维汉神经机器翻译代词性别偏见改进方法

史学文[1,2]，黄河燕[1,2]，鉴萍[1,2]*，唐翼琨[1,2]

（1. 北京理工大学计算机学院，北京市，100081；
2. 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，北京市，100081）

摘要：在利用神经机器翻译进行维吾尔语到汉语的翻译时，由于维吾尔语代词不区分性别，这给翻译模型在汉语端使用正确的代词带来了很大的挑战。另一方面，由于训练数据本身在不同性别的代词使用场景上频率的偏差，神经机器翻译倾向于输出阳性代词而不是更恰当的代词。为缓解上述问题，本文先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；之后，分别提出引入性别标记和翻译、性别预测联合建模两种方法，将代词性别预测目标显式地融入神经机器翻译模型的训练过程。我们在多个维汉翻译测试集上进行了实验验证，结果表明，我们提出的方法相对于基线系统，在不影响翻译总体效果的情况下缓解了神经机器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：TP391.2　　**文献标志码：** A

1 引言

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愈加受到机器翻译领域的关注，以及针对资源稀缺语言的 NMT 技术的发展[6,7]，NMT 也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，除了数据规模之外，语料中的一些偏置也会对翻译的结果产生影响[9,10]。特别地，对于维吾尔语 (以下简称“维”) 到汉语的机器翻译，源语言维语端代词不区分阴阳性，而目标语言汉语书写时，代词有阴阳性的区分（例如 “她/他”），这给维-汉机器翻译带来了很大的挑战。另一方面，机器翻译模型的训练数据本身在不同性别的代词使用场景上也有频率的偏差，这也增加了维-汉机器翻译的难度。以 CCMT2019 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）端出现的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。在表 1 中，对于 NMT，“训练集” 指利用 NMT 模型重新翻译训练集语料而得到的数据。从表 1 可以看出，在原始语料中，阴性代词与阳性代词的比例接近于 1：5，而在模型输出的翻译结果中，该比例缩小到约为 1：15，产生这种现象的原因主要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。

表 1 维-汉机器翻译原始数据集与 NMT 翻译结果在代词的使用上的对比

Tab.1 Comparison of the use of pronouns between the original dataset and NMT translation results on Uighur-Chinese dataset

数据	训练集			开发集		
	阴性	阳性	比值	阴性	阳性	比值
数据集	2,771	16,550	1:5.97	29	149	1:5.14
神经机器翻译生成	1,053	16,438	1:15.61	9	141	1:15.67

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

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[11-13]，并且，研究者们提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[11,12]。与前人工作不同，在本文中，一方面我们希望缓解宏观上因为不同性别相关的上下文出现频率的偏差造成的 NMT 性别偏见问题（在具体性别不明确的情况下倾向于使用阳性词汇的问题）；而另一方面，我们希望在微观上，利用具体的上下文与不同性别代词的共现关系，在不引入额外知识的情况下，推断应该选用的代词性别。

本文以 CCMT2019 维汉翻译任务提供的数据为例，研究了维汉翻译中代词性别偏见问题，并提出了缓解该问题的简单有效的方法。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），因此我们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，我们利用汉语单语语料构造伪数据对翻译训练数据进行了扩展，引入的伪数据缓和了语料中代词性别不平衡的问题。同时，我们提出了两种 NMT 模型显式融合性别信息的方法：i）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 模型中显式的融入了性别信息，而解码时，句首的性别标记可以约束后面生成代词的选择；ii）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。本文的实验所用的双语数据均来自 CCMT2019 维汉新闻翻译任务，测试数据除使用给出的开发集数据外，还为代词性别预测专门抽取了开发集和测试集。最终实验结果表明，我们使用的数据扩展方法和性别预测方法均有效缓解了维汉翻译中代词性别的偏见问题。

本文的主要贡献简要总结如下：

- NMT 性别偏见问题主要由训练数据的不平衡导致，因此，我们利用汉语单语数据和反向翻译模型对维汉翻译训练数据进行了微量的扩展，以平衡不同性别代词的使用。此外我们从训练数据中抽取出了用于评价代词性别预测效果的开发集和测试集。
- 汉语代词本身虽然具有区分性别的能力，但还兼具了其他的语法功能，因此，本文提出了在 NMT 模型中显示地融入性别预测任务的方法，使得 NMT 模型训练时可以明确关于性别预测部分的梯度。我们提出了两种方法，分别是 i）融入性别标志，和 ii）NMT 与性别预测联合建模。
- 我们在多个测试数据集上进行了实验验证。实验结果表明，与基线系统对比，我们提出的方法均可以有效改善维汉机器翻译代词性别偏见的问题：在不影响翻译表现的情况下，我们的方法在代词性别预测的召回率等指标上有显著的提升，尤其在原始语料中出现较少的阴性代词上提升明显。
- 我们对实验结果和样例进行了细致的分析，结果表明，尽管我们的方法对性别偏见问题有所缓解，但 NMT 模型仍会放大训练数据的性别偏差。另外，我们发现容易翻译的句子中，往往代词性别的使用也基本准确，我们认为其现象可能与数据分布有关。最后我们给出了一个翻译案例，简单说明了可以成功预测代词性别的原因。

2 相关工作

近年来，性别偏见问题在机器翻译领域逐渐引起重视[11-14]。首先，NMT 模型通常用大规模的语料进行训练，语料本身的特点结合模型的训练目标特性，均会影响 NMT 的翻译表现。Koehn et al.[9] 在关于 NMT 面临的若干挑战的报告中指出，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[15,16]。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 et al.[10] 在 Koehn et al.[9] 提出的问题的基础上针对数据集不确定性进行了更详细的研究，指出了性别等信息容易在翻译过程中丢失的问题。

另一方面，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使用正确语法的性别代词的问题。以维-汉翻译为例，维吾尔语中代词并不区分阴阳性，而汉语代词区分，这给代词的翻译带来了挑战。Michel et al.[14] 针对上述问题，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，实现调整性别、职业等相关信息翻译的目的。Vanmassenhove et al.[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 et al.[12] 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，并在 Winogender[18] 和 WinoBias[19] 构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 et al.[11] 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 et al.[13] 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，该方法有效缓解了部分机器翻译的性别偏见问题。

本文研究了维汉翻译性别偏见和代词性别预测问题。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[5,12,13] 更加简化和具体。我们希望通过我们的工作实现两个目的：i）缓解 NMT 输出的目标语言端代词性别失衡问题；ii）希望能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。

3 背景介绍

给定一个源语言序列 $\boldsymbol{x} = \{x_1, x_2, \dots, x_{T_x}\}$ 和对应的目标语言序列 $\boldsymbol{y} = \{y_1, y_2, \dots, y_{T_y}\}$，神经机器翻译模型（NMT）[2,5] 通常直接表示为下述条件概率建模：

$$p(\boldsymbol{y}|\boldsymbol{x};\theta) = \prod_{t=1}^{T_y} p(y_t|\boldsymbol{y}_{<t}, \boldsymbol{x};\theta) \tag{1}$$

其中 $\boldsymbol{y}_{<t}$ 表示 t 时刻之前的目标语言序列，θ 代表神经网络参数集合。在基于编码器解码器架构[20] 的神经机器翻译模型中，概率 $p(y_t|\boldsymbol{y}_{<t}, \boldsymbol{x};\theta)$ 的计算由解码器得到：

$$p(y_t|\boldsymbol{y}_{<t}, \boldsymbol{x};\theta) = Dec(y_{t-1}, s_t, c_t), \tag{2}$$

其中 y_{t-1} 表示解码器上一时刻的输出，而 s_t 和 c_t 则分别表示目标语言和源语言端的上下文向量。c_t 通常由注意力机制 [2,5] 在编码器的输出上计算得到。

NMT 模型通常采用最大似然估计的方法进行优化，给定训练集平行句对 $(\boldsymbol{x}, \boldsymbol{y})$，最大似然估计在单一平行句对下的损失函数定义为：

$$\mathcal{L}_{nmt}(\boldsymbol{x}, \boldsymbol{y}, \theta) = \sum_{t=1}^{T_y} -log\, p(y_t|\boldsymbol{y}_{<t}, \boldsymbol{x};\theta) \tag{3}$$

在本文中，所提出的方法适用于目前已有的基于编码器-解码器[20] 框架的 NMT 模型[2-5] 。为验证我们的模型的有效性，我们采用目前最先进的模型代表之一——Transformer [5]作为本文神经机器翻译的基线模型架构。Transformer 的编码器部分包含 6 层独立的网络结构，其中每一层网络均包含两部分：i）一个包含多感知头的注意力模型，以及 ii）一个位置相关的全连接前向网络。层与层之间有残差连接，每层输出前还需要经过层正则化操作[21] 。Transformer 的解码器部分同样包含 6 层独立的网络结构，与编码器中每一层不同的是，解码器还额外引入了一个子层，该子层用来在编码器的输出端执行多头的注意力机制，以获取源语言的上下文向量。本文的模型设置均以 Transformer base 作为参考，其中包括编码器解码器各 6 层，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 2048。

4 方法

由于翻译语料通常以单句为单位，很少出现多个句子组成的段落或多个主语的情况，因此大部分训练集句子中，代词的性别在句内是一致的。事实上，我们通过对 CCMT 2019 维汉翻译训练数据的统计发现，只包含单一性别的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%。基于上述发现，我们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。这相当于我们只考虑两种特殊情况：i) 句子中所有代词均指代同一对象；ii) 句子语境中只关于一种性别，即只包含一种性别。上述做法将某一个代词的性别预测简化成了对整个句子性别预测的问题，避免了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。

4.1 伪数据构建

从表 1 可以看出，在 CCMT2019 训练集数据中，不同性别的代词使用极不平衡，模型很容易受这种数据偏见影响。与此同时，包含有代词的句子总数占数据的总数很少，机器翻译模型很难得到充足的训练，另一方面，也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，我们从 CCMT2019 的训练数据中抽取出部分数据作为测试代词性别翻译准确性的开发集（PseDev）和测试集数据（PseTest）；之后，我们利用 LDC 汉语语料和反向翻译方法[6] 对其余的训练数据（ResTrain）进行了扩展，并在此基础上构建了伪训练数据（PseTrain）。具体的伪数据构造流程如图 1 所示：首先，利用数据 ResTrain 训练一个基于 Transformer[5]的汉-维翻译模型；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吾尔语，构成伪平行句对；最后，将伪平行句对融入 ResTrain 数据中，并随机打乱顺序，得到伪训练数据（PseTrain）。扩展后的数据统计信息见章节 5.1 中的表 2 。

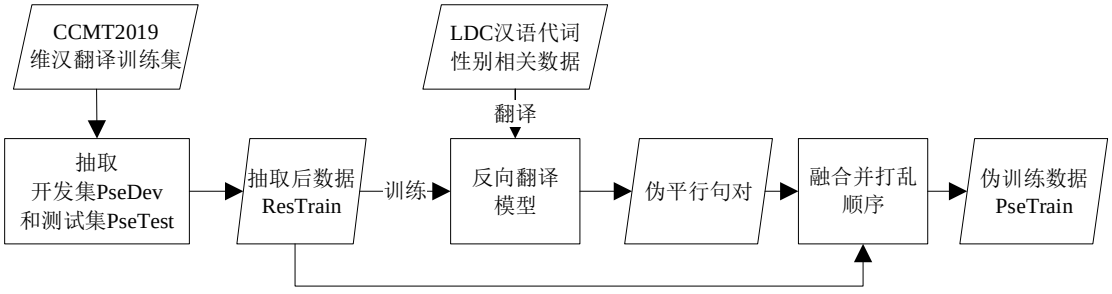

图 1 伪数据构造流程示意图

Fig. 1 Schematic diagram of pseudo data construction process

4.2 插入性别标志

在神经机器翻译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[7,22]。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。在本文中，我们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。我们引入三种性别标志：“<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，这样做可以为其后面的词汇提供性别标志的信息，图 2 给出了一个在平行句对中引入性别标记的例子。

<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 。

图 2 性别标志插入实例

Fig. 2 Example of the gender signs insertion

4.3 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，将目标语言句子使用的代词性别的分类融入到神经机器翻译模型的建模中，这样，公式 (1) 融合了性别预测后则改变为公式 (4) ：

$$p(\boldsymbol{y}, c|\boldsymbol{x};\theta) = p(c|\boldsymbol{x};\vartheta)\prod_{t=1}^{T_y} p(y_t|\boldsymbol{y}_{<t}, \boldsymbol{x};\theta) \tag{4}$$

其中 c 代表性别类别，共三种，分别是中性、阴性和阳性，θ 是神经机器翻译模型和性别预测模型参数的集合。

我们将神经机器翻译模型中编码器最顶层的输出状态 $H = \{h_1, \ldots, h_{T_x,}\}$ 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性性别分类。

$$p(c|\boldsymbol{x};\theta) = p\big(c\big|H;\theta_g\big) = \text{softmax}(W_o s), \tag{5}$$

其中，$W_o \in \mathbb{R}^{3\times d}$，$\text{s} \in \mathbb{R}^d$，通过自注意力机制得到：

$$s = H \times \text{softmax}(\tanh(W_a H + b_a)), \tag{6}$$

其中 $W_a \in \mathbb{R}^{1\times d}$，偏置量 $b_a \in \mathbb{R}$。对于分类器的损失函数，我们同样采用对数似然损失函数：

$$\mathcal{L}_{gen}(\boldsymbol{x}, c;\theta) = -log\, p(c|\boldsymbol{x};\theta)。 \tag{7}$$

最后，模型整体的损失函数为：$\mathcal{L} = \mathcal{L}_{nmt} + \lambda\mathcal{L}_{gen}$，其中 λ 用来调节 $\mathcal{L}_{gen}$ 的占比，在本文中系数 λ 设置为 0.1。

在进行翻译解码时，由于性别预测模型 D_g 与神经机器翻译模型的解码器不产生关联，因此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。

4.4 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[23]。Lin et al.[24]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。在本文中，我们参考焦点损失函数的方法，在公式 (7) 中引入了调节因子 $-(1-p)^\gamma$，其中 $\gamma \geq 0$ 称为聚焦参数。将 $p(c|\boldsymbol{x};\theta)$ 简写为 p，这样负对数似然损失函数 $\mathcal{L}_{gen}$ 则被修改为 $\mathcal{L}_{gen+focal}$：

$$\mathcal{L}_{gen+focal} = -(1-p)^\gamma \log p \ 。 \tag{8}$$

在公式(8) 中，当 γ 大于 1 时，p 越大，调节因子将越趋近于 0，这样容易预测的样本产生的损失在训练中的占比就会被缩小；相反地，难以预测的样本的损失函数值被缩小的程度相对较轻。在本文中，我们将 γ 设置为 2。

5 实验

5.1 数据集

本文的维-汉双语数据来自于 CCMT2019[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看章节 4.1。语料中的中文部分首先采用 LTP 中文分词工具[25]进行了分词预处理。之后，维吾尔语和汉语部分均利用 Moses[26] *tokenizer.perl* 脚本进行了进一步切分，其中我们对维吾尔语部分采用了针对英文的切分规则。最后，我们分别对源语言和目标语言使用字节对编码（byte-pair encoding, BPE）[15]。经过字节对编码处理后，我们在训练集上得到的词表（包含词和词切片）规模约为：维吾尔语 2.7 万，以及汉语 3.2 万。我们采用抽取得到的 PseDev 作为开发集，测试数据采用 PseTest、CWMT2018 提供的开发集（CWMT2018）以及 CCMT2019 提供的开发集（CCMT2019）。

在构建伪数据时，我们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 以及 LDC2002E18。上述语料中我们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 数据集统计信息

Tab.2 The statistics of the datasets

数据集	句对	阴性	阳性	中性
抽取后训练集 ResTrain	167,944	1,127	12,621	154,246
扩展后训练集 PseTrain	199,140	16,700	28,194	154,246
构造的开发集 PseDev	1,000	500	500	0
构造的测试集 PseTest	1,000	500	500	0
CWMT2018 开发集 CWMT2018	1,000	11	84	905
CCMT2019 开发集 CCMT2019	1,000	17	97	886

5.2 实验设置和评价指标

基线系统：我们采用 Transformer base[5] 作为神经机器翻译的基线模型。为了验证扩展后得到的伪训练数据（PseTrain）的有效性，我们还对比了在 ResTrain 上训练的基线模型，记为 Transformer+ResTrain。训练时，我们将模型的 dropout 设置为 0.3，采用 Adam 算法[27] 最优化模型，其中 $\beta_1 = 0.9$，$\beta_2 = 0.08$，$\epsilon = 10^{-9}$ ，初始学习率设置为 1.0。

+PosEdit: 我们训练了一个基于两层 Transformer[5] 架构的文本分类器 D_g，用于预测目标语言代词性别，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，其他设置均与神经机器翻译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[27] ，其中$\epsilon = 10^{-9}$ ，初始学习率为 0.008. 分类模型采用了 Focal Loss 的损失缩放方法，γ 设置为 2。

评价指标：本文机器翻译任务采用 BLEU[28] 作为评价指标，利用 Moses[26]工具中 *multi-bleu.pl* 脚本作为打分。为消除汉语分词对实验结果的影响，我们以汉字为单位进行打分（英文单词、阿拉伯数字等不拆分）。对于目标语言性别预测任务，我们采用精确率（P ）、召回率（R）、F_1 值作为评价指标。

5.3 机器翻译结果

本节将展示机器翻译任务的实验结果，各个实验设置在不同测试数据上的 BLEU 值在表 3 中给出。在表 3 中，“模型” 表示不同的实验设置，“PseTest”、“CWMT2018”、 “CCMT2019” 代表三个不同的测试数据。

在表 3 中对比第 1 行和第 3 行，我们发现，使用扩展后的语料对翻译的 BLEU 值产生的影响并没有很大。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。对比第 1 行和第 2 行以及第 3 行和第 4 行，我们发现利用句子分类器进行性别预测后进行译后编辑，对翻译结果负面影响居多，这可能是由于分类器本身的准确率决定的，具体的性别预测表现将在章节 5.4 中给出。

表 3 中第 5~7 行是本文提出直接融入目标语言性别信息的方法。第 5 行中的 “GenTok” 对神经机器翻译模型和训练方式没有改动，模型在生成过程中要优先生成出目标语言的性别标记。第 6 行“$\mathcal{L}_{gen}$”和第 7 行“$\mathcal{L}_{gen+focal}$” 的解码过程则与神经机器翻译基线模型一致。从翻译结果上看，我们提出的方法在三个测试集上的 BLEU 值均有所提升。对于以汉字为单位的评价指标，代词性别正确与否在 BLEU 值上的影响并不不大，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在章节 5.4 给出。

表 3 机器翻译结果

Tab.3 Machine translation results

#	模型	PseTest	CWMT2018	CCMT2019
1	Transformer（ResTrain）	29.50	48.44	33.59
2	+PosEdit	29.53	48.43	33.59
3	Transformer（PseTrain）	29.73	48.17	33.90
4	+PosEdit	29.73	48.15	33.89
5	+GenTok	29.96	48.52	34.03
6	+$\mathcal{L}_{gen}$	29.93	48.50	33.92
7	+$\mathcal{L}_{gen+focal}$	29.92	48.72	33.98

5.4 代词性别预测结果

本节我们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，所以我们的实验结果在 PseTest 测试集得出。表 4 分别给出了两种代词性别以及测试集整体的分类评价指标，对于测试集整体精确率和召回率相等，故只给出精确率指标。

在表 4 中，第 1 行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。而第 3 行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第 1 行有大幅提升，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。

第 2 行是译后编辑使用的分类器的结果，该分类器与第 3~6 行中模型均在 PseTrain 训练集上训练。从实验结果可以看出，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了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，我们提出的其他方法均对性别预测问题有所改善，第 6 行引入焦点损失函数对于

阴性代词的预测有所提升，但是比较第 5 行和第 6 行可以发现差别并不明显。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 “+$\mathcal{L}_{gen}$” 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。

表 4 代词性别预测结果

Tab.4 Prediction results of pronoun gender

#	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	D_g	77.23	20.40	32.28	53.29	68.00	59.75	44.20
3	Transformer（PseTrain）	73.18	26.20	38.59	57.04	66.40	61.83	46.30
4	+GenTok	75.43	26.40	39.11	59.46	66.60	62.03	46.50
5	+$\mathcal{L}_{gen}$	74.59	27.00	39.65	58.26	67.00	62.33	47.00
6	+$\mathcal{L}_{gen+focal}$	74.32	27.20	39.82	58.19	66.80	62.20	47.00

6 分析

6.1 性别预测结果对翻译结果影响

本节我们主要分析了神经机器翻译模型翻译过程中，对目标语言代词使用情况的正确预测与否与翻译结果的 BLEU 值之间的关系。表 5 给出了在不同测试集下，在代词性别使用正确和错误两种情况下机器翻译的 BLEU 得分。其中 PseTest 只包含阴阳两种目标语言类别，而 CWMT2018 和 CCMT2019 中均包含三种情况，并以中性为主（参见表 2 ）。从表 5 中可以明显看出，被正确预测代词性别的翻译结果在整体上 BLEU 得分远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：i）使用了正确的代词性别有利于对整体翻译质量的提升；ii）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。由于第 1 行和第 2 行中模型本身并没有显式的目标语言性别预测机制，因此我们认为，产生该现象的原因更有可能是（ii）。

表 5 代词性别预测结果与机器翻译结果的关系

Tab.5 The relation between the prediction results of pronoun gender and the machine translation performances

#	数据	PseTest		CWMT2018		CCMT2019	
		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	+GenTok	33.28	27.33	49.29	34.50	39.97	25.57
4	+$\mathcal{L}_{gen}$	33.26	27.31	49.15	34.73	39.89	25.66
5	+$\mathcal{L}_{gen+focal}$	33.28	27.30	49.32	34.61	39.91	25.78

6.2 翻译结果代词性别对比

在本节中，我们展示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目对比，在表 6 中给出。其中第 1 行是数据集本身的统计信息，第 2~6 行为不同设置下模型输出的结果。从

表 6 中可以看出，翻译结果中两种性别的代词失衡问题在我们提出的方法（第 3~6 行）中有所缓解。在 PseTrain 数据集上，阴性阳性代词数量比例约等于 1:1.7，我们的模型在 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 翻译结果在代词的使用上的对比

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

#	数据	PseTest			CWMT2018			CCMT2019		
		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	109	-
3	Transformer（PseTrain）	179	574	1:3	22	99	1:5	20	103	1:5
4	+GenTok	175	560	1:3	14	88	1:6	30	86	1:3
5	$+\mathcal{L}_{gen}$	181	575	1:3	16	87	1:5	28	87	1:3
6	$+\mathcal{L}_{gen+focal}$	183	574	1:3	16	88	1:5	31	90	1:3

6.3 翻译案例

在本节，我们给出了一个翻译案例，如图 3 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，代词部分用加粗红色字体标记，对应的句子级别 BLEU 值在括号中给出。可以看出，第 4~7 行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，我们提出的方法均可以正确地预测出代词性别。

1	源语言	بىر قېتىم ئاپىسى دوختۇرخانىدا يېتىپ قالغاندا ، شېن خاۋ ئۇنى يوقلاپ باردى ، ئاپىسى كېسەل يېتىپمۇ شېن خاۋنىڭ شياۋگاڭدىكى خىزمىتىدىن ئەنسىرىگەن ئىدى
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 。（BLEU：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 。（BLEU：33.69）
5	+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 。（BLEU：47.68）
6	$+\mathcal{L}_{gen}$	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 。（BLEU：46.16）
7	$+\mathcal{L}_{gen+focal}$	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 。（BLEU：51.61）

图 3 翻译案例

Fig. 3 Case study of translation

7 总结

本文研究了维吾尔语-汉语翻译中代词性别使用不平衡的问题，以 CCMT2019 维-汉新闻翻译任务的数据为研究样本，提出了三种缓解代词性别问题的方法：i）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；ii）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；iii）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。我们构建了 PseTrain 维-汉翻译伪训练数据，并将 CWMT2018、CCMT2019 以及我们自己构建 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。我们提出的显式融合性别知识方法可以进一步地提升代词性别预测的精度。

在未来的工作中，我们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。另外我们将在更多与维-汉翻译相似的语言上对我们的方法进行验证。另外，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，我们希望在未来的工作中能识别出这些情况，以改善训练集、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[C]. Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada. [S.l.: s.n.], 2014: 3104–3112.

[2] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [C]. 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. [S.l.: s.n.], 2015.

[3] Wu Y, Schuster M, Chen Z, et al. Google’s neural machine translation system: Bridging the gap between human and machine translation[J]. CoRR, 2016, abs/1609.08144.

[4] Gehring J, Auli M, Grangier D, et al. Convolutional sequence to sequence learning[C]. Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017. [S.l.: s.n.], 2017: 1243–1252.

[5] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]. Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA. [S.l.: s.n.], 2017: 6000 – 6010.

[6] Sennrich R, Haddow B, Birch A. Improving neural machine translation models with monolingual data[C]. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers. The Association for Computer Linguistics, 2016.

[7] Johnson M, Schuster M, Le Q V, et al. Google’s multilingual neural machine translation system: Enabling zero-shot translation[J]. TACL, 2017, 5: 339–351.

[8] Yang M, Hu X, Xiong H, et al. Ccmt 2019 machine translation evaluation report[M]. [S.l.: s.n.], 2019.

[9] Koehn P, Knowles R. Six challenges for neural machine translation[C]. Proceedings of the First Workshop on Neural Machine Translation, NMT@ACL 2017, Vancouver, Canada, August 4, 2017. [S.l.: s.n.], 2017: 28–39.

[10] Ott M, Auli M, Grangier D, et al. Analyzing uncertainty in neural machine translation[C]. Proceed- ings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018. [S.l.: s.n.], 2018: 3953–3962.

[11] Cho W I, Kim J W, Kim S M, et al. On measuring gender bias in translation of gender-neutral pronouns[C]. Proceedings of the First Workshop on Gender Bias in Natural Language Processing. Florence, Italy: Association for Computational Linguistics, 2019: 173–181.

[12] Stanovsky G, Smith N A, Zettlemoyer L. Evaluating gender bias in machine translation[C]. Proceed- ings of the 57th

Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers. [S.l.: s.n.], 2019: 1679–1684.

[13] Saunders D, Byrne B. Reducing gender bias in neural machine translation as a domain adaptation problem[C]. Jurafsky D, Chai J, Schluter N, et al. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020. [S.l.]: Association for Computational Linguistics, 2020: 7724–7736.

[14] Michel P, Neubig G. Extreme adaptation for personalized neural machine translation[C]. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers. 2018: 312–318.

[15] Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units[C]. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers. [S.l.: s.n.], 2016.

[16] Luong T, Sutskever I, Le Q V, et al. Addressing the rare word problem in neural machine translation [C/OL]. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers. The Association for Computer Linguistics, 2015: 11–19. https://doi.org/10.3115/v1/p15-1002.

[17] Vanmassenhove E, Hardmeier C, Way A. Getting gender right in neural machine translation[J]. CoRR, 2019, abs/1909.05088.

[18] Rudinger R, Naradowsky J, Leonard B, et al. Gender bias in coreference resolution[C/OL]. Walker M A, Ji H, Stent A. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers). Association for Computational Linguistics, 2018: 8–14. https://doi.org/10.18653/v1/n18-2002.

[19] Zhao J, Wang T, Yatskar M, et al. Gender bias in coreference resolution: Evaluation and debi- asing methods[C/OL]. Walker M A, Ji H, Stent A. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Tech- nologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers). Association for Computational Linguistics, 2018: 15–20. https://doi.org/10.18653/v1/n18-2003.

[20] Cho K, van Merrienboer B, Gülçehre Ç, et al. Learning phrase representations using RNN encoder- decoder for statistical machine translation[C]. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meet- ing of SIGDAT, a Special Interest Group of the ACL. [S.l.: s.n.], 2014: 1724–1734.

[21] Ba L J, Kiros R, Hinton G E. Layer normalization[J]. CoRR, 2016, abs/1607.06450.

[22] Sennrich R, Haddow B, Birch A. Controlling politeness in neural machine translation via side con- straints[C]. Knight K, Nenkova A, Rambow O. NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Tech- nologies, San Diego California, USA, June 12-17, 2016. [S.l.]: The Association for Computational Linguistics, 2016: 35–40.

[23] Anand R, Mehrotra K G, Mohan C K, et al. An improved algorithm for neural network classification of imbalanced training sets[J]. IEEE Trans. Neural Networks, 1993, 4(6): 962–969.

[24] Lin T, Goyal P, Girshick R B, et al. Focal loss for dense object detection[C]. IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017. [S.l.]: IEEE Computer Society, 2017: 2999–3007.

[25] Che W, Li Z, Liu T. LTP: A chinese language technology platform[C]. COLING 2010, 23rd In- ternational Conference on Computational Linguistics, Demonstrations Volume, 23-27 August 2010, Beijing, China. [S.l.: s.n.], 2010: 13–16.

[26] Koehn P, Hoang H, Birch A, et al. Moses: Open source toolkit for statistical machine translation [C]. ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic. [S.l.: s.n.], 2007.

[27] Kingma D P, Ba J. Adam: A method for stochastic optimization[C/OL]. Bengio Y, LeCun Y. 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. 2015. http://arxiv.org/abs/1412.6980.

[28] Papineni K, Roukos S, Ward T, et al. Bleu: a method for automatic evaluation of machine translation [C]. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA. [S.l.: s.n.], 2002: 311–318.

Reducing Gender Bias of Pronouns in Uyghur to Chinese Neural Machine Translation

SHI Xuewen[1, 2], Huang Heyan[1, 2], JIAN Ping[1, 2] *, TANG Yi-Kun[1, 2]

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

Abstract: Pronouns in Uyghur are gender-insensitive, which makes a challenge for neural machine translation (NMT) models to translate Uyghur into Chinese accurately. On the other hand, there are significant bias between pronouns in different grammatical genders in the training corpus, and it makes NMT tends to generate pronoun in the gender of male. In this paper, to alleviate the above problems, we expand the original training corpus by constructing a pseudo data with Chinese monolingual data. The gender bias in the new constructed training data is less obvious. We also introduce two branches of methods to incorporate gender prediction into NMT explicitly: 1) adding a special gender token, and 2) modeling gender prediction and NMT jointly. We conduct our experiments on three Uyghur-to-Chinese translation test sets. The experimental results show that the proposed method performs less gender bias in machine translation and gains better gender prediction results.

Keywords: neural machine translation, gender bias, pseudo data, gender prediction