. l.]: NIPS, 2016: 820-828.

[9] SENNRICH R, HADDOW B, BIRCH A. Improving neural machine translation models with monolingual data [EB/OL]. [2018-11-08]. https://arxiv.org/pdf/1511.06709.

[10] SENNRICH R, HADDOW B, BIRCH A. Neural machine translation of rare words with subword units[EB/OL]. [2018-11-08]. https://arxiv.org/pdf/1508.07909.

[11] ARTETXE M, LABAKA G, AGIRRE E. Learning bilingual

word embeddings with (almost) no bilingual data[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. [S. l.]: ACL, 2017,1:451-462.

[12] 李良友,贡正仙,周国栋. 机器翻译自动评价综述[J]. 中文信息学报,2014,28(3):81-91.

[13] 李亚超,江静,加羊吉,等. TIP-LAS:一个开源的藏文分词词性标注系统[J]. 中文信息学报,2015,29(6):203-207.

[14] 韩冬,李军辉,熊德意,等. 基于子字单元的神经机器翻译未登录词翻译分析[J]. 中文信息学报,2018,32(4):74-79,119.

[15] ZOPH B, YURET D, MAY J, et al. Transfer learning for low-resource neural machine translation[EB/OL]. [2018-11-08]. https://arxiv.org/pdf/1604.02201.

[16] HE D, XIA Y C, QIN T, et al. Dual learning for machine translation[C]. [S. l.]: NIPS, 2016:820-828.

[17] 李亚超,熊德意,张民,等. 藏汉神经网络机器翻译研究[J]. 中文信息学报,2017,31(6):103-109.

[18] 位素东. 基于短语的藏汉在线翻译系统研究[D]. 兰州:西北民族大学,2015.

[19] 罗延根,李晓,蒋同海,等. 基于词向量的维吾尔语词项归一化方法[J]. 计算机工程,2018,44(2):220-225.

[20] 潘一荣,李晓,杨雅婷,等. 面向汉维机器翻译的调序表重构模型[J]. 计算机应用,2018,38(5):1283-1288.

[21] 哈里旦木·阿布都克里木,刘洋,孙茂松. 神经机器翻译系统在维吾尔语-汉语翻译中的性能对比[J]. 清华大学学报(自然科学版),2017,57(8):878-883.

[22] ARTETXE M, LABAKA G, AGIRRE E, et al. Unsupervised neural machine translation [EB/OL]. [2018-11-08]. https://arxiv.org/pdf/1710.11041.

[23] LAMPLE G, CONNEAU A, DENOYER L, et al. Unsupervised machine translation using monolingual corpora only[EB/OL]. [2018-11-08]. https://arxiv.org/pdf/1711.00043.

[24] YANG Z, CHEN W, WANG F, et al. Unsupervised neural machine translation with weight sharing [EB/OL]. [2018-11-08]. https://arxiv.org/pdf/1804.09057.

[25] GUZMÁN F, CHEN P J, OTT M, et al. Two new evaluation datasets for low-resource machine translation: Nepali-English and Sinhala-English[EB/OL]. [2019-02-27]. https://arxiv.org/pdf/1902.01382.

Machine translation algorithm of low-resource languages based on semi-supervised learning

LU Wenjie[1], TAN Ruxin[1], LIU Gongshen[1,2*], SUN Huanrong[2]

(1. Shanghai Jiao Tong University, School of Electronic Information and Electrical Engineering, Shanghai 200240, China;
2. Shanghai Jiao Tong University-Shanghai Songheng Information Content Analysis Joint Lab, Shanghai 200240, China)

Abstract: Recent years, neural machine translation has achieved great development. However, its requirement for large-scale parallel corpora, translating low-resource languages fluently becomes a big challenge. This paper first briefly introduces the encoder-decoder framework and attention mechanism. Next, we propose a semi-supervised neural network model based on dual-learning, which can translate low-resource languages using some monolingual corpora and small parallel corpora. Finally, results show that semi-supervised neural machine translation can achieve reasonable results with parallel corpora which are insufficient to train a common neural model. However, the semi-supervised model requires a large number of monolingual corpora to achieve great performance.

Keywords: semi-supervised learning; low-resource language; machine translation