基于迭代知识精炼的对偶学习蒙汉机器翻译

孙 硕，侯宏旭*，乌尼尔，常鑫，贾晓宁，李浩然

（内蒙古大学计算机学院，呼和浩特 010020）

摘要：深度学习方法凭借对语义的深度理解能力在机器翻译领域取得长足的进步。然而，对于低资源语言，一个难以攻克的问题是大规模双语语料的缺乏导致的模型过拟合。本文针对低资源神经机器翻译数据稀疏的问题，提出了一种迭代知识精炼的对偶学习训练方法，利用回译扩充双语平行语料，通过迭代调整伪语料和真实语料比例在学习语言表征的同时降低噪声风险，最后结合译文质量及流利度奖励在源-目标，目标-源两个方向上优化模型参数，从而达到提升译文质量的目的。我们在 CWMT2019 蒙古语-汉语翻译任务上进行了多项实验，结果表明本文方法相比基线提高显著，充分证明该方法的有效性。

关键词：神经机器翻译；低资源语言；对偶学习；回译；知识精炼

近年来，基于深度学习的神经机器翻译(NMT)[1-2]发展迅速，它主要以交叉熵为训练准则，最小化机器译文和参考译文的熵值。因此，平行语料的质量及规模决定了模型的最终翻译水平，也成为评判监督学习方法好坏的重要指标。然而，当平行资源规模较小时，则会由于数据稀疏导致训练过拟合问题，无法很好的实现双语映射，交叉熵准则也不能发挥最大作用。

针对这一问题通常包括两种解决思路：缩小建模粒度和无监督学习。由于语料资源的匮乏，会产生大量集外词(OOV)现象，即很多测试集中的待翻译片段无法通过模型找到匹配的译文，因此 Rico 等[3]采用了一种细粒度单元的建模方式，将单词切分为子词或字符粒度，极大的缩小了词典规模，缓解了 OOV 问题，同时在向量空间中也拉近词向量之间的线性距离。然而粒度的尺寸选择和不同粒度间的对齐关系学习也影响了模型质量。Lample 等[4]提出一种无监督方法来训练机器翻译模型，他们的工作是在一个初始的翻译模型上利用大规模单语语料实现伪平行语料的扩充，并从扩充后的语料中学习语义的表达和模型的构建。但是伪语料中的数据没有对应标签的正确性指导，导致伪语料包含大量噪声，制约着模型的优化。为了缓解上述问题，HE 等[5]提出一种新的学习范式 -- 对偶学习(Dual-Learning,DL)，这种方法利用人工智能任务的对称属性使其获得更有效的反馈或正则化，从而引导和加强在数据量较少情况下的模型学习过程。机器翻译中，对偶学习可以有效地利用源语言或目标语言的单语数据进行学习。通过使用该学习机制，单语数据与平行双语数据承担相似的作用。使得在模型训练过程中，可以显著降低对平行双语数据的要求。

本文采用对偶学习方法结合强化学习方法，在蒙汉双语语料资源稀缺的前提下，通过回译(Back Translation)方法扩充语料，并对蒙汉神经机器翻译建模。实现源语-目标语和目标语-源语两个方向的翻译模型，为了评测译文的质量，我们使用大规模单语语料构建语言模型来计算模型的置信度奖励以及生成译文的语言模型得分，通过设定奖励阈值来控制模型的收敛，生成最终模型。此外，我们提出一种知识精炼方法训练模型，以缓解由于通过回译得到的大规模伪双语语料中大量“有害”噪声。我们进行了多项实验验证本文方法，首先通过对蒙古语和汉语的不同预处理操作进行对比实验，采用最优的预处理方法参与模型的训练，之后采用消融学习的方法对本文方法的各个组件进行验证，最后我们分析了句子长度对翻译质量的影响并给出翻译例句。实验结果显示本文方法始终优于基线系统，证明该方法可以有效提升蒙汉神经机器翻译的性能。

基金项目：内蒙古自治区科技成果转化(2019CG028)；内蒙古自治区自然科学基金项目(2018MS06005)

作者简介：孙 硕(1996—)， 女， 硕士研究生， 主要研究领域为自然语言处理，E-mail:sunshuo07@126.com

1 神经机器翻译

神经机器翻译(NMT)主要利用编码器-解码器结构实现对源语言的语义编码和对目标语言的预测。具体方式是利用一个 Encoder 将输入的源语言 $X=(x_1,x_2,...,x_N)$ 编码成一个固定的向量，然后利用 Decoder 对该向量进行解码，最终得到目标语言。对于 y_t，已知其之前的单词序列 $y_{<t}$ 和源语言句子 X，使用 $P(y|x)$ 来确定当前目标词的概率 $P(y_t|y_{<t},x)$。具体的计算过程如下所示：

$$P(y_t|y_{<t},x)\propto \exp(y_t;r_t;C_t) \tag{1}$$

其中，r_t 是神经机器翻译模型 Encoder 中 RNN 网络在时刻 t 的隐藏层状态。C_t 是根据 Encoder 的隐藏层节点状态定义的生成词 y_t 的上下文状态信息。

NMT 采用最大似然估计(MLE)训练。给定 J 个训练句子对 $\{X_i,Y_i\}_{i=1}^{J}$，在每个时间步长，NMT 通过最大化在源句子 X 上的翻译概率生成目标词 y_t。训练目标是最大化：

$$L_{mle}=\sum_{i=1}^{N}\sum_{t=1}^{M}\log\, p(y_t^i|y_1^i,...,y_{t-1}^i,x^i) \tag{2}$$

2 方法

在本节，我们详细的介绍提出的方法，图 1 是本文的整体框架图。首先采用回译技术对语料扩充，生成大规模伪语料，在对偶学习基础上使用大规模单语语料得到目标端语言模型，结合奖励机制对译文的流利度和相似度打分，通过最大化总奖励来优化模型。此外，由于扩充双语语料存在大量的噪声，我们提出一种迭代知识精炼的方法来降低“有害”噪声。下面我们主要介绍回译技术、对偶学习方法以及训练方式等。

2.1 迭代回译

由于蒙汉双语平行语料稀缺，单纯利用双语训练势必会因语义稀疏导致过拟合。传统方法是在训练中采用早停或正则化等方法缓解过拟合问题，但是这两种方法也无法很好拟合目标分布。为此，我们在低资源双语语料的基础上引入

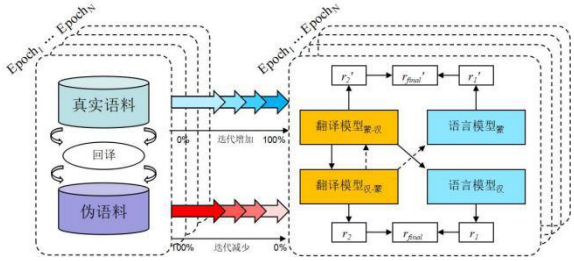

图 1 整体框架图

Fig.1 An illustration of the overall model architecture.

回译方法生成大规模伪双语语料，将生成的伪语料和已有的双语语料合成总语料一起进行接下来的模型训练。

在资源匮乏的情况下，回译是一种扩充语料的有效途径。具体来说，它基于一个初始的翻译模型，将大规模的源端单语翻译成目标端的带噪单语($x\rightarrow M_{s\rightarrow t}(x)\rightarrow y^*$)，同时将带噪目标语言翻译回含有噪声的源端单语($y^*\rightarrow M_{t\rightarrow s}(y)\rightarrow x^*$)，以此类推。

2.2 对偶学习神经机器翻译

对偶学习的最主要的内容就是根据语言模型的反馈信息来指导翻译模型的参数优化，因此本小节主要对所涉及的语言模型和翻译模型奖励如何设定进行说明。具体过程如图 2 所示。

2.2.1 语言模型奖励

模型首先通过语言模型给两个方向上的翻译模型提供流利度奖励反馈，然后通过提出的对偶学习算法，在训练中基于反馈回来的奖励来提升两个翻译模型的性能。本文采用蒙古语语料 D_{MO} 和汉语语料 D_{ZH} 两个单语语料来训练两个语言模型。这两个语料分别包含蒙古语句子和汉语句子并且不需要相互对齐。通过已经训练好的两个语言模型 $LM_{MO}(\cdot)$ 和 $LM_{ZH}(\cdot)$，每个语言模型获取目标端句子作为输入，输出一个表示这个句子流利度的奖励 r_1，公式如下：

$$r_1=\log P_T(t)=\sum_{i=1}^{M}\log P_T(t_i|t_1,...,t_{i-1}) \tag{3}$$

2.2.2 翻译置信度奖励

对偶学习翻译任务包含前向翻译步骤和反向翻译步骤。可以将句子从蒙古语翻译到汉语，反之亦然。由迭代知识精炼方法训练得到了两个

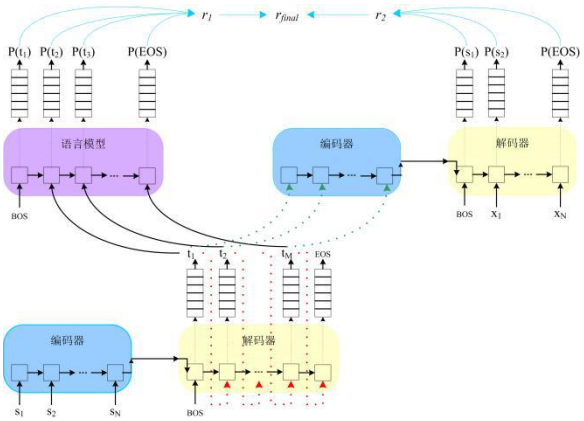

图 2 对偶学习过程

（自下而上，从左至右）

Fig.2 Dual-learning procedure.

(bottom to top, left to right)

初始翻译模型，接下来我们采用真实的蒙汉双语语料进行对偶学习。假设双语语料中语料蒙古语语料W_A有N_A个句子，汉语语料W_B有N_B个句子。两种语料需要对齐。定义$P(\cdot|T,\theta_{AB})$和$P(\cdot|T,\theta_{BA})$是两个神经机器翻译模型。其中，θ_{AB}和θ_{BA}是对应方向上的模型参数。如图 2 所示，对偶学习过程从W_A或W_B中的句子S开始，定义T作为中间翻译输出，再通过翻译模型将中间翻译T还原回S,将S的对数似然概率作为对偶置信度的奖励，其中置信度奖励公式：

$$r_2 = \log P(S|T;\theta) = \sum_{i=1}^{N} P(s_i|s_1,...,s_{i-1};\theta) \tag{4}$$

通过简单的线性组合将由语言模型产生的流利度奖励r_1和由翻译模型产生的通信奖励r_2组合作为整体反馈奖励，公式如下：

$$r_{final} = \lambda r_1 + (1-\lambda) r_2 \tag{5}$$

其中，整体反馈奖励中λ是超参数。

2.3 训练

本文通过将对偶学习机制作为蒙汉神经机器翻译的主要训练方法来改进翻译模型。为获得初始翻译模型，我们利用 2.1 节中的回译技术生成的伪语料和双语语料一起训练双向机器翻译基准模型。但是由于伪数据的质量问题，单纯的联合训练会带来大量的噪声，为精炼模型的训练数据，本文采用迭代知识精炼方法逐步细化模型性能，如图 3 所示，在每一个训练周期结束时增加一定比例的真实语料并且降低一定比例的伪数据来逐步细化基准翻译模型，直到伪双语语料为 0。这样在扩充语料使模型学习到更多的语言特征的同时，还减少了由于伪数据的质量问题带来的噪声。

由于总体奖励可以视为S，T以及翻译模型$P(\cdot|S,\theta_{AB})$和$P(\cdot|S,\theta_{BA})$的函数。因此，本文可以通过策略梯度方法来优化翻译模型中的参数，从而达到奖励的最大化。在该训练方法中，首先基于翻译模型$P(\cdot|S,\theta_{BA})$采样出T。接着

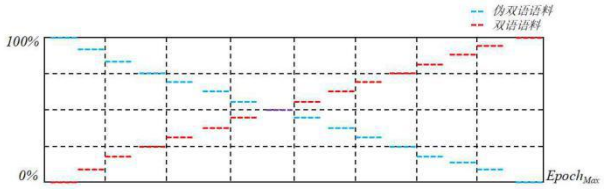

图 3 迭代知识精炼示意图

Fig.3 Iterative knowledge refinement training method.

计算期望奖励$E[r]$关于参数θ_{AB}和θ_{BA}的梯度。根据策略梯度定理得到两个方向上的训练梯度，具体如下所示：

$$\nabla_{\theta_{AB}} E[r] = E[r_{final} \nabla_{\theta_{AB}} \log P(T|S;\theta_{AB})]$$
$$\nabla_{\theta_{BA}} E[r] = E[(1-\lambda) \nabla_{\theta_{BA}} \log P(S|T;\theta_{BA})] \tag{6}$$

模型更新：

$$\theta_{AB} \leftarrow \theta_{AB} + \gamma_{1,t} \nabla_{\theta_{AB}} \overline{E}[r]$$
$$\theta_{BA} \leftarrow \theta_{BA} + \gamma_{2,t} \nabla_{\theta_{BA}} \overline{E}[r] \tag{7}$$

本文考虑到随机采样将会带来非常大的方差，并且会导致机器翻译中出现不合理的结果，针对梯度计算，本文使用束搜索(Beam-Search)的方法来获取更加合理的中间翻译输出，通过贪婪方法产生 Top-K 个高概率的中间翻译输出，然后使用束搜索的平均值来近似真实的梯度。

3 实验与分析

3.1 数据集和设置

本文实验数据共计约 310 万语料（蒙古语和汉语各 155 万），其中采用内蒙古自治区蒙古文智能信息处理重点实验室提供的各 130 万的蒙汉单语语料来训练两个语言模型，采用 CCMT2019 会议提供的近 25 万的双语平行语料用于翻译模型的训练。CWMT2017dev 作为验证集，CWMT2017test 作为测试集。为降低模型困

惑度，上述数据均删除了长度大于 50 的句子。我们采用传统的 RNNSearch[1] 和最新的 Transformer[6] 模型作为基线系统。对于 Transformer，我们的设置与原始论文的设置完全相同。dropout=0.1，词嵌入维数为 512，头数设置为 8. 编码器和解码器均具有 6 层堆栈。我们使用通过平均最近 20 个检查点获得的单个模型并且利用束搜索方法来对目标 token 进行采样，其中 beam_size=4 且长度损失α= 0.6。我们使用 Adam 算法优化，$\gamma_{1,t}=0.0002$ 并且 $\gamma_{2,t}=0.02$。对于 RNNSearch，它是一种基于 RNN 的具有注意力机制的编码器-解码器框架，我们设定编码器-解码器的隐藏层节点和词向量维度均为 512。设置 dropout=0。学习率和检查点的设置和 Transformer 保持一致。考虑到实验的硬环境和模型复杂度，我们采用 Elmo[8] 来构建和训练语言模型，它包含 2 层，512 个单元的 LSTM，词嵌入维度设置为 512。本文以句子级评估指标 BLEU[9] 作为评测标准。所有模型均在 TensorFlow 中实现，实验采用单核 NVIDIA TITAN X 显卡进行训练。当模型对开发集的评估没有任何改善时将停止训练。

3.2 语料预处理结果

本文首先在基础的 NMT 上对蒙古语和汉语分别进行不同的预处理操作进行实验，选取最优的方法训练对偶学习 NMT 模型。我们在汉语端分别进行分字，分词，BPE[3]处理。在蒙古语端，由于蒙古语本身自带天然分隔符，因此单纯采用 BPE 处理。不同切分方法示例如表 1 所示。实验结果如表 2 所示。

表 1 使用不同切分方法的结果示例

Tab. 1 An example of the sentence segmentation results by different methods

蒙古语源句	ᠬᠡᠪᠡ <200d>ᠡᠵᠡ ᠮᠠᠨᠳ ᠨᠢ ᠬᠦᠮᠦᠨ <200d>ᠡᠳ ᠪᠤ ᠲᠡᠷᠡᠭᠦᠨ ᠦᠭᠡᠢᠯᠡᠭᠰᠡᠨ ᠲᠤᠭᠠᠯᠠᠭ ᠤᠨ ᠬᠦᠮᠦᠵᠢᠯ ᠲᠦ ᠨᠠᠭᠠᠭᠠᠳ ᠠᠪᠴᠠᠷᠠᠯᠭᠦᠢ ᠪᠡᠨ ᠦᠨᠢᠳᠡᠭᠡᠨ ᠮᠥᠨ ᠃᠃
BPE 切分	ᠬᠡᠪᠡ <200d>ᠡᠵᠡ ᠮᠠᠨᠳ ᠨᠢ ᠬᠦᠮᠦᠨ@@ ᠦᠭᠡ <200d>ᠡᠳ ᠪᠤ ᠲᠡᠷᠡᠭᠦᠨ@@ ᠡ ᠦᠭᠡᠢᠯᠡᠭᠰᠡᠨ ᠲᠤᠭᠠᠯᠠᠭ ᠤᠨ ᠬᠦᠮᠦᠵᠢᠯ@@ ᠲᠦ ᠨᠠᠭᠠᠭᠠᠳ ᠠᠪᠴᠠᠷᠠᠯᠭᠦᠢ@@ ᠪᠡᠨ@@ ᠤ ᠨᠢᠳᠡᠭᠡᠨ@@ ᠮᠥᠨ@@ ᠃᠃
汉语源句	他的父母亲对儿子拒绝与这样一位美貌的姑娘成亲感到惊诧。
分字	他 的 父 母 亲 对 儿 子 拒 绝 与 这 样 一 位 美 貌 的 姑 娘 成 亲 感 到 惊 诧 。
分词	他 的 父母亲 对 儿子 拒绝 与 这样 一 位 美貌 的 姑娘 成亲 感到 惊诧 。
BPE 切分	他的@@ 父母@@ 亲@@ 对@@ 儿子@@ 拒绝@@ 与@@ 这样@@ 一位@@ 美@@ 貌@@ 的@@ 姑娘@@ 成@@ 亲@@ 感到@@ 惊@@ 诧@@ 。
分词+BPE 切分	他 的 父@@ 母亲 对 儿子 拒绝 与 这样 一 位 美@@ 貌 的 姑娘 成@@ 亲 感到 惊@@ 诧 。

表 2 不同预处理操作的 BLEU 分数

Tab.2 BLEU scores for different segmentation.

模型	蒙古语预处理	汉语预处理	BLEU
基线 1	BPE	分字	26.9
基线 2	BPE	分词	26.1
基线 3	BPE	BPE	27.2
基线 4	BPE	分词+BPE	28.5

从表 2 可以看到，基线 4 将蒙古语 BPE 处理，汉语分词+BPE 处理取得最优的结果。我们分析将汉语分词会产生数据稀疏问题影响翻译模型的性能，而分字的形式会导致语料无法保留汉语词所能包含的语义信息。在接下来的实验中，我们采用基线 4 的方法对语料预处理。

3.3 主要结果和分析

为了验证本文提出的不同组件的有效性，我们在基线系统 RNNSearch 和 Transformer 上做了一系列实验，实验结果如表 3 所示。其中超参数 $\lambda=0.005$。在 Transformer 上使用回译技术扩充双语语料在蒙古语-汉语 BLEU 值相较于未扩充语料提

升 0.4，达到 33.8，在汉语-蒙古语上也提升 0.3。在 RNNSearch 也显示了同样的趋势。这充分证明了在低资源翻译任务上使用数据增强方法扩充语料可以有效提升翻译质量。对偶学习在本文方法中扮演重要的角色，在传统的 RNNSearch 上相较于基线提升 1.7 个 BLEU 值，在 Transformer 提升更多。相较于监督学习需要大规模双语语料的局限性，对偶学习这种半监督学习范式打破了数据稀少的问题，同时相较于无监督学习完全没有对应标签指导的缺陷，对偶学习利用少量双语平行语料来训练翻译模型可以为模型提供正确性指导，以此提升了翻译质量。此外，两种模型采用迭代知识精炼方法在蒙古语-汉语语上分别提升 1.0 和 1.2，在同时采用回译对偶学习和迭代知识精炼后也分别较基线系统提升 2.7 和 3.4，分别达到 31.2 和 36.8。实验结果表明如果单纯的扩充语料而不考虑伪数据的噪声问题，虽然也可以提升模型性能，但是效果不明显。采用迭代知识精炼方法精炼模型，减少噪声对模型的影响可以显著提升模型性能。本实验进行了几次典型实验后选取每轮迭代增加 0.1%的真实双语语料同时降低 0.1%的伪语料来训练模型。

表 3　不同模型翻译结果

Tab.3　Translation results of different models.

模型	CCMT2019			
	蒙-汉	提升	汉-蒙	提升
RNNSearch	28.5	-	26.4	-
+回译	28.9	0.4	26.6	0.2
+对偶学习	30.2	1.7	27.7	1.3
+迭代衰减	31.2	2.7	28.5	2.1
Transformer	33.4	-	29.8	-
+回译	33.8	0.4	30.1	0.3
+对偶学习	35.6	2.2	31.4	1.6
+迭代衰减	36.8	3.4	32.2	2.4

3.4 句长影响

为了验证本文方法在长句上的表现，我们依照 Bahadanau[1]等人的做法将蒙古语-汉语任务的开发集数据和测试集数据按照句子长度进行划分。图 4 展示了不同的句子长度的 BLEU 值。从图中我们可以看到，当句子的长度在 10 到 20 时，在 Transformer 上的 BLEU 值达到最高为 38.7。随着句子长度的进一步扩大，模型的质量则有了一定的下降，但总体维持在 35 左右。在 RNNSearch 上也表现了同样的趋势。通过上述实验可以看出，利用对偶学习进行训练的 Dual-NMT 模型无论在翻译性能上还是在不同长度句子的翻译质量上相比较传统的 NMT 模型都有了提高。充分说明了方法的有效性，在少资源语料生成的弱翻译器存在先天翻译性能低下的条件下，通过大规模单语语料的训练支持，能够进一步加强模型的翻译质量。

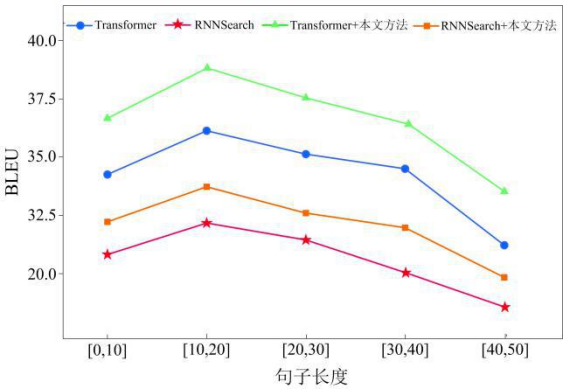

图 4 不同长度的句子的 BLEU 分数折线图

Fig.4 BLEU scores over different lengths of sentences

4 结论

本文针对蒙汉双语语料稀缺导致模型过拟合问题，采用一种半监督对偶学习方法对蒙汉神经翻译建模。通过反馈流利度奖励和通信奖励来不断地优化模型性能。此外，我们采用回译方法生成大量伪双语语料一起参与到模型训练中，为减少伪语料中大量有害噪声，提出迭代衰减方法精炼模型，减少噪声。虽然模型能够在一定程度上提高翻译质量，但是依然存在一些未登录，错译和漏译等现象。因此本文未来的任务是需要针对具体问题进行具体的分析，主要包括模型的重构、弱翻译器的结构和训练参数的优化以及语言模型的结构优化。充分实现两端语料语义特征的学习和表示，同时在词向量的处理上进一步考虑语言的自身特性和词性特点，以达到预期的翻译效果。

参考文献:

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.

[2] Sutskever I, Vinyals O, Le Q. V.: Sequence to sequence learning with neural networks. In: Neural Information Processing Systems (NIPS). pp. 3104-3112.(2014)

[3] Rico Sennrich, Barry Haddow, and Alexandra Birch.Neural machine translation of rare words with subword units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers, 2016.

[4] Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Phrase-based & neural unsupervised machine translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, pages 5039–5049, 2018.

[5] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu,Tie-Yan Liu, and Wei-Ying Ma. Dual learning for machine translation. In Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain, pages 820–828, 2016.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, pages 5998–6008.

[7] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettle moyer. 2018. Deep contextualized word representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers), pages 2227–2237.

[8] 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, July 6-12, 2002, Philadelphia, PA, USA

Dual-Learning Mongolian-Chinese machine translation based on iterative knowledge refining

Abstract: Deep learning has made great progress in the field of machine translation with its deep understanding of semantics. However, for low-resource languages, the lack of large-scale bilingual corpus leads to overfitting of models. Aiming at the problem of sparse data in low-resource neural machine translation, this paper proposes a dual-learning based on iterative knowledge refining training method, using back translation to expand bilingual parallel corpus, and the proportion of pseudo corpus and real corpus is adjusted iteratively to reduce the noise risk while learning language representation. Finally, combining the translation quality and fluency rewards to optimize model parameters in two directions: source-target and target-source, so as to achieve the purpose of improving translation quality. We conducted a number of experiments on the CWMT2019 Mongolian-Chinese translation task. The results show that the method in this paper has a significant improvement compared with the baselines, which fully proves the effectiveness of the method.

Keywords: neural machine translation; low-resource language; dual-learning; back translation; knowledge refining