计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210800254

基于语种关联度课程学习的多语言神经机器翻译

于 东[1] 谢婉莹[1] 谷舒豪[2,3] 冯 洋[2,3]
1 北京语言大学信息科学学院 北京 100083
2 中国科学院计算技术研究所 北京 100190
3 中国科学院大学 北京 100049
(yudong@blcu.edu.cn)

摘 要 近年来,使用单一模型实现多语言神经机器翻译的方法受到了广泛关注。然而,现有方法多将所有语种语料直接混合作为训练语料,未能利用多种语言之间关联和相似的信息。此外,模型训练涉及语言种类多、数据量大、整体训练难度大、耗时长等问题。针对以上两个问题,文中提出了一种基于语种关联度的课程学习方法来提高多语言神经机器翻译的整体性能和收敛速度。具体来说,提出了两种度量语种关联度的指标:使用奇异向量典型相关分析对不同语言进行排序以及使用余弦相似度对特定语言中的不同句子进行排序。进一步,文中提出以验证集损失为课程替换标准的课程学习策略,使模型训练由整体训练转化为一系列课程上的训练,降低了训练难度。该方法填补了课程学习策略在多语言神经机器翻译领域的空白。文中在平衡和非平衡的IWSLT多语言数据集和Europarl语料库数据集上进行了实验,结果表明,所提方法优于多语言基线翻译系统,最多可使训练时间缩短64%。

关键词:机器翻译;多语言;课程学习;关联度评估;语种排序;句子排序
中图法分类号 TP391

Similarity-based Curriculum Learning for Multilingual Neural Machine Translation

YU Dong[1],XIE Wan-ying[1],GU Shu-hao[2,3] and FENG Yang[2,3]
1 College of Information Sciences,Beijing Language and Culture University,Beijing 100083,China
2 Institute of Computing Technology,Chinese Academy of Sciences,Beijing 100190,China
3 University of Chinese Academy of Sciences,Beijing 100049,China

Abstract Multilingual neural machine translation (MNMT) with a single model has drawn more attention due to its capability to deal with multiple languages. However,the current multilingual translation paradigm does not make use of the similar features embodied in different languages,which has already been proven useful for improving the multilingual translation. Besides,the training of multilingual model is usually very time-consuming due to the huge amount of training data. To address these problems,we propose a similarity-based curriculum learning method to improve the overall performance and convergence speed. We propose two hierarchical criteria for measuring the similarity,one is for ranking different languages (inter-language) with singular vector canonical correlation analysis,and the other is for ranking different sentences in a particular language (intra-language) with cosine similarity. At the same time,the paper proposes a curriculum learning strategy that takes the loss of validation set as the curriculum replacement standard. We conduct experiments on balanced and unbalanced IWSLT multilingual data sets and Europarl corpus datasets. The results demonstrate that the proposed method outperforms strong multilingual translation systems and can achieve up to a 64% decrease in training time.

Keywords Machine translation,Multilingual,Curriculum learning,Similarity evaluation,Language ranking,Sentence ranking

1 引言

机器翻译是使用计算机将源语言自动翻译成目标语言的技术,是自然语言处理中的一项重要研究任务。近年来,神经机器翻译(Neural Machine Translation,NMT)由于其优越的性能引起了广泛关注[1]。通常,一个神经机器翻译模型只能从源语言翻译为目标语言,不能翻译其他语言,但世界上有数千种语言,在实际应用中经常面临多种语言相互翻译的情

到稿日期:2021-08-28 返修日期:2021-10-18
基金项目:教育部人文社会科学研究青年基金项目(19YJCZH230);北京语言大学研究生创新基金资助项目(20YCX138)
This work was supported by the Humanity and Social Science Youth Foundation of Ministry of Education(19YJCZH230) and Research Funds of Beijing Language and Culture University(20YCX138).
通信作者:冯洋(fengyang@ict.ac.cn)

况[2-3]。目前针对单一语言对的神经机器翻译模型可以取得很好的效果，但是，由于资源有限，不可能为世界上所有的语言对都训练一个单独的翻译模型。多语言神经机器翻译(Multilingual Neural Machine Translation，MNMT)使用单一模型同时处理多种语言，能够有效减小参数大小和降低训练成本，因此，这种方法正在成为新的发展方向[4-6]。

2017 年，Johnson 等[7]提出了多语言神经机器翻译系统，该系统仅包含一个共享编码器和一个共享解码器，并在输入句中添加了一个特殊的标签来确定目标语言。在训练过程中，将不同语言的训练数据组合在一起，用混合数据对模型进行训练。近年来，多语言神经机器翻译工作均采用类似的结构，而该体系结构已成为当前最流行的多语言翻译模型。

然而，这个主流模型仅仅将不同语言的数据混合在一起进行训练，未能充分考虑关联程度高的语言共同训练和关联程度低的语言共同训练的差异性。Tan 等[8]已经证明，将关联度更高的语言一起训练有助于提高模型的整体性能，因为关联度高的语言中往往体现了语种间更多的共同特征，如相似的词源、句法结构或者外来词，甚至也有相似的文化历史和地域影响。例如，训练中文和法文到英文的翻译系统较为困难，而训练德文和法文到英文的翻译系统就较为简单，也能获得更好的翻译结果。这说明，在多语言翻译中，关联度高的语言共同训练较为容易，关联度低的语言共同训练则比较困难。

此外，多语言神经机器翻译模型需要从大量双语数据中学习翻译知识，不同语言和样本的难度显然是不一样的。有的语言和句子共同训练比较困难，不易于模型学习，有的语言和句子比较简单，易于模型学习。这非常类似于 Bengio 等[9]提出的课程学习思想，也就是使模型像人类学习那样，从易到难地学习训练样本，以期获得一个平滑的目标函数，从而达到最优的局部最优点。Kocmi 等[10]的工作已经证明，这种思路在单一语言机器翻译上取得了较好的性能。不仅如此，通过获得平滑的目标函数，课程学习能够减少模型在收敛过程中的震动，有效加快收敛速度。本文的实验结果也验证了课程学习在优化模型训练时间上具有巨大贡献。

基于此，本文尝试将课程学习思想应用到 MNMT 中，提出了一种基于语种关联度的课程学习方法。本文将语种间的关联度和语种内的关联度联合建模，在随机分块的多语言训练数据上按照关联度得到分块数据的二维排序，再采用课程学习策略训练模型，最终模型能够由易到难地学习不同语言和数据，在提升性能的同时大大缩短了训练时间。在此基础上，本文提出了两个度量关联程度的维度，用于对语种和句子分别进行排序，即文中将要介绍的语种间排序和语种内排序。本文通过奇异向量典型相关分析方法，对粗粒度的语种进行排序(语种间排序)，使模型能够尽早学习关联度高的语言，同时通过余弦相似度方法，对细粒度的句子进行排序(语种内排序)，使模型能够尽早学习关联度高的句子。分别对粗粒度和细粒度两个维度进行排序，既保证了文中应用语种关联度的整体宏观框架，又能保证模型在每个批次实际学习时都是由易到难的。本文提出了一种特定的课程学习策略，将所有的数据分为不同的碎片逐步学习，并使用验证集损失作为课程替换标准，当前课程碎片的验证集损失不再下降后，进入下一课程碎片的学习。本文实验验证了该方法的有效性，在翻译性能和收敛速度上均有提升。

2　相关工作

2.1　多语言翻译

机器翻译已经从翻译单语言对扩展到使用多语言机器翻译模型翻译多语言对，通常多语言翻译使用具有共享注意力机制、编码器或解码器的框架[11-12]。近年来，研究人员对多语言翻译模型的结构进行了很多探索。Dong 等[13]于 2015 年在多种源语言翻译成多种目标语言的场景中，针对源语言应用了共享编码器，针对每种目标语言应用了单独的解码器。Johnson 等[7]于 2017 年使用通用编码器-解码器结构来同时处理多种源语言和目标语言，并在输入表示中使用特殊标签来确定输出为哪种语言。Wang 等[14]于 2019 年提出了一种紧凑且语言敏感的方法，通过使用具有语言敏感的嵌入、注意力和鉴别器的框架来替换编码器和解码器。虽然上述工作都专注于更好地设计多语言翻译模型以减小模型大小和参数，但忽略了收敛速度也是值得研究的一个问题[15]。

2.2　课程学习

本文方法为 MNMT 引入了课程学习方法。课程学习一经提出就被广泛应用到机器翻译训练中。由于能够更平滑地优化训练目标，翻译模型往往能够找到参数空间中更好的局部最优点，并且模型的收敛速度大大加快。课程学习的相关研究已经持续多年，并且在神经机器翻译任务上有明显的效果，表现了课程学习提高翻译质量和收敛速度的潜力[16-18]。Platanios 等[19]于 2019 年使用单词稀有度和句子长度作为衡量不同句子排名的方法，缩短了训练时间。Zhang 等[20]于 2019 年根据未标记域训练样本与域内数据的相似性对其进行排名，提高了域内数据的翻译质量。课程学习已经被引入到神经机器翻译的各个特定领域并取得了较好的效果，如领域自适应、强化学习[21]和非自回归神经机器翻译[22]。多语言翻译也是应用课程学习的合适任务，但是以往的工作忽略了这一点。因此，本文基于语种间和语种内的关联度研究了多语言神经机器翻译的课程学习方法。

3　基于语种关联度的课程学习方法

本文提出了一种基于关联度的多语言神经机器翻译课程学习新方法，如图 1 所示。该方法首先进行关联度评估，其中包括语种间排序和语种内排序，然后使用课程学习策略进行模型的训练，每个训练阶段包含不同关联度的数据。课程学习需要解决两个主要问题：如何对训练实例进行排序，以及如何设计训练策略。针对第一个问题，本文认为输入数据的关联度越高，模型就越容易学习。因此，本文将介绍两个用于关联度评估的方法，分别从不同语种间和同一语种内两个层次来对训练数据进行排序。针对第二个问题，本文将介绍一种基于验证集损失的训练策略，使模型能够及时在不同难度的训练样本之间进行学习。

3.1　语种关联度评估

3.1.1　语种间排序(inter-language ranking)

为了度量不同语种之间的关联度，本文使用奇异向量典

型相关分析[23](Singular Vector Canonical Correlation Analysis,SVCCA)。它可以帮助人们理解各种深度神经网络在训练过程中的内部表征,来检验模型学习到的不同语言的表征之间的关系[24]。

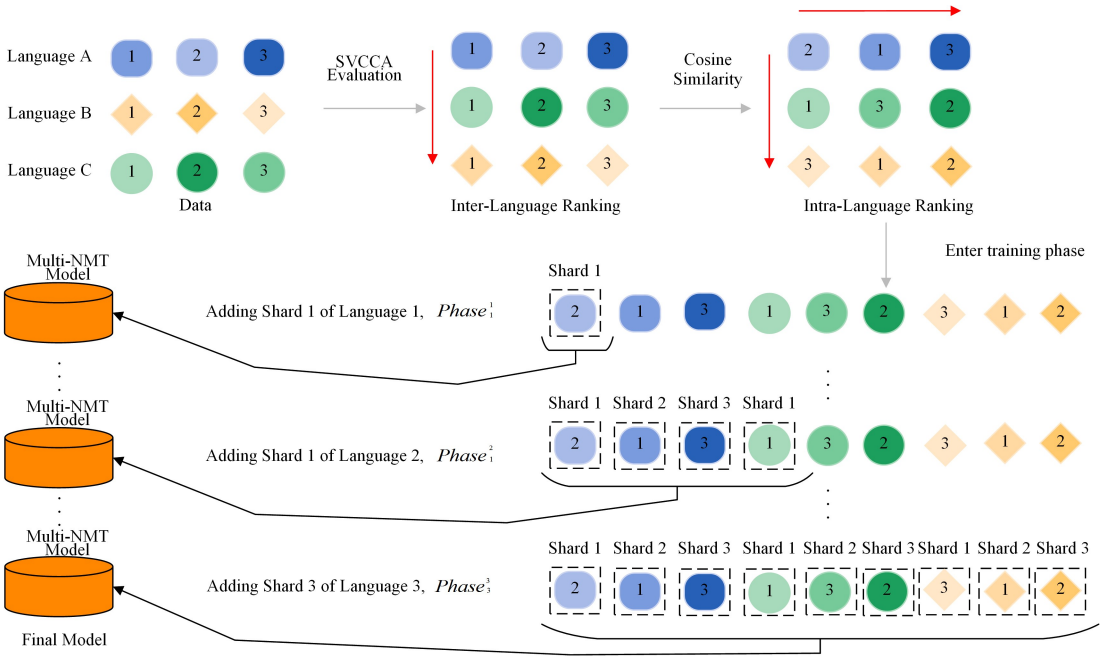

图 1 所提方法的整体训练流程

Fig. 1 Whole training process of the proposed method

假设有 N 个语种,本文使用 $L=\{l_1,l_2,\cdots,l_N\}$ 来表示这 N 个语种的训练语料库,使用 m_i 表示 l_i 语种内的句子数量。为了比较不同语种表征的关联关系,需要得到每个语种内每个句子的向量表示。通常词向量表示比较容易获得,当前有多种获得词向量表示的方法,如 Word2Vec,GloVe 和 BERT。本文使用多语言 BERT 模型获得每个词语的向量表示,并通过在句子长度上进行平均来得到一个句子的句子向量表示。那么给定语种 l_1 和 l_2 的数据集,它们分别具有数量为 m_1 和 m_2 的训练样本,本文将这两个数据集的语料分别表示为 $l_1=\{s_1^{l_1},\cdots,s_{m_1}^{l_1}$ 和 $l_2=\{s_1^{l_2},\cdots,s_{m_2}^{l_2}\}$,其中 $s_{m_1}^{l_1}$ 是语种 l_1 的第 m_1 个样本的句子向量表示,$s_{m_2}^{l_2}$ 是语种 l_2 的第 m_2 个样本的句子向量表示。同时,d_1 表示语种 l_1 内句子向量的维度,d_2 表示语种 l_2 内句子向量的维度,那么语种 l_1 的整体表示的维度服从 $l_1\in R^{m_1\times d_1}$,语种 l_2 的整体表示的维度服从 $l_2\in R^{m_2\times d_2}$。

已知某两个语种的数据集分别表示为 l_1 和 l_2,那么使用奇异向量典型相关分析进行如下两步操作。

第一步 为了降低噪声并提高计算效率,需要找到最重要的奇异值来代表整个表示。本文使用奇异值分解(Singular Value Decomposition,SVD)分别对 l_1 和 l_2 进行计算,以获得包含原始空间最重要方向的子空间 l_1' 和 l_2',其中 $l_1'\in R^{m_1'\times d_1'}$ 和 $l_2'\in R^{m_2'\times d_2'}$,子空间的维度 m_1',d_1',m_2' 和 d_2' 远小于原来未经过奇异值分解的 m_1,d_1,m_2 和 d_2。通过奇异值分解,原先非常大的矩阵被分解为具有原空间重要方向的小矩阵,达到了降维和去噪的目的,以便接下来的关联度计算。

第二步 计算 l_1' 和 l_2' 的典型相关系数,以反映两组指标之间的整体相关性,以衡量两个语种之间的关联度[25]。首先对 l_1' 和 l_2' 分别进行线性变换:

$$\widetilde{l_1}=W_1 l^{1\prime} \tag{1}$$

$$\widetilde{l_2}=W_2 l_2' \tag{2}$$

典型相关分析希望求出向量 $\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$,使得随机变量 $\widetilde{l_1}$ 和 $\widetilde{l_2}$ 的相关性最大,相关性 ρ 的定义如下:

$$\rho=\frac{cov(\widetilde{l_1}',\widetilde{l_2}')}{\sqrt{var(\widetilde{l_1}')var(\widetilde{l_2}')}} \tag{3}$$

其中,$var(\cdot)$ 代表对应向量的方差,$cov(\cdot)$ 代表对应向量的协方差。使相关性 ρ 最大的随机变量 $\widetilde{l_1}$ 和 $\widetilde{l_2}$ 是第一对典型变量,其相关性表示为 ρ_1。然后,寻求一个最大化相关性 ρ 但与第一对典型变量不相关的向量,这样就得到了第二对典型变量。持续进行这个步骤 $\min(m_1',m_2')$ 次,最终得到相关性:

$$corr=\{\rho_1,\cdots,\rho_{\min}(m_1',m_2')\} \tag{4}$$

遵循 Raghu 等[23]的做法,本文使用相关性的均值 $\bar{\rho}$ 作为所求语种表征相关性的近似表示,本文将其称为 SVCCA 分数。针对多种源语言对一个目标语言的翻译场景,本文首先计算目标语言和每种源语言之间的 SVCCA 分数,确定最相关的语言,并将其作为第一个被训练的语种。本文将模型此时能看到的样本称为数据池,那么此时就将最相关的语种的数据放入数据池中。然后,计算还未用于训练的语言与已进入数据池中的语言之间的 SVCCA 分数,以确定第二个被训练的语种,以此类推,完成所有语种的排序。针对一种源语言到多个目标语言翻译场景的排序与上述流程相似,不同之处在于开始时是计算源语言和每种目标语言之间的 SVCCA 分数。最终,得到已经排序完成的语种为 $\{l_1,l_2,\cdots,l_N\}$,l_1 最为相关,首先将其加入数据池中进行训练,l_N 最不相关,最后将其加入数据池中进行训练。

3.1.2 语种内排序(intra-language ranking)

在上述跨语言的语种排序之后,可以得到从关联度高到关联度低的语种的训练顺序。然而,以往的课程学习方法都

是针对句子级别的，如果只对语种进行排序仍然是粗粒度的，因此本文进一步对更细粒度的句子层面进行探索，根据关联度对每种特定语言内的句子进行排序。遵循前文的内容，输入的数据越相关，模型就越容易学习。针对某个句子，文中计算数据池中的所有句子和这个句子之间的向量表示的平均余弦相似度来测量关联度。

假设已完成$l_{1:i-1}$的排序，那么此时数据池中已有$l_{1:i-1}$的所有样本数据，接下来将进行语种l_i中句子的关联度衡量。因此，对于l_i中的第m个句子，将其句子向量表示为$\boldsymbol{s}_m^{l_i}$，它与之前训练句子的关联度得分如式(5)所示：

$$score(\boldsymbol{s}_m^{l_i})=\frac{\sum_{j=1}^{i-1}\sum_{d=1}^{m_j}\cos(s_d^{l_j},s_m^{l_i})}{\sum_{j=1}^{i-1}m_j} \tag{5}$$

其中，$l_j\in l_{1:i-1}$，$s_d^{l_j}$是l_j中第d个句子的向量表示，$cos(\cdot)$表示余弦相似度计算。然后，根据关联度分数对l_i中的句子进行排序，分数更大意味着该句子与已加入数据池的语言和句子更相关，因此更早地将这个句子加入数据池中。

需要说明的是，在多对一翻译场景中，对于最相关的语种l_1，这个语种内每个句子的关联度分数是根据目标语言的句子计算的；在一对多翻译场景中，最相关语种l_1内每个句子的关联度分数是根据源语言的句子计算的。

经过语种间排序和语种内排序，不同的语言和每种语言中的不同句子就按照它们的关联度进行排序。

3.2　课程学习训练策略

排序完成后，本文提出了以下课程学习策略，如算法1所示。对于有序语言$L=(l_1,l_2,\cdots,l_N)$，每种语言内排序后的句子被划分为不同的碎片，因此每个碎片内句子的关联度相差较小。本文用$H=(h_1^{l_i},h_2^{l_i},\cdots,h_o^{l_i})$表示语言$l_i$中的碎片，$o$是$l_i$语种内碎片的数量。

算法1　课程学习算法

输入：Ordered languages L，shards H，and hyper parameter λ that controls updates for moving to next phase

输出：The final MNMT model

1. for language l=1，…，N do
2.　for shard h=1，…，O in language l do
3.　　$Phase_h^l$：Add shard h to existing data
4.　　$\tilde{t}$ is the best validation loss updates
5.　　t is the current updates
6.　　while training step $t-\tilde{t}<\lambda$ do
7.　　Train model
8.　　end while
9.　　Move to next stage：$phase_{h+1}^l$
10.　end for
11.　Move to next stage：$phase_1^{l+1}$
12. end for.

本文将训练过程的每个单元定义为阶段，其中每个阶段只有部分碎片可用于训练，阶段$phase_o^n$表示刚刚加入的碎片是第n个语种的第o个碎片。在这个连续的阶段中，第一个阶段$phase_1^1$的数据池中只包含最关联的语种的最简单的碎片$h_1^{l_1}$，当进入下一阶段$phase_2^1$时，数据池中包含最相关语言的第二个最简单的碎片$h_2^{l_1}$。在添加了最相关语言l_1的所有碎片后，进入阶段$phase_1^2$，数据池中将添加第二相关语言的最简单碎片$h_1^{l_2}$。

为了确定模型是否可以进入下一阶段，本文使用验证集的损失作为度量。如果当前验证集的损失大于最佳的验证集损失，并且已经持续λ次的更新步数均大于最佳的验证集损失，则表示模型已经学会了处理当前数据，那么训练过程将进入下一阶段，即在数据池中添加下一碎片。

综上所述，训练策略共有$N*O$个阶段，其中N是语言的数量，O是每种语言的碎片数量。

4　实验

4.1　数据准备

本文在多对一和一对多两个场景中进行了实验，两种场景的数据分别如下。

(1)多对一场景(Many-to-One)。本文在语料平衡数据集和语料不平衡数据集上分别测试了所提方法。平衡数据集选用2017年国际顶级口语机器翻译评测大赛(The International Conference on Spoken Language Translation-2017，IWSLT-2017)中的翻译数据集，包括法语、意大利语、罗马尼亚语、荷兰语、德语这5种语言翻译为英语的数据，本文分别将其简称为Fr，It，Ro，Nl，De，每种语言对分别有23.28万、23.16万、22.05万、23.72万和20.61万平行句对。本文分别选择test2016和test2017作为开发集和测试集。所有语言的句子都由Moses脚本进行分词，并使用字节对编码(Byte Pair Encoder，BPE)规则对所有语言联合进行32 000合并操作，以进一步将词语分割为子词符号。

对于不平衡数据集，本文同样在IWSLT-2017数据集上进行实验，选择法语、德语、中文、罗马尼亚语、日语这5种语言翻译为英语的平行语料库，分别简称为Fr，De，Zh，Ro，Ja。为了构造不平衡场景，本文从罗马尼亚语和日语中分别采样2万个句子进行实验，因此这些语种的平行句对数量分别为23.28万、20.61万、23.13万、2万和2万。使用中文词法分析工具包(THU Lexical Analyzer for Chinese，THULAC)对中文句子进行分词，使用Mecab对日语句子进行分词，使用Moses对其他句子进行分词。使用BPE规则对所有语言对进行40 000合并操作。

(2)一对多场景(One-to-Many)。本文使用第7次发布的欧洲议会平行语料库(European Parliament Proceedings Parallel Corpus-v7)训练数据来评估本文方法的性能。选取英语翻译为12种语言的平行句对：捷克语、芬兰语、希腊语、匈牙利语、立陶宛语、拉脱维亚语、波兰语、葡萄牙语、斯洛伐克语、斯洛文尼亚语、瑞典语和西班牙语(本文中分别简称为Cs，Fi，El，Hu，Lt，Lv，Pl，Pt，Sk，Sl，Sv和Es)。对于每个语言对，随机抽取60万平行句对作为训练语料库，因此整个实验总共包含720万平行句对。使用dev2006作为验证集，使用devtest2006作为测试集。对于没有可用于开发集和测试集的语言对，本文从相应的训练集中分别随机筛选1 000个未见过的句子对作为开发集和测试集。本文使用Moses方法对句子进行分词，并从所有源端和目标端合并一起的训练数据联合学习9万步大小的BPE词表。

4.2　实验设置

(1)模型参数。本文使用Facebook发布的开源工具包

Fairseq-py 来实现所提方法，使用 Transformer_base[26] 的配置作为基本模型结构。实验选用 Adam 作为优化器，其中 $\beta_1=0.9$，$\beta_2=0.98$，并且 $\epsilon=10^{-9}$，设置失活率（dropout）为 0.3。在推理过程中，所有语言的光束搜索大小设置为 5，长度惩罚设置为 $\alpha=1.4$。最终翻译结果被去分词化，然后使用 4-gram 并且区分大小写的 SacreBLEU 工具来评估质量。

（2）关联度评估。本文使用开源的多语言模型 multilingual cased BERT-Base 来得到每个句子的词嵌入表示。对于语种内的句子排序，由于计算余弦相似度非常耗时，根据对当前数据池 $l_{1,i-1}$ 中的句子采样出 2 万个句子用于计算。

（3）课程学习。本文尝试了不同的碎片数量，最终发现在每个语种内包含 3 个碎片能够获得最佳效果，因此多对一场景实验将 5 种语言共分成 15 个碎片，一对多场景实验将 12 种语言共分成 36 个碎片。为了使模型更加关注新添加的语言和相应的句子，本文对它们的训练数据进行了过采样（例如，在 $Phase_i^l$ 中对第 l 种语言的数据进行过采样），给模型足够的时间来吸收新语言的信息内容。同样，每个碎片中的句子数量并不相等，而是以 3:2:1 的比例递减。在训练时，每个碎片中的句子将被打乱，因此它们被随机输入模型。对于不平衡数据集，将高资源语言对和低资源语言对分为两部分，先确定高资源语料上的语言顺序，然后确定低资源语料的语言顺序，通过后训练低资源语言对来使它们获得多资源语言对的知识。

4.3 实验系统

本文基于 Transformer 框架进行了实验，并将本文方法与 3 个基线系统进行了比较。

（1）单模型基线（Individual）：使用 Transformer 模型在每个语言对上进行训练。N 语言对会有 N 个不同的模型。

（2）多语言基线（Multilingual）：该方法在包含一个编码器和一个解码器的单个模型中同时处理多种语言，并且使用一个特定的语言标记来确定翻译的方向。

（3）基于能力的方法（Competence）：该方法使用句子长度和单词稀有度来衡量句子的难度，因此定义每个句子的难度时没有考虑语言差异。这是一种对 NMT 有效的课程学习方法，但不是为 MNMT 设计的。在这个系统中，来自不同语言的所有句子同时进行排序。

（4）本文提出的方法（Our Method）：模型设置与多语言基线相同，但根据提议的课程学习策略进行训练。

4.4 实验结果

首先报告 SVCCA 对语种顺序的评估结果。在多对一任务中，对于平衡数据集，语种顺序是 Fr-It-Ro-Nl-De，对于不平衡数据集，语种顺序是 Fr-De-Zh-Ro-Ja。对于一对多任务，语种顺序是 Es-Pt-Sv-Pl-Cs-Sk-Sl-El-Fi-Lv-Lt-Hu。

多对一场景（Many-to-One）中，平衡数据集的实验结果如表 1 所列。本文提出的课程学习方法始终比 Multilingual 的表现更好，*BLEU* 值在 Ro 上可以实现高达 0.93 个百分点的提升。与基于能力的课程学习方法（Competence）相比，本文方法也可以获得更好的翻译效果。相比之下，在模型参数数量相同的情况下，所提方法显著缩短了训练时间，最终仅为 Multilingual 的 67%。

表 1 多对一场景中平衡数据集的实验结果

Table 1 Experimental results of balanced data set in many-to-one scenarios

System	*BLEU*/%						Time
	Fr	It	Ro	NL	De	AVE	
Individual	37.39	37.92	32.85	32.12	28.10	33.68	0.58
Multilingual	37.04	37.79	32.42	32.63	27.83	33.54	1.00
Competence	37.68	38.00	32.65	32.67	27.94	33.79	0.71
Our Method	37.85	38.30	33.35	32.85	28.37	34.14	0.67

对于不平衡数据集，实验结果如表 2 所列，本文方法也能取得比多语言基线更好的结果，尤其是在低资源语言对 Ro 和 Ja 中。此外，还可以观察到，与单模型基线（Individual）相比，Ro 和 Ja 的翻译性能分别可以达到 14.36（26.10 对比 11.74）和 6.81（9.03 对比 2.22）的提升。然而，尽管本文方法可以在低资源语言对上获得优秀的结果，但在其他语言对的翻译质量上并没有超过单模型基线。这可能是因为多语言翻译模型可能会牺牲高资源语言对的翻译质量来提高低资源语言对的翻译质量。此外，本文方法还极大地缩短了训练时间，仅为多语言基线的 36%，并且同时获得了更好的性能。本文方法的训练时间与单模型基线相当，但总的参数量大大减少。这证明了本文方法在平衡和不平衡数据集上的有效性。

表 2 多对一场景中不平衡数据集的实验结果

Table 2 Experimental results of unbalanced data set in many-to-one scenarios

System	*BLEU*/%						Time
	Fr	De	Zh	Ro	Ja	AVE	
Individual	37.39	28.10	21.34	11.74	2.22	20.16	0.35
Multilingual	35.76	26.44	20.58	24.74	7.70	23.04	1.00
Competence	36.02	26.76	20.55	24.98	7.83	23.23	0.62
Our Method	36.41	27.21	20.89	26.10	9.03	23.93	0.36

一对多场景（One-to-Many）的实验结果如表 3 所列。单模型基线（Individual）是最快的，但翻译效果一般，因为每个语言对的数据量不足以训练出令人满意的模型。多语言基线（Multilingual）比单模型基线消耗了更多的时间进行训练，但效果稍好。本文方法在平均翻译 BLEU 值上比其他基线略有提升，并且训练时间是所有多语言模型中最短的。

表 3 一对多场景的实验结果

Table 3 Experimental results in one-to-many scenarios

System	*BLEU*/%													Time
	Cs	El	Es	Fi	Hu	Lt	Lv	Pl	Pt	Sk	Sl	Sv	AVE	
Individual	36.14	39.86	41.16	22.95	31.75	32.31	38.12	32.95	35.57	40.51	43.83	33.23	35.70	0.43
Multilingual	37.87	40.34	41.58	23.03	31.10	33.11	39.22	32.67	36.20	42.05	44.76	33.16	36.26	1.00
Competence	37.61	40.50	41.94	22.91	31.41	32.32	39.12	33.60	36.03	41.93	44.22	32.94	36.21	0.91
Our Method	38.05	40.06	41.75	23.30	31.67	32.94	39.55	32.79	36.22	41.94	44.56	32.75	36.30	0.85

5　分析

5.1　消融分析

实验结果如表4所列，首先消去语种内的排名，即忽略了不同句子的难度，只考虑语种的顺序。最终翻译性能大大下降，这表明语种内排序有助于提升翻译性能。实验还试图放弃语种间排序，即将每种语言中包含容易句子的碎片的集合作为训练模型的早期数据池。观察实验结果发现，翻译性能下降会更剧烈，这意味着基于关联度的语种间排序在设计课程学习策略时更加全面和合理。接下来，实验排除了这两种排序，并且发现性能进一步下降，这表明这两个层次的排序方式对模型的训练均具有影响。

表4　消融实验的结果

Table 4　Results of ablation study

System	Fr	It	Ro	Nl	De	AVE
Full	37.85	38.30	33.35	32.85	28.37	34.14
-Intra	37.68	38.12	32.83	32.26	27.97	33.77
-Inter	37.40	37.82	32.19	32.66	28.08	33.63
-Both	37.04	37.79	32.42	32.63	27.83	33.54

5.2　课程策略比较

课程学习策略定义了将不同关联度的句子和语言对呈现给模型的顺序。除了本文方法外，还有其他课程学习策略。

(1)Our Method：更相关的语言和包含更多相关句子的碎片将最先被加入数据池中。

(2)Shards_reverse：颠倒碎片的呈现顺序但不改变语种的顺序，即包含不相关句子的碎片将最先被加入数据池中。

(3)Languages_reverse：颠倒语种的呈现顺序但不改变每个语种的碎片顺序。

(4)Languages_random：随机排列语种顺序但不调整碎片顺序。这样加入数据池的语种与语言关联度无关，而是以随机顺序加入的。

实验结果如表5所列，本文方法可以在所有语言对中获得更高的*BLEU*分数，并且花费的训练时间更短。这表明最相关的语种和样本先训练的课程学习想法是有效的。

表5　不同课程策略的实验结果

Table 5　Results of different curriculum strategies

System	Fr	It	Ro	Nl	De	Time
Our Method	37.85	38.30	33.35	32.85	28.37	1.00
Shards_reverse	37.30	37.28	32.79	32.25	27.44	1.14
Languages_reverse	36.58	37.39	32.39	32.49	28.02	1.12
Languages_random	36.69	37.54	32.95	32.28	27.52	1.11
Multilingual	37.04	37.79	32.42	32.63	27.83	1.49

5.3　收敛表现

为了检查所提方法是否在性能和速度这两个指标上进行了改进，本文对比了所提方法和多语言基线(Multilingual)在验证集上的损失。

如图2所示，本文方法在验证集上的损失大致为五阶梯状分布，这是因为根据提出的训练策略，训练数据是逐步呈现的。由于翻译模型会吸收更多的信息内容，在添加新语言的训练数据时，验证损失大大减少，故呈现阶梯状。与添加新语种相比，在一种语言中添加新的碎片只会导致验证集损失略有减少，这表明语种间的顺序比语种内的顺序更加有效。

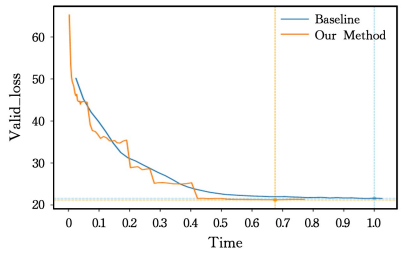

图2　验证集损失的对比

Fig. 2　Comparison of validation set loss

总的来说，本文方法在收敛后可以获得比多语言基线更低的损失值。较低的损失表明可以实现更好的翻译质量。此外，本文方法的损失下降很快，收敛时间仅为多语言基线训练时间的67%，这意味着可以大大加快收敛速度。

结束语　多语言神经机器翻译是同时处理多个语言对的典型方法，但在大多数情况下，多语言模型的性能较差而且训练非常耗时。为了解决这些问题，本文引入了一种基于关联度的课程学习方法来对语种和句子进行排序，从而使模型以与人类相似的方式学习数据。本文将SVCCA应用到语言间排序，将余弦相似度应用到语种内排序。在多个数据集上进行了实验，结果证明所提方法能够用更短的训练时间获得更好的翻译性能，这点在训练数据量较少的语种中表现得更加明显。然而，本文方法仅仅是课程学习在多语言翻译方向的一个尝试，未来可以在衡量关联度的不同方式以及其他课程学习训练策略上进行更多的探索。

参 考 文 献

[1] KALCHBRENNER N, BLUNSOM P. Recurrent continuous translation models[C]// Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013: 1700-1709.

[2] AHARONI R, JOHNSON M, FIRAT O. Massively multilingual neural machine translation[J]. arXiv: 1903.00089, 2019.

[3] ARIVAZHAGAN N, BAPNA A, FIRAT O, et al. 2019. Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges[J]. arXiv: 1907.05019, 2019.

[4] HA T L, NIEHUES J, WAIBEL A. Toward multilingual neural machine translation with universal encoder and decoder[J]. arXiv: 1611.04798, 2016.

[5] XUE Q T, LI J H, GONG Z X. Multi-language unsupervised neural machine translation[J]. Journal of Xiamen University (Natural Science), 2020, 59(2): 192-197.

[6] FIRAT O, CHO K, BENGIO Y. Multi-Way, Multilingual Neural Machine Translation with a Shared Attention Mechanism[C]// Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016: 866-875.

[7] JOHNSON M, SCHUSTER M, LE Q V, et al. Google's multilingual neural machine translation system: Enabling zero-shot translation[J]. Transactions of the Association for Computational Linguistics, 2017, 5: 339-351.

[8] TAN X, CHEN J, HE D, et al. Multilingual Neural Machine Translation with Language Clustering[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: 963-973.

[9] BENGIO Y, LOURADOUR J, COLLOBERT R, et al. Curriculum learning[C]// Proceedings of the 26th Annual International Conference on Machine Learning. 2009: 41-48.

[10] KOCMI T, BOJAR O. Curriculum Learning and Minibatch Bucketing in Neural Machine Translation[C]// Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2017). 2017: 379-386.

[11] LU Y, KEUNG P, LADHAK F, et al. A neural interlingua for multilingual machine translation[C]// Proceedings of the Third Conference on Machine Translation: Research Papers. 2018: 84-92.

[12] GU J, HASSAN H, DEVLIN J, et al. Universal Neural Machine Translation for Extremely Low Resource Languages[C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). 2018: 344-354.

[13] DONG D, WU H, HE W, et al. Multi-task learning for multiple language translation[C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2015: 1723-1732.

[14] WANG Y, ZHOU L, ZHANG J, et al. A compact and language-sensitive multilingual translation method[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 1213-1223.

[15] DABRE R, FUJITA A. Recurrent stacking of layers for compact neural machine translation models[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33(1): 6292-6299.

[16] WANG S, FAN Y X, GUO J F, et al. Dynamic Learning Method of Neural Machine Translation Based on Sample Difficulty[J]. Journal of Guangxi Normal University (Natural Science Edition), 2021, 39(2): 13-20.

[17] WANG W, WATANABE T, HUGHES M, et al. Denoising Neural Machine Translation Training with Trusted Data and Online Data Selection[C]// Proceedings of the Third Conference on Machine Translation: Research Papers. 2018: 133-143.

[18] KUMAR G, FOSTER G, CHERRY C, et al. Reinforcement Learning based Curriculum Optimization for Neural Machine Translation[C]// Proceedings of NAACL-HLT. 2019: 2054-2061.

[19] PLATANIOS E A, STRETCU O, NEUBIG G, et al. Competence-based Curriculum Learning for Neural Machine Translation[C]// Proceedings of NAACL-HLT. 2019: 1162-1172.

[20] ZHANG X, SHAPIRO P, KUMAR G, et al. Curriculum Learning for Domain Adaptation in Neural Machine Translation[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019: 1903-1915.

[21] ZHAO M, WU H, NIU D, et al. Reinforced Curriculum Learning on Pre-Trained Neural Machine Translation Models[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(5): 9652-9659.

[22] GUO J, TAN X, XU L, et al. Fine-tuning by curriculum learning for non-autoregressive neural machine translation[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 7839-7846.

[23] RAGHU M, GILMER J, YOSINSKI J, et al. SVCCA: Singular Vector Canonical Correlation Analysis for Deep Learning Dynamics and Interpretability[C]// Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems. 2017: 6076-6085.

[24] KUDUGUNTA S, BAPNA A, CASWELL I, et al. Investigating Multilingual NMT Representations at Scale[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: 1565-1575.

[25] HARDOON D R, SZEDMAK S, SHAWE-TAYLOR J. Canonical correlation analysis: An overview with application to learning methods[J]. Neural computation, 2004, 16(12): 2639-2664.

[26] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems. 2017: 5998-6008.

YU Dong, born in 1982, Ph.D, associate professor, master supervisor, is a member of China Computer Federation. His main research interests include natural language processing and artificial intelligence.

FENG Yang, born in 1982, Ph.D, professor, Ph.D supervisor, is a member of China Computer Federation. Her main research interests include natural language processing, machine translation and dialogue.

(责任编辑:李亚辉)