doi:10.6043/j.issn.0438-0479.202011023

# 抑制维汉神经机器翻译代词性别偏见的方法

史学文,黄河燕,鉴 萍*,唐翼琨
(北京理工大学计算机学院,北京市海量语言信息处理与云计算应用工程技术研究中心,北京 100081)

**摘要**:利用神经机器翻译进行维吾尔语到汉语的翻译时,维吾尔语中的代词不区分性别,给翻译模型在汉语端使用正确的代词带来了挑战.另外,由于训练数据集中不同性别的代词使用频率差异明显,神经机器翻译倾向于输出阳性代词而不是更恰当的代词.基于此,利用汉语单语语料构造伪平行数据以扩展原训练集,缓解训练集本身的代词不平衡问题;并分别引入性别标记和翻译、性别预测联合建模两种方法,将代词性别预测显式地融入神经机器翻译的训练过程.在多个维汉翻译测试集上进行实验验证,结果表明该方法相对于基线系统,在不影响翻译质量的情况下缓解了翻译输出结果的性别偏见问题,在代词性别预测的精度上也有显著提升.

**关键词**:神经机器翻译;性别偏见;伪数据;性别预测
**中图分类号**:TP 391.2 **文献标志码**:A **文章编号**:0438-0479(2021)04-0693-08

近年来,端到端的神经机器翻译(neural machine translation, NMT)[1-2]在诸多语言的翻译任务上取得了令人瞩目的成果[3-5].随着少数民族语言机器翻译在机器翻译领域越来越受关注,以及针对资源稀缺语言的 NMT 技术的发展[6-7],NMT 逐渐成为我国少数民族语言到汉语翻译的主要技术手段[8].

NMT 模型的翻译表现通常与训练数据的特征密切相关,除数据规模外,语料中的一些数据不平衡也会对翻译结果产生影响[9-10].机器翻译训练集中的低频词相对于高频词更难以正确翻译[11-12].样本分布的不平衡是造成机器翻译性别偏见问题的重要原因,具体到维汉翻译,汉语端阴性样本出现频率较低,模型难以从中学习到更多关于性别区分的有效信息.Ott 等[10]在针对数据集不确定性的研究中指出,性别等信息容易在翻译过程中丢失.此外,不同语言在语法上对性别的区分程度不同,也造成了翻译中难以正确使用性别代词.以维吾尔语(以下简称"维语")到汉语的机器翻译为例,源语言维语端代词不区分阴阳性,而目标语言汉语的代词区分阴阳性(例如"她/他"),这给维汉机器翻译带来了很大的挑战.以第 15 届全国机器翻译大会(CCMT 2019)维汉新闻机器翻译任务[8]为例,该任务的训练数据中在目标语言(汉语)端出现阳性代词的频率远高于阴性代词,造成了代词的性别偏见问题,并最终反映在测试时的翻译结果上,如表 1 所示.表 1 中"训练集"指利用 NMT 模型重新翻译训练集语料而得到的数据,可以看出,在原始语料中,阴性代词与阳性代词的比例约为 1∶5,而在模型输出的翻译结果中,该比例约为 1∶15.产生这种现象的主要原因是数据中阳性代词出现的频率远高于阴性代词,造成 NMT 模型在解码时更倾向于给阳性代词分配更高的估计概率.

近年来,机器翻译系统的性别偏见问题逐渐引起研究者们的重视[13-16],并且研究者们给出了专门针对机器翻译性别偏见问题的评价指标和测试数据[13-14].Michel 等[16]提出了一种个性化翻译方法,该方法通过引入用户向量来控制不同用户输出的翻译特征,实现了调整性别、职业等相关信息翻译的目的.Vanmassenhove 等[17]为 10 种欧洲语言翻译任务人工标记了源语言端的性别信息,将性别信息作为 NMT 的输入,以控制 NMT 生成的目标语言性别极性.Stanovsky 等[14]研究了机器翻译语料中性别偏见的问题,设计了衡量性别偏见的方法,并利用数据集 Winogender[18]和 WinoBias[19]构

**收稿日期**:2020-11-14 **录用日期**:2021-04-17
**基金项目**:国家重点研发计划(2017YFB1002103);国家自然科学基金(61732005)
* **通信作者**:pjian@bit.edu.cn
**引文格式**:史学文,黄河燕,鉴萍,等.抑制维汉神经机器翻译代词性别偏见的方法[J].厦门大学学报(自然科学版),2021,60(4):693-700.
**Citation**:SHI X W,HUANG H Y,JIAN P,et al. The method for reducing gender bias of pronouns in Uyghur to Chinese neural machine translation[J]. J Xiamen Univ Nat Sci,2021,60(4):693-700. (in Chinese)

表 1 原始与 NMT 生成的维汉数据集中不同极性代词的占比

Tab. 1 The proportion of different polatity pronouns in the original and NMT generated Uighur-Chinese dataset

| 数据 | 训练集 | | | 开发集 | | |
|---|---|---|---|---|---|---|
| | 阴性 | 阳性 | 阴性与阳性的比 | 阴性 | 阳性 | 阴性与阳性的比 |
| 原始数据集 | 2 771 | 16 550 | 1∶5.97 | 29 | 149 | 1∶5.14 |
| NMT 生成 | 1 053 | 16 438 | 1∶15.61 | 9 | 141 | 1∶15.67 |

建了用于评价机器翻译系统性别偏见情况的数据集 WinoMT. 类似地,Cho 等[13]针对以韩国语为源语言的机器翻译提出了性别偏见问题的评价指标,并构建了相应的评测数据. Saunders 等[15]将机器翻译中的性别偏见纠正问题视为领域适应问题,并构建了小规模的用于 NMT 领域自适应的语料,在 WinoMT[12]数据集上的实验结果表明,该方法有效地缓解了部分机器翻译的性别偏见问题.

在目标语言汉语端,只在第三人称代词上区分性别使用,因此需要考虑的情况相对于前人的工作[5,14-15]更加简化和具体. 本研究工作期望实现两个目的:1) 缓解 NMT 输出的目标语言端代词性别失衡问题;2) 利用有限的源语言信息自动地预测目标语言的代词性别,而不引入额外的控制信息.

经统计发现,语料中 99%的句子中使用的代词性别是一致的(即只出现阳性代词或只出现阴性代词),因此本文将正确预测某一个代词性别的问题简化为预测目标语言句子性别的问题. 首先,利用汉语单语语料构造伪数据对翻译训练数据进行扩展,引入的伪数据缓和了语料中代词性别不平衡的问题. 同时提出两种 NMT 模型显式融合性别信息的方法:1) 在目标语言句首引入性别标记,在不改动模型本身的架构的前提下在 NMT 模型中显式地融入了性别信息,而解码时,句首的性别标记可以约束后面生成代词的选择;2) 对机器翻译任务和目标语言代词性别预测任务联合建模,使 NMT 模型的编码器可以显式地学习代词性别的相关上下文信息.

## 1 性别偏见抑制方法

由于翻译语料通常以句子为单位,很少出现多个句子组成的段落或多个主语的情况,所以大部分训练集句子中代词的性别在句内是一致的. 通过对 CCMT 2019 维汉翻译训练集数据的统计发现,只包含单一性别代词的目标语言句子占所有包含代词的目标语言句子总数的 99.07%. 因此将词级别的代词性别预测问题简化为以句子为单位的性别预测问题. 这相当于只考虑两种特殊情况:1) 句子中所有代词均指代同一对象;2) 句子语境中只关于一种性别,即只包含一种性别. 这可避免引入指代消解等复杂的问题,同时也符合机器翻译语料以单个句子为主的内在特性.

### 1.1 伪数据构建

从表 1 可以看出,在 CCMT 2019 训练集数据中,不同性别的代词使用极不平衡,模型很容易受其影响. 与此同时,包含有代词的句子总数相对于总数据占比很少,机器翻译模型很难得到充足的训练,且缺少可以有效评价代词性别是否翻译正确的测试数据. 为解决上述问题,首先,从 CCMT 2019 的训练集数据中随机抽取出部分包含代词的数据,作为测试代词性别翻译准确性的开发集(PseDev)和测试集(PseTest)数据;然后利用语言数据联盟(linguistic data consortium, LDC)汉语语料和反向翻译方法[6]对其余的训练数据(ResTrain)进行扩展,并在此基础上构建了伪训练数据(PseTrain). 具体的 PseTrain 构造流程如图 1 所

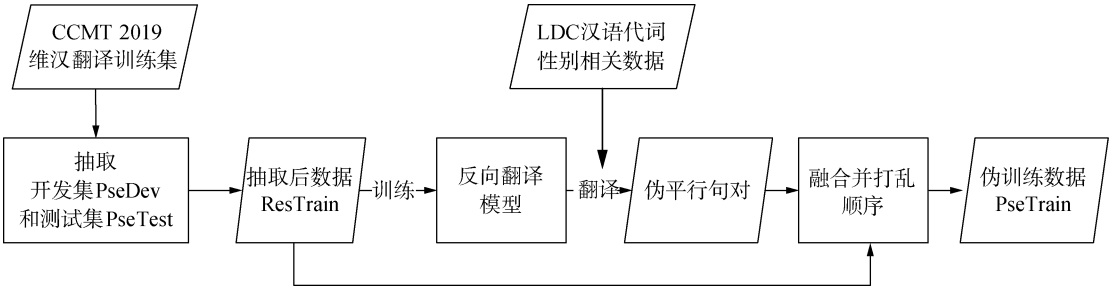


图 1 PseTrain 的构造流程示意图

Fig. 1 Schematic diagram of construction process of PseTrain

示:首先,利用 ResTrain 训练一个基于 Transformer[5] 的汉维翻译模型;然后,从汉语语料中随机抽取出数量相等的只包含阳性代词或阴性代词的汉语句子,并利用训练好的汉维翻译模型将其翻译成维语,构成伪平行句对;最后,将伪平行句对融入 ResTrain 中,并随机打乱顺序,得到 PseTrain. 扩展后的数据统计信息见 2.1 节中的表 2.

### 1.2　插入性别标志

在 NMT 中加入特殊标志是一种简单高效的融入附加信息的技术手段[7,20]. 这种方法不需要对模型进行改动,只需要在训练用的平行句对中加入所需要的标识符(token),即可达到融入附加信息的目的. 本文在目标语言端引入了一个表示性别信息的标志(GenTok),用以标识汉语端句子的性别信息. 引入 3 种性别标志:<NL>代表中性,<ML>代表阳性,<FL>代表阴性. 性别标志设置在目标语言句子的句首,可为其后面的词汇提供性别标志的信息,例如:

<NL>　木星 是 太阳系 中 质量 最 大 的 行星.

<FL>　她 当即 和 丈夫 决定:要 完成 儿子 的 遗愿,去 内蒙古 种 树 治沙.

<ML>　他们 无论 男女老少,都 有 一个 共同 的 名字——兵团 人.

### 1.3　性别预测与机器翻译联合建模

本文提出了性别预测与机器翻译联合建模的方法,将目标语言句子中使用的代词性别的分类融入到 NMT 模型的建模中. 融合性别预测后,NMT 模型输出的条件概率可表示为:

$$p(\boldsymbol{y},c \mid \boldsymbol{x};\theta) = p(c \mid \boldsymbol{x};\theta)\prod_{t=1}^{T_y} p(\boldsymbol{y}_t \mid \boldsymbol{y}_{<t},\boldsymbol{x};\theta). \tag{1}$$

其中:$\boldsymbol{x}$ 和 $\boldsymbol{y}$ 分别源端和目标端句子的向量表示;$\boldsymbol{y}_{<t}$ 为目标端句子前 $t-1$ 个词的向量表示;$T_y$ 为目标端句子的单词数;$c$ 代表性别类别,共 3 种,分别是中性、阴性和阳性;$\theta$ 是 NMT 模型和性别预测模型参数的集合.

将 NMT 模型中编码器最顶层的输出状态 $\boldsymbol{H} = [h_1, h_2, \cdots, h_{T_x}]$ 作为性别预测模型 $D_g$ 的输入,使用单层基于注意力机制的多分类器进行性别分类.

$$p(c \mid \boldsymbol{x};\theta) = p(c \mid \boldsymbol{H};\theta_g) = \mathrm{softmax}(\boldsymbol{W}_o\boldsymbol{s}), \tag{2}$$

其中,模型参数 $W_o \in \mathbf{R}^{3\times d}$,$\boldsymbol{s} \in \mathbf{R}^d$. 通过自注意力机制得到解码器状态:

$$\boldsymbol{s} = \boldsymbol{H} \times \mathrm{softmax}(\tanh(\boldsymbol{W}_a\boldsymbol{H} + b_a)), \tag{3}$$

其中,模型参数 $\boldsymbol{W}_a \in \mathbf{R}^{1\times d}$,偏置量 $b_a \in \mathbf{R}$. 对于分类器的损失函数,同样采用对数似然损失函数:

$$L_{\mathrm{gen}}(\boldsymbol{x},c;\theta) = -\log p(c \mid \boldsymbol{x};\theta). \tag{4}$$

最后,模型整体的损失函数为 $L = L_{\mathrm{nmt}} + \lambda L_{\mathrm{gen}}$. 其中:$L_{\mathrm{nmt}}$ 为 NMT 模型的损失函数;$\lambda$ 用来调节 $L_{\mathrm{gen}}$ 的占比,在本文中系数 $\lambda$ 设置为 0.1.

在进行翻译解码时,由于性别预测模型 $D_g$ 与 NMT 模型的解码器不产生关联,所以可以选择不执行性别预测步骤,这样模型解码的效率则同基线系统一致.

### 1.4　焦点损失函数

利用最大似然估计进行优化通常会受到类别不平衡问题的影响,即低频的类别由于在语料中占比小,造成该类别下的误差占比过小,影响模型训练的效果[21]. Lin 等[22] 提出了焦点损失函数(focal loss),用来缓解图像目标检测任务中前景与后景类别不平衡造成的负面影响. 本文参考焦点损失函数的方法,在式(4)中引入调节因子 $-(1-p)^{\gamma}$,其中 $\gamma \geqslant 0$ 称为聚焦参数. 将 $p(c \mid \boldsymbol{x};\theta)$ 简写为 $p$,这样负对数似然损失函数 $L_{\mathrm{gen}}$ 则被修改为 $L_{\mathrm{gen+focal}}$:

$$L_{\mathrm{gen+focal}} = -(1-p)^{\gamma}\log p. \tag{5}$$

当 $\gamma > 1$ 时,$p$ 越大则调节因子越趋近于 0,这样容易预测的样本产生的损失在训练中的占比就会缩小;相反地,难以预测的样本的损失函数值缩小的程度相对较轻.

## 2　实　验

### 2.1　数据集

本文的维汉双语数据来自于 CCMT 2019[8] 维汉新闻翻译任务,而数据扩展使用的汉语单语语料来自于 LDC 语料库. 语料的划分和扩充方法参见 1.1 节. 语料中的汉语部分首先采用 LTP 中文分词工具[23] 进行分词预处理;然后,维语和汉语部分均利用 Moses[24] tokenizer.perl 脚本进一步切分,其中维吾尔语部分采用了针对英文的切分规则;最后,分别对源语言和目标语言使用字节对编码(byte-pair encoding,BPE)[11]. 经过 BPE 处理后,在训练集上得到的词表(包含词和词切片)中维语约 2.7 万,汉语约 3.2 万. 采用抽取得到的 PseDev 作为开发集,测试数据采用 PseTest、第 14 届全国机器翻译大会(CWMT 2018)提供的开发集(CWMT 2018)以及 CCMT 2019 提供的开发集(CCMT 2019).

在构建伪数据时,采用的汉语语料来自于 LDC 语料库,具体的语料编号为 LDC2005T10、DC2003E14、

LDC2004T08 和 LDC2002E18. 上述语料中选取汉语中包含代词且代词性别在句内一致的句子作为候选. 最终选取的语料阴性和阳性比例为 1∶1.

实验中使用到的各部分数据集的统计信息如表 2 所示,其中阴性和阳性分别代表目标语言句子中只包含阴性代词和阳性代词,而中性则表示其他情况.

表 2 数据集统计

Tab. 2 The statistics of the datasets

| 数据集 | 句对 | 性别 | | |
|---|---|---|---|---|
| | | 阴性 | 阳性 | 中性 |
| ResTrain | 167 944 | 1 127 | 12 621 | 154 246 |
| PseTrain | 199 140 | 16 700 | 28 194 | 154 246 |
| PseDev | 1 000 | 500 | 500 | 0 |
| PseTest | 1 000 | 500 | 500 | 0 |
| CWMT 2018 | 1 000 | 11 | 84 | 905 |
| CCMT 2019 | 1 000 | 17 | 97 | 886 |

## 2.2 实验设置和评价指标

基线系统采用 6 层编码解码器构成的 Transformer[5] 作为 NMT 的基线模型,每层网络输出维度为 512,网络中的注意力模型包含 8 个注意力读写头,内部的全连接层神经元个数为 2 048. 为了验证扩展后得到的 PseTrain 的有效性,对比了在 ResTrain 上训练的基线模型,记为 Transformer(ResTrain). 训练时,将模型的失活率(dropout)设置为 0.3,采用 Adam 算法[25] 最优化模型,其中 $\beta_1=0.9$,$\beta_2=0.08$,$\epsilon=10^{-9}$,学习率的设置方法和自适应变化策略与 Transformer[5] 文献中所提到的方法一致,热启动训练步骤设置为 4 000.

+Balanced:Saunders 等[15] 提出的伪数据构建方法之一,扩展后的伪数据除包含有性别指示代词的原始数据外,还包含与原始数据对应的经过性别反转的数据. 根据 Saunders 等[15] 的方法,NMT 模型在原始数据集训练后,在伪数据上进行领域自适应调优. 本文用于调优的数据为含阳性代词的数据及其经过性别反转后得到的数据共同构建的性别无偏的数据,共 10 000 条.

+PosEdit:训练一个基于 2 层 Transformer[5] 架构的文本分类器 $D_g$,用于预测目标语言的代词性别,再利用预测的性别对翻译结果进行译后编辑. 分类器的输入为源语言句子,输出为目标语言性别. 分类器除层数外,其他设置均与 NMT 基线模型的编码器相同. 分类器在训练时,dropout 设置为 0.3,优化方法为 Adam 算法[25],其中 $\epsilon=10^{-9}$,初始学习率为 0.008. 分类模型采用焦点损失函数的损失缩放方法,$\gamma$ 设置为 2.

评价指标:本文机器翻译任务采用双语互译评估(bilingual evaluation understudy,BLEU)[26] 值作为评价指标,利用 Moses[27] 工具中 multi-bleu.pl 脚本作为打分工具. 为消除汉语分词对实验结果的影响,以汉字为单位进行打分(英文单词、阿拉伯数字等不进行拆分). 对于目标语言性别预测任务,采用准确率($P$)、召回率($R$)、$F_1$ 值作为评价指标.

## 2.3 机器翻译结果

表 3 为各模型在不同测试数据上的 BLEU 值.

对比表 3 的实验 1 和 4 可以发现,是否使用扩展后的语料对翻译模型进行训练,对 BLEU 值的影响并不大. 其中,在针对代词性别的测试数据 PseTest 上,使用扩展后的训练集会对翻译表现稍有提升,因为扩展的语料均采用为代词性别预测选取的数据.

Balanced 方法更适合对于名词阴阳性的降偏执而不是其逆过程,与本文的应用场景不十分吻合,故在本文的测试集上提升并不明显. 该方法的性别预测实验结果将在 2.4 节中给出.

表 3 机器翻译结果

Tab. 3 Machine translation results %

| 实验 | 模型 | BLEU 值 | | |
|---|---|---|---|---|
| | | Pse-Test | CWMT 2018 | CCMT 2019 |
| 1 | Transformer (ResTrain) | 29.50 | 48.44 | 33.59 |
| 2 | a+Balanced[15] | 29.57 | 48.45 | 33.54 |
| 3 | a+PosEdit | 29.53 | 48.43 | 33.59 |
| 4 | Transformer (PseTrain) | 29.73 | 48.17 | 33.90 |
| 5 | b+PosEdit | 29.73 | 48.15 | 33.89 |
| 6 | b+GenTok | 29.96 | 48.52 | 34.03 |
| 7 | b+ $L_{\text{gen}}$ | 29.93 | 48.50 | 33.92 |
| 8 | b+ $L_{\text{gen+focal}}$ | 29.92 | 48.72 | 33.98 |

注:a 表示 Transformer(ResTrain),b 表示 Transformer(PseTrain),下同.

实验 6~8 为本文提出直接融入目标语言性别信息的方法:GenTok 对 NMT 模型和训练方式没有改动,模型在生成过程中要优先生成目标语言的性别标记;$L_{\text{gen}}$ 和 $L_{\text{gen+focal}}$ 的解码过程则与 NMT 基线模型一

致.从翻译结果上看,本文提出的方法在 3 个测试集上的 BLEU 值均有微弱提升.对于以汉字为单位的评价指标,代词性别正确与否对整个数据集的 BLEU 值影响并不不大,因此可以看到在不同的实验设置下翻译任务上的表现相近.然而,由于训练数据和训练目标的不同,各个模型在代词性别预测上的表现大相径庭,代词性别预测任务的实验结果将在 2.4 节中给出.

### 2.4　代词性别预测结果

本节给出了各个实验设置下模型在目标语言性别预测上的实验结果.由于其他测试集包含的与代词性别相关的样本过少,所以实验仅在 PseTest 测试集上进行.表 4 分别给出了两种代词性别以及测试集总体的分类评价指标.由于测试集总体 $P$ 值和 $R$ 值相等,所以只给出 $P$ 值.

表 4　代词性别预测结果

Tab.4　Prediction results of pronoun gender　　%

| 实验 | 模型 | 阴性 | | | 阳性 | | | 总体 |
|---|---|---|---|---|---|---|---|---|
| | | $P$ | $R$ | $F_1$ | $P$ | $R$ | $F_1$ | $P$ |
| 1 | Transformer (ResTrain) | 80.95 | 3.40 | 6.52 | 51.97 | 71.20 | 60.08 | 37.30 |
| 2 | a+Balanced[15] | 68.29 | 28.00 | 39.72 | 60.00 | 60.60 | 60.30 | 44.30 |
| 3 | $D_g$ | 77.23 | 20.40 | 32.28 | 53.29 | 68.00 | 59.75 | 44.20 |
| 4 | Transformer (PseTrain) | 73.18 | 26.20 | 38.59 | 57.04 | 66.40 | 61.83 | 46.30 |
| 5 | b+GenTok | 75.43 | 26.40 | 39.11 | 59.46 | 66.60 | 62.03 | 46.50 |
| 6 | b+ $L_{\text{gen}}$ | 74.59 | 27.00 | 39.65 | 58.26 | 67.00 | 62.33 | 47.00 |
| 7 | b+ $L_{\text{gen+focal}}$ | 74.32 | 27.20 | 39.82 | 58.19 | 66.80 | 62.20 | 47.00 |

实验 1 采用的训练集代词性别偏差极大(阳性 12 621,阴性 1 127),因此模型解码输出以阳性代词为主,造成阴性代词的召回率极低.实验 4 的基线模型采用扩展语料进行训练,在代词性别预测的表现相对于实验 1 有大幅提升,证明引入代词性别平衡的伪数据确实可以改善性别偏见的问题.

Balanced 方法是 NMT 模型训练完成后,在人工构造的性别平衡的小规模数据上进行领域适应调优.从表 4 中可以看出,调优后的模型在性别平衡方面表现较好,阴性代词的召回率有较大的提升,同时阳性代词的召回率有所下降.但由于领域适应所用的数据是性别无偏的,该数据本身对代词性别推断的影响并不明确,这可能导致代词性别预测的精度损失.

实验 3 是译后编辑使用的分类器的结果,该分类器与实验 4～7 中的模型均在 PseTrain 训练集上训练.实验 3 中由于扩展后的语料性别比例仍有偏差,文本分类模型表现出比较明显的性别偏置问题.该现象说明单纯的文本分类更容易受到数据中性别不平衡的影响,$F_1$ 值在两种类别中差异很大.

除语料扩展外,本文提出的其他方法均对性别预测问题有所改善,实验 6 引入焦点损失函数对于阴性代词的预测有所提升,但是对比实验 6 和 7 可以发现二者差别并不明显.值得注意的是,通过在 PseDev 和 PseTest 数据集上的实验结果观察发现,无论是"+GenTok"还是"+$L_{\text{gen}}$"方法,输出的性别标记或类别与翻译出目标语言句子所使用的代词性别都是一致的,并不需要根据模型本身预测的性别标记或类别对目标语言句子进行译后编辑.

## 3　分　析

### 3.1　性别预测结果对翻译结果的影响

本节抽取测试集中含代词的语料进行测试,主要分析 NMT 模型翻译过程中,对目标语言代词使用情况的正确预测与否同翻译结果的 BLEU 值之间的关系.表 5 给出了在不同测试集下,代词性别使用正确和错误两种情况下机器翻译的 BLEU 值.其中 PseTest 只包含阴阳两种目标语言类别,而 CWMT 2018 和 CCMT 2019 中均包含 3 种情况,并以中性为主(表 2).从表 5 可以明显看出,被正确预测代词性别的翻译结果在整体上 BLEU 值明显高于目标语言性别预测错误的翻译.产生该现象的原因可能有两种:1) 使用了正确的代词性别,有利于对整体翻译质量的

提升;2) 容易得到高分翻译的源语言样本,同时也容易被预测出目标语言性别种类. 由于实验 1 和 2 中模型本身并没有显式的目标语言性别预测机制,所以认为产生该现象的原因更有可能是后者.

表 5 代词性别预测结果与翻译结果 BLEU 值之间的关系

Tab. 5 The relation between the prediction results of pronoun gender and the BLEU values of translation results %

| 实验 | 模型 | PseTest | | CWMT 2018 | | CCMT 2019 | |
|---|---|---|---|---|---|---|---|
| | | 正确 | 错误 | 正确 | 错误 | 正确 | 错误 |
| 1 | Transformer (ResTrain) | 35.29 | 25.04 | 49.37 | 34.32 | 39.50 | 23.81 |
| 2 | Transformer (PseTrain) | 32.29 | 27.36 | 48.88 | 34.45 | 39.88 | 25.66 |
| 3 | b+GenTok | 33.28 | 27.33 | 49.29 | 34.50 | 39.97 | 25.57 |
| 4 | b+$L_{\text{gen}}$ | 33.26 | 27.31 | 49.15 | 34.73 | 39.89 | 25.66 |
| 5 | b+$L_{\text{gen+focal}}$ | 33.28 | 27.30 | 49.32 | 34.61 | 39.91 | 25.78 |

## 3.2 翻译结果中的代词性别对比

表 6 给出了测试数据集和模型输出的翻译结果在表示不同性别的代词上的数量对比,其中实验 1 是数据集本身的统计信息,实验 2～6 为不同设置下模型输出的结果. 可以看出:翻译结果中两种性别的代词失衡问题在本文提出的方法(实验 3～6)中有所缓解. 在 PseTrain 数据集上,阴性阳性代词数量比例约为 1∶1.7,本文模型在 PseTest 测试集上得到的该比例约为 1∶3,说明模型仍存在放大性别偏见的问题.

表 6 翻译结果在代词使用上的对比

Tab. 6 Comparison of translation results in the use of pronouns

| 实验 | 数据集或模型 | PseTest | | | CWMT 2018 | | | CCMT 2019 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 阴性 | 阳性 | 比值 | 阴性 | 阳性 | 比值 | 阴性 | 阳性 | 比值 |
| 1 | 数据集 | 500 | 500 | 1∶1 | 11 | 84 | 1∶8 | 17 | 97 | 1∶6 |
| 2 | Transformer (ResTrain) | 21 | 685 | 1∶33 | 0 | 99 | 0 | 0 | 109 | 0 |
| 3 | Transformer (PseTrain) | 179 | 574 | 1∶3 | 22 | 99 | 1∶5 | 20 | 103 | 1∶5 |
| 4 | b+GenTok | 175 | 560 | 1∶3 | 14 | 88 | 1∶6 | 30 | 86 | 1∶3 |
| 5 | b+$L_{\text{gen}}$ | 181 | 575 | 1∶3 | 16 | 87 | 1∶5 | 28 | 87 | 1∶3 |
| 6 | b+$L_{\text{gen+focal}}$ | 183 | 574 | 1∶3 | 16 | 88 | 1∶5 | 31 | 90 | 1∶3 |

## 3.3 翻译案例

本节给出了一个翻译案例,如表 7 所示. 在该案例中,单从源语言信息容易推断句子中的代词指代对象为前文中提到的“母亲”,因此应该用阴性代词“她”. 在给出的翻译例子中,代词部分用粗体字标记. 可以看出,实验 4～7 采用了本文提出的方法,均在目标语言句子中使用了正确的代词,这说明在上下文信息明确的情况下,本文提出的方法均可以正确地预测出代词性别.

# 4 结 论

本文研究了维汉翻译中代词性别使用不平衡的问题,以 CCMT 2019 维汉新闻翻译任务的数据为研究样本,提出了 3 种缓解代词性别问题的方法:1) 选取与代词性别相关的汉语单语语料,采用反向翻译方法对原始训练集进行扩展;2) 在训练数据的目标语言端引入代词性别标记,在不改动模型的情况下,将性别信息显式地融入翻译任务中;3) 同时对机器翻译和代词性别预测建模,使模型同时学习翻译和性别预测任务. 构建了 PseTrain 维汉翻译伪训练数据,并将 CWMT 2018、CCMT 2019 以及本文构建的 PseTest 作为测试集以验证模型表现. 实验结果表明,通过对数据进行扩展,可以有效缓解维汉翻译中的代词性别偏见问题. 本文提出的显式融合性别知识方法可以进一步提升代词性别预测的精度.

表 7　翻译案例

Tab. 7　Case study of translation

| 实验 | 来源 | 句子 | BLEU/% |
|---|---|---|---|
| 1 | 源语言 | ئەدناغلاق‌پەيتىيادىناخرۇتخودىسىپاىمىتىقىرىب ، ىدرابپلاقويىنۇىڭاخرۇتخودىسىپاىمىتىقىرىب ، ىدىيىنەگىرىسنەيىنىدىتىمزىخىكىدلگاگۇايشلگىنڭاخرۇمۇمپەيتىيىلىسىكىسىپاى | |
| 2 | 参考译文 | 一次母亲住院,沈浩去看**她**,老母亲病中还在挂念着沈浩在小岗村的工作. | |
| 3 | Transformer (ResTrain) | 一次当母亲赶到医院时,申浩只好去看望**他**,母亲病了,也担心沈浩在小岗的工作. | 23.40 |
| 4 | Transformer (PseTrain) | 有一次母亲住在医院时,沈太太赶来看望**她**,母亲病也担心沈浩在小岗的工作. | 33.69 |
| 5 | b+GenTok | 有一次母亲住院,沈浩前往看望**她**,母亲病也担心了沈浩在小岗的工作. | 47.68 |
| 6 | b+$L_{gen}$ | 有一次母亲住院,沈浩前往看望**她**,母亲病了,也担心了沈浩在小岗的工作. | 46.16 |
| 7 | b+$L_{gen+focal}$ | 有一次母亲住院,沈浩去看望**她**,母亲也担心沈浩在小岗的工作. | 51.61 |

在未来的工作中,将考虑句子中出现不同性别的代词情况,将代词预测方法从句子级别改进为词级别.另外将在更多与维汉翻译相似的语言上对本文方法进行验证.在很多情况下,源语言可能并未包含足够的上下文信息用以判断代词的性别,希望在未来的工作中能识别出这些情况,以改善训练集和测试集的构建方法.

## 参考文献:


[1] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[EB/OL]. [2020-11-08]. https://arxiv.org/pdf/1409.3215.pdf.

[2] BAHDANAU D, CHO K H, BENGIO Y. Neural machine translation by jointly learning to align and translate[EB/OL]. [2020-11-08]. https://arxiv.org/pdf/1409.0473.pdf.

[3] WU Y H, SCHUSTER M, CHEN Z F, et al. Google's neural machine translation system: bridging the gap between human and machine translation[EB/OL]. [2020-11-08]. https://arxiv.org/pdf/1609.08144v1.pdf.

[4] GEHRING J, AULI M, GRANGIER D, et al. Convolutional sequence to sequence learning[EB/OL]. [2020-11-08]. http://arxiv.org/abs/1705.03122v3.

[5] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. [2020-11-08]. https://arxiv.org/pdf/1706.03762v5.pdf.

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

[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] YANG M, HU X, XIONG H, et al. CCMT 2019 machine translation evaluation report[M]// Communications in Computer and Information Science. Singapore: Springer, 2019: 105-128..

[9] KOEHN P, KNOWLES R. Six challenges for neural machine translation[C]// Workshop on Neural Machine Translation. Vancouver: ACL, 2017: 28-39.

[10] OTT M, AULI M, GRANGIER D, et al. Analyzing uncertainty in neural machine translation[EB/OL]. [2020-11-08]. https://arxiv.org/pdf/1803.00047.pdf.

[11] SENNRICH R, HADDOW B, BIRCH A. Neural machine translation of rare words with subword units[EB/OL]. [2020-11-08]. https://arxiv.org/pdf/1508.07909v2.pdf.

[12] LUONG M T, SUTSKEVER I, LE Q V, et al. Addressing the rare word problem in neural machine translation[EB/OL]. [2020-11-08]. https://arxiv.org/pdf/1410.8206.pdf.

[13] CHO W I, KIM J W, KIM S M, et al. On measuring gender bias in translation of gender-neutral pronouns[C]// Workshop on Gender Bias in Natural Language Processing. Florence: ACL, 2019: 173-181.

[14] STANOVSKY G, SMITH N A, ZETTLEMOYER L. Evaluating gender bias in machine translation[EB/OL]. [2020-11-08]. https://arxiv.org/pdf/1906.00591.pdf.

[15] SAUNDERS D, BYRNE B. Reducing gender bias in neural machine translation as a domain adaptation problem[EB/OL]. [2020-11-08]. https://arxiv.org/pdf/2004.04498.pdf.

[16] MICHEL P, NEUBIG G. Extreme adaptation for personalized neural machine translation[EB/OL]. [2020-11-08]. https://arxiv.org/pdf/1805.01817.pdf.

[17] VANMASSENHOVE E, HARDMEIER C, WAY A. Getting gender right in neural machine translation[C]// Conference on Empirical Methods in Natural Language Processing. Brussels: ACL, 2018: 3003-3008.

[18] RUDINGER R, NARADOWSKY J, LEONARD B, et al. Gender bias in coreference resolution[C]// Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans: ACL, 2018: 8-14.

[19] ZHAO J Y, WANG T L, YATSKAR M, et al. Gender bias in coreference resolution: evaluation and debiasing methods[C] // Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Techno-logies. New Orleans: ACL, 2018: 15-20.

[20] SENNRICH R, HADDOW B, BIRCH A. Controlling politeness in neural machine translation via side constraints [C]// Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego: ACL, 2016: 35-40.

[21] ANAND R, MEHROTRA K G, MOHAN C K, et al. An improved algorithm for neural network classification of imbalanced training sets[J]. IEEE Transactions on Neural Networks, 1993, 4(6): 962-969.

[22] LIN T Y, GOYAL P, GIRSHICK R B, et al. Focal loss for dense object detection[C] // IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, 2017: 2999-3007.

[23] CHE W X, LI Z H, LIU T. LTP: A chinese language technology platform[C] // International Conference on Computational Linguistics: Demonstrations. Beijing: ACL, 2010: 13-16.

[24] KOEHN P, HOANG H, BIRCHA, et al. Moses: open source toolkit for statistical machine translation[C]// Annual Meeting of the Association for Computational Linguistics. Prague: ACL, 2007: 177-180..

[25] KINGMA D P, BA J L. Adam: a method for stochastic optimization[EB/OL]. [2020-11-08]. https: // arxiv.org/pdf/1412.6980v9.pdf.

[26] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: a method for automatic evaluation of machine translation [C]// Annual Meeting of the Association for Computational Linguistics. Philadelphia: ACL, 2002: 311-318.

# The method for reducing gender bias of pronouns in Uyghur to Chinese neural machine translation

SHI Xuewen, HUANG Heyan, JIAN Ping*, TANG Yikun

(Beijing Engineering Research Center on High Volume Language Information Processing and Cloud Computing Applications, School of Computer Science &Technology, Beijing Institute of Technology, Beijing 100081, China)

**Abstract**: The gender insensitivity of pronouns in Uyghur faces a challenge for neural machine translation (NMT) models to translate Uyghur to Chinese accurately. Furthermore, significant bias of usage rate between pronouns exists in different genders in the training corpus, prompting NMT to generate pronouns in the male gender but proper gender. To circumvent these problems, we expand the original training corpus by constructing pseudo data with Chinese monolingual data. The gender bias in the new constructed training data becomes less obvious. We also introduce two branches of methods to incorporate gender prediction into NMT explicitly by adding a special gender token and modeling the gender prediction and NMT jointly. We conduct our experiments related to three Uyghur-to-Chinese translation test sets. Experimental results show that the proposed method performs with less gender bias without affecting the quality of translation and gains more satisfactory gender prediction results.

**Keywords**: neural machine translation; gender bias; pseudo data; gender prediction