基于情感语义增强编解码的神经机器翻译方法

万　飞
(合肥工业大学 管理学院,安徽 合肥 230009)

摘　要:针对目前神经机器翻译模型仅依赖平行语料训练而无法充分挖掘深层语言知识的问题,提出一种基于情感语义增强编解码的神经机器翻译方法,旨在通过引入额外的情感语义,提高模型对语言深层次信息的理解能力。首先,利用 word2vec 技术获取语料中所有单词的词嵌入,将其输入到一个融合模型中进行训练。该融合模型结合了基于 GRU 和文档嵌入的机制,以获取单词级别和文档级别的情感语义表征;其次,在情感融合阶段,采用加权公式将单词级别和文档级别的情感语义有机地融合,形成更为综合的情感语义表征;最后,将此表征与上下文语义表征按位相加,以全面引入情感信息,并将其作为输入传递到机器翻译模型的编码器和解码器中。在多个基准数据集上的实验显示,相较于传统的 Transformer 模型,该方法在 IWSLT 数据集上性能显著提升,BLEU 值增加 1.3 至 1.62。在 WMT 数据集上也取得良好性能,证实了融合情感语义在机器翻译中的有效性。

关键词:情感语义;增强编解码;神经机器翻译;Transformer;平行语料
中图分类号:TP391　　**文献标识码**:A　　**文章编号**:1673-629X(2024)09-0094-08
doi:10.20165/j.cnki.ISSN1673-629X.2024.0159

Neural Machine Translation Method Based on Emotional Semantics Enhanced Encoding and Decoding

WAN Fei
(School of Management, Hefei University of Technology, Hefei 230009, China)

Abstract: To address the problem that current neural machine translation models rely solely on parallel corpus training and cannot fully tap into deep linguistic knowledge, we propose a neural machine translation method based on emotional semantic enhancement coding and decoding to improve the model's ability to understand deep linguistic information by introducing additional emotional semantics. Firstly, word2vec technology is used to obtain word embeddings for all words in the corpus, which are then input into a fusion model for training. This fusion model combines mechanisms based on GRU and document embedding to obtain emotional semantic representations at the word and document levels. Secondly, in the emotional fusion stage, a weighted formula is used to integrate the emotional semantic representations at the word and document levels organically, forming a more comprehensive emotional semantic representation. Finally, this representation is added bitwise with contextual semantic representations to fully introduce emotional information, and passed as input to the encoder and decoder of the machine translation model. Experiments on multiple benchmark datasets show that compared to traditional Transformer models, the proposed method significantly improves performance on the IWSLT dataset, with BLEU values increasing by 1.3 to 1.62. It also achieves good performance on the WMT dataset, confirming the effectiveness of integrating emotional semantics in machine translation.

Key words: emotional semantics; enhanced encoding and decoding; neural machine translation; Transformer; parallel corpus

0　引　言

随着深度学习技术的不断进步和计算资源的显著提升,神经机器翻译(Neural Machine Translation, NMT)方法已成为当前机器翻译领域的主导范式[1]。然而,传统的 NMT 方法局限于使用双语平行语料,仅对词级别的关系进行建模,未能充分利用语言学先验知识。不同语言中的等价词通常具有相似的语义分布[2],平行语料中的源语和目标语之间也存在共同的情感语义,但主流基于注意力机制的 NMT 模型往往未能充分挖掘其中的情感信息。以英文句子“He's a sweet salesman”为例,其翻译可能是“他是个糖果推销员”或“他是个和蔼可亲的推销员”。如果知道源语句

收稿日期:2023-12-28　　**修回日期**:2024-04-30
基金项目:安徽省高校自然科学研究重点项目(KJ2021A1253)
作者简介:万　飞(1988-),男,博士研究生,CCF 会员(T0915M),研究方向为图计算、自然语言处理。

子所包含的情感语义是“工作辛苦”,则模型更可能选择前者的译文;而如果情感语义是“顾客满意”,则模型可能更倾向于选择后者的译文。因此,将情感语义融入神经机器翻译中,有助于使模型能够根据不同语境正确地进行翻译。然而,在平行语料资源有限的情况下,NMT 模型往往难以充分训练,无法捕捉到复杂语料中隐藏的情感信息。因此,利用外部的情感分析语料训练分类模型,并将其应用于平行语料以获取情感语义信息,从而对 NMT 模型进行情感语义增强,成为提高机器翻译质量的有效途径。

1　相关工作

神经机器翻译将一个长度不确定的源语言序列映射为另一个长度未知的目标语言序列,通常采用编码-解码架构。该架构的核心理念在于,通过编码器将源语料编码成一个固定长度的向量,并利用解码器将该向量解码成目标语料。随着研究的不断深入,关于编码-解码架构的改进方法也日益涌现。其中,注意力机制[3]被引入到编码-解码架构中,取代了传统的使用固定编码器输出向量的方式,极大地提升了翻译性能。例如,Gehring 等[4]采用基于卷积神经网络的编解码器,显著提高了模型的训练速度。此外,李梦洁等[5]提出在机器翻译过程中引入注意力机制,以解决输入序列的长距离依赖问题。另一项重要的突破是由 Vaswani 等提出的 Transformer 模型[6],该模型完全采用自注意力机制和交叉注意力机制,摒弃了传统的 RNN 编码方式,极大地提升了机器翻译的质量。

然而,Transformer 模型仅依赖平行语料进行训练,忽略了语言具有的语法、句法等信息,从而导致翻译出现语义表征的偏差。杨丹等[7]指出,在平行语料匮乏的情况下,机器翻译的效果较差,并强调采用同义词替换进行数据增强的重要性。另外,亢晓勉等[8]指出,目前仅依赖平行语料的上下文信息的机器翻译方法存在局限,难以充分挖掘篇章级别的结构化语义信息。

因此,当前针对 Transformer 方法改进的重要研究方向之一是如何更深入地挖掘平行语料中的语义信息。一种常见的方法是通过句法分析来提取平行语料中的语义信息。例如,王振晗等[9]利用深度优先遍历获取源语言句法解析树的向量表示,并将句法向量与源语言词嵌入相加作为输入,用于训练翻译模型。除此之外,构建语义图并利用图神经网络技术来挖掘图中蕴含的语义信息也成为当前的研究热点。例如,薛媛[10]采用图循环神经网络对源语生成的抽象语义表示(Abstract Meaning Representation, AMR)图进行编码,并验证了将 AMR 语义知识纳入模型有助于提高汉蒙机器翻译的质量。普浏清等[11]引入图编码器,对源语的依存结构图进行向量化编码,以引导模型在解码过程中生成翻译。在计算机视觉领域特征提取技术不断发展的背景下,将视觉信息与 NLP 文本信息相融合成为了新的研究热点。例如,黄鑫等[12]提出将完整的图片输入到翻译模型中,显式对齐文本与图像中的实体,达到实体级的跨模态语义融合。这些方法的探索为解决 NMT 中语义信息挖掘的难题提供了新的途径和有效策略。

情感语义作为语义表征的一种形式,已经被广泛研究,并证明将其作为词嵌入表征的额外补充可以有效提升 NLP 相关任务上的模型性能[13-14]。例如,朱星浩等[13]提出通过关联度计算自然语言和音乐的情感语义,从而更有效地提供更精准的音乐和文本检索以及推荐服务。徐月梅等[14]引入源语言的情感监督信息以获得源语言情感感知的词向量表示,用于跨语言文本的情感预测。尽管当前的研究取得了一定的进展,然而在源语言情感资源匮乏以及细粒度情感任务的场景下,仍然缺乏有效的解决途径。

针对上述问题,该文在 Transformer 方法的基础上融合了额外的情感语义表征。首先训练了一种结合 GRU 和文档嵌入的融合模型。该融合模型旨在推断平行语料中存在的细粒度情感,并将源语和目标语中的情感语义映射到不同的语料空间。随后,将映射得到的情感语义与原始词嵌入进行融合,以实现对神经机器翻译模型的细粒度情感语义增强。主要贡献包括以下三个方面:

(1)提出了一种基于情感语义增强编解码的神经机器翻译方法(Emotional Semantic Enhancing Encoding and Decoding,ESEED)。该方法引入情感语义以优化源语料的表征,并深入研究了情感语义对 NMT 模型性能的影响。

(2)在不同情感粒度以及不同情感强度下,对该方法进行了性能探索,并对可能的原因进行了深入分析。

(3)通过在公开数据集 IWSLT 和 WMT 上进行多组实验,证明该方法在不引入额外训练参数的情况下显著提升了 Transformer 方法的性能。

2　模型介绍

2.1　模型整体结构

图 1 展示了 ESEED 模型的整体结构。相较于 Transformer 模型,该模型包含以下两个方面的改进:

(1)引入额外的情感分析语料。通过训练基于 GRU 和文档嵌入的融合模型,从平行语料中挖掘情感语义。

(2)在源语和目标语的词嵌入空间分别生成源语和目标语的情感语义编码,然后将这些编码与源语/目标语词嵌入以及相应位置编码相加,实现对编码器和解码器中输入表征的情感语义增强。

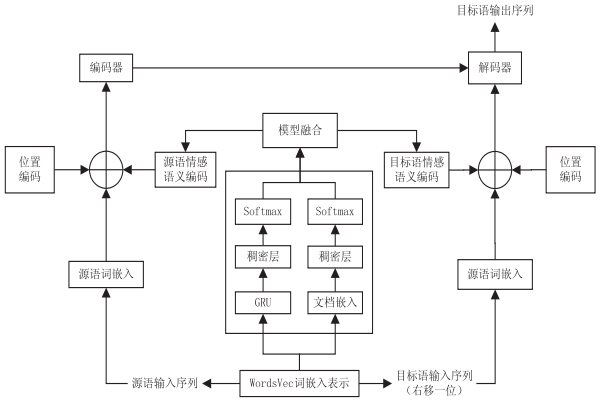

图 1 模型整体结构

2.2 平行语料情感语义表征获取

基于 Seq2Seq 结构的 Transformer[6] 被认为是当前性能最优的 NMT 方法之一。该方法仅依赖于自注意力机制和交叉注意力机制,通过自动学习平行语料的上下文信息进行翻译。然而,尽管在自动化学习方面取得了巨大成功,但这种方法缺乏句法和语义知识的直接引导,在语料规模较小的低资源场景下,可能会导致译文质量的下降。

为了解决这一问题,该文利用德英翻译和英越翻译中的英文情感分类语料,训练了一个融合单词级别和文档级别情感分类的模型。该融合模型被用来显式地指导 Transformer 获取情感语义表征,提升在低资源情况下的翻译质量。

假设情感分类的输入语料为 $S=\{x_1,x_2,\cdots,x_i,\cdots,x_n\}$,其中 x_i 表示语料 S 中的第 i 个单词。为了降低计算复杂度,每个单词通过加载静态预训练模型 word2vec 获取词嵌入。

对于输入语料 S 中的每个单词,被映射为 $S=\{e_1,e_2,\cdots,e_i,\cdots,e_n\}$,其中 e_i 表示语料 S 中的第 i 个词嵌入,将 S 输入 GRU[15](Gated Recurrent Units,门控循环单元)进行语料的特征提取,如公式 1 所示。GRU 与 LSTM[16] 在性能上相似,但将 LSTM 中的遗忘门和候选门合并为一个更新门,减少了模型的参数量,进而提升了训练速度。通过 GRU 获取每个时刻隐藏层的输出,如公式 1 所示。

$$h_i=\mathrm{GRU}(h_{i-1},e_i) \tag{1}$$

其中,h_i 表示第 i 个时刻 GRU 的隐藏层输出,e_i 表示第 i 个时刻的词嵌入,即第 i 个时刻 GRU 的输入。

分析整个语料的情感需要综合考虑每个单词表达的情感。因此,对 GRU 的所有隐藏层输出进行展平操作,输入到一个稠密层中,进一步提取相对低维的情感特征。这个过程可以表示为公式 2。

$$\mathrm{Dense}(W_h,\theta_h)=\boldsymbol{W}_h[\mathrm{flatten}(\mathrm{concat}(h_0,h_1,\cdots,h_i))]+\theta_h \tag{2}$$

其中,$\boldsymbol{W}_h$ 表示稠密层的权重矩阵,θ_h 表示稠密层的偏置项,flatten 是深度神经网络中对张量的展平操作。

将提取的低维情感特征输入到单词级别的 softmax 层,预测输入语料的情感倾向,如公式 3 所示。

$$P(T\mid e)=\mathrm{softmax}(\mathrm{Dense}(W_h,\theta_h)\boldsymbol{W}_s+\theta_s) \tag{3}$$

其中,$P(T\mid e)$ 表示当前输入语料通过 GRU 模型得到的情感概率分布,$\boldsymbol{W}_s$ 表示 softmax 层的权重矩阵,θ_s 表示 softmax 层的偏置项。

训练过程中使用交叉熵损失函数计算 GRU 模型的 loss 值,即比较模型的预测情感概率分布与语料的真实情感概率分布之间的距离,如公式 4 所示。

$$L_{\mathrm{word}}=-\frac{1}{|K_b|}\sum_{i=1}^{|K_b|}\sum_{T\in[-1,1]}P(T\mid h_i)\log P(T\mid e_i) \tag{4}$$

其中,L_{word} 表示一个 batch 的语料中单词级别的交叉熵损失,$|K_b|$ 表示一个 batch 的语料中单词的数目。T 表示情感极性,等于 1 时表示积极情感,等于-1 时表示消极情感。每个单词 h 的真实情感分布概率 $P(T\mid h)$ 通过源语生成的词典计算得到,h_i 表示当前 batch 的第 i 个单词,e_i 表示模型预测的第 i 个单词嵌入,$P(T\mid e)$ 表示单词通过模型预测得到的情感分布概率。

对于输入语料 $S=\{e_1,e_2,\cdots,e_n\}$,设 e_d 表示当前语料的文档嵌入,则 e_d 是对所有输入的嵌入表示求均值,如公式 5 所示。

$$e_d=\frac{1}{|K|}\sum_{i=1}^{|K|}e_i \tag{5}$$

其中,$|K|$ 表示每条语料 S 中单词的数目,e_i 表示语料中每个单词对应的词嵌入。

获得的文档嵌入表示语料的整体情感倾向,但由于其维度较高,需要将其输入一个稠密层进行降维和特征提取,如公式 6 所示。

$$\mathrm{Dense}(W_d,\theta_d)=\boldsymbol{W}_d*e_d+\theta_d \tag{6}$$

其中,$\boldsymbol{W}_d$ 表示稠密层的权重矩阵,θ_d 表示稠密层的偏置项。

将提取的低维情感特征输入到文档级的 softmax 层,预测输入语料的情感倾向,如公式 7 所示。

$$P(T\mid e_d)=\mathrm{softmax}(\mathrm{Dense}(W_d,\theta_d)\boldsymbol{W}_{s'}+\theta_{s'}) \tag{7}$$

其中,$P(T\mid e_d)$ 表示当前输入语料通过文档级情感分类模型得到的情感概率分布,$\boldsymbol{W}_{s'}$ 表示 softmax 层的权重矩阵,$\theta_{s'}$ 表示 softmax 层的偏置项。

训练过程中使用交叉熵损失函数计算文档级情感

分类模型的 loss 值,即比较模型的预测情感概率分布与语料的真实情感概率分布之间的距离,如公式 8 所示。

$$L_{\text{doc}} = -\frac{1}{|K_d|}\sum_{i=1}^{|K_d|} P(T \mid h_{di}) \log P(T \mid e_{di}) \tag{8}$$

其中,L_{doc} 表示一个 batch 语料中文档级别的交叉熵损失,$|K_d|$ 表示一个 batch 语料中文档嵌入的数目。T 表示情感极性,等于 1 时表示积极情感,等于-1 时表示消极情感。每个文档嵌入 h_d 的真实情感分布概率 $P(T \mid h_{di})$ 通过源语生成的词典计算得到,$P(T \mid e_{di})$ 表示文档向量通过模型预测得到的情感分布概率。

提取的情感特征需要同时考虑到输入语料的局部情感特征和整体情感特征,因此该文将单词级别和文档级别训练的情感分类模型加以融合,综合考虑两个分类子模型产生的交叉熵损失值,定义如公式 9 所示。

$$L = \alpha L_{\text{word}} + (1-\alpha) L_{\text{doc}} \tag{9}$$

其中,$\alpha \in [0,1]$ 表示两个子模型损失值的权重系数,α 越大表示越重视局部情感,α 越小则表示越重视整体情感。

Demszky D 等[17]认为传统情感分析的 6 种基本分类已经不能满足复杂情感的表达,因此提出包含 28 种不同情绪的细粒度情感数据集 GoEmotion。受其启发,该文没有直接使用融合后的模型进行情感分类,而是将模型最后一个 softmax 层的值取出,作为情感倾向的评分。在情感粒度为 n 时,将平行语料的每一条语句划分到不同的情绪下,其中包含 $(n-1)/2$ 个正面情绪和 $(n-1)/2$ 个负面情绪,以及 1 个中立情绪。

通过给不同的情绪打上特有的标记,将源语和目标语分别映射到不同的语料空间中,进而获得平行语料的情感语义表征,如公式 10 所示。

$$E^{\text{emo}} = W^{\text{emo}} \bullet x^{\text{emo}} \tag{10}$$

其中,x^{emo} 表示当前语料中情绪标记的 token 编码,W^{emo} 表示训练情感语义的线性变换参数。

2.3　获取情感语义增强的神经机器翻译过程

在 Transformer 的单个编码器中,包含一个自注意力层和一个前馈网络层。设输入的源语料为 $S = \{x_1, x_2, \cdots, x_n\}$,$n$ 表示语料的长度。通过源语的向量化表示构建自注意力层的原始输入,如公式 11 所示。

$$S_{\text{embed}} = \{E_1, E_2, \cdots, E_n\} \tag{11}$$

其中,E_n 表示输入语料第 n 个位置的词嵌入表示,语料词嵌入维度设为 len * dim,len 表示最大语料长度,dim 为词嵌入维度。

对于每一个词嵌入 E,将其与 3 个不同的权重矩阵 $\boldsymbol{W}^Q$、$\boldsymbol{W}^K$ 和 $\boldsymbol{W}^V$ 相乘,每个权重矩阵的维度为 dim * d。相乘的结果分别为 $\boldsymbol{Q}$(查询向量)、$\boldsymbol{K}$(键向量)和 $\boldsymbol{V}$(值向量),维度均为 len * dim。自注意力机制使用查询向量 Q 乘以键向量 K,获取当前值向量 V 的自注意力评分,再使用自注意力评分乘以值向量,作为当前编码器的输出,同时也是下一个编码器的输入,如公式 12 所示。

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d}}\right)\boldsymbol{V} \tag{12}$$

其中,$\sqrt{d}$ 是对权重矩阵 $\boldsymbol{W}^Q$、$\boldsymbol{W}^K$ 和 $\boldsymbol{W}^V$ 的列维度 d 求开方运算。

该文在 Transformer 的基础上融入了情感语义编码,将 2.2 中获取的情感语义表征与词嵌入表征按位相加。这样做的好处在于可以使用原本编码器的权重矩阵 $\boldsymbol{W}^Q$、$\boldsymbol{W}^K$ 和 $\boldsymbol{W}^V$ 进行自注意力的运算,没有引入额外的参数。使用词表征相加的方式与词表征拼接的方式,在深度学习训练过程中的作用一致,且减少了训练参数。设要训练的权重参数为 $\boldsymbol{W}$,推导过程如公式 13 所示。

$$\begin{aligned}\boldsymbol{W} \bullet x &= [W^I, W^{\text{emo}}] \bullet [[x^i]^{\text{T}}, [x^{\text{emo}}]^{\text{T}}]^{\text{T}} = \\ &W^I \bullet x^i + W^{\text{emo}} \bullet x^{\text{emo}} = \\ &E^I + E^{\text{emo}}\end{aligned} \tag{13}$$

其中,W^I 表示训练词嵌入的线性变换参数,W^{emo} 表示训练情感语义的线性变换参数,中括号表示拼接操作。可见对于原始表征的拼接操作,与词嵌入和情感语义编码的直接相加是等价的。由于该文使用原编码器的权重矩阵,因此 x^i 和 x^{emo} 相同,从而减少了训练参数。

对于第一个编码器的输入,使用词嵌入表征、情感语义编码和位置编码相加的方式,如公式 14 所示。

$$E = E^I + \beta \bullet E^{\text{emo}} + E^{\text{pos}} \tag{14}$$

其中,β 是情感强度因子,表示情感的强度。E^{pos} 表示位置编码,其在不同位置的取值如公式 15 所示。

$$\begin{cases} E^{(\text{pos},2i)} = \sin(\text{pos}/10\,000^{2i/\text{dim}}) \\ E^{(\text{pos},2i+1)} = \cos(\text{pos}/10\,000^{2i/\text{dim}}) \end{cases} \tag{15}$$

为了解决梯度消失问题,在每一个编码器的最后都包含了一个残差神经网络和层归一化结构,接受自注意力层的输出和原始词嵌入输入。

在解码器中,使用和编码器相同的词嵌入表征构建方式,但在层归一化和前馈网络层之间添加了交叉注意力机制,用于获取编码器的隐藏输出,且通过注意力分数的计算改变输出值。进入前馈网络层的输入分别是编码器的输出查询向量 $\boldsymbol{Q}_e$、键向量 $\boldsymbol{K}_e$ 和解码器的输出 $\boldsymbol{V}_d$,如公式 16 所示。

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}_e\boldsymbol{K}_e^{\text{T}}}{\sqrt{d}}\right)\boldsymbol{V}_d \tag{16}$$

在经过前馈网络层的线性变换和激活函数后,通过线性层和 softmax 层将每个单词的概率分布映射到目标语的词汇表空间中。

在生成翻译时,通过计算解码器当前时刻输出单

词的概率分布,获取条件概率最大的翻译生成序列,如公式 17 所示。

$$P(Y \mid C) = \begin{cases} P(y_i \mid C_i) & i = 0 \\ \prod_{i=1}^{N_y} P(y_i \mid C_i, y_0, y_1, \cdots, y_i) & i \geqslant 1 \end{cases} \tag{17}$$

其中,i 表示解码器输入语料所处的时刻,y_i 表示第 i 时刻解码器的输出,C_i 表示第 i 时刻编码器传递的隐藏状态,N_y 表示输入解码器的语料最大长度,Y 表示解码器找出的翻译序列。

3 实验与分析

3.1 实验数据集

为了验证情感语义对 NMT 模型产生的影响,该文使用 IMDB(Internet Movie Database,互联网电影资料库)数据集训练一个融合 GRU 和文档嵌入的多输入模型。IMDB 数据集包含 10 万条 IMDB 影评,其中 5 万条影评含有情感分类标签,评论被标记为二元情感。

将 80% 的语料用于训练,20% 的语料用于验证,具体分割情况见表 1。预训练词嵌入使用谷歌[18]发布的 GoogleNews-vectors-negative300. bin。

表 1 情感分类 IMDB 数据集参数

语料总数	训练集	验证集
50 000	40 000	10 000

NMT 模型的训练数据集包括以下平行语料:IWSLT14(International Conference on Spoken Language Translation,国际口语机器翻译大会)中的德-英和 IWSLT15 英-越数据集,以及 WMT14(Workshop on Statistical Machine Translation,统计机器翻译研讨会)中的英-法和 WMT17 英-德数据集。具体划分如表 2 所示。

表 2 神经机器翻译平行语料数据集规模统计

数据集	IWSLT14 德-英	IWSLT15 英-越	WMT14 英-法	WMT17 英-德
训练集	320 K	266 K	9 174 K	7 921 K
验证集	14 570	3 108	6 878	80 120
测试集	13 502	2 538	6 008	6 008

3.2 评价指标

采用 BLEU4[19]作为生成译文的评估指标。在统计语言模型中,N-gram 表示 n 个连续的单词序列,假设 N-gram 的 BLEU 评分为 P_n,其计算方法如公式 18 所示。

$$P_n = \frac{\sum_{c \in \text{candidates}} \sum_{n\text{-gram} \in c} \text{Count}_{\text{clip}}(n\text{-gram})}{\sum_{c' \in \text{candidates}} \sum_{n\text{-gram} \in c'} \text{Count}(n\text{-gram}')} \tag{18}$$

其中,C 表示 NMT 译文中出现的 N-gram,C' 表示在参考译文中出现的 N-gram,candidates 表示参考译文,$\text{Count}_{\text{clip}}(n\text{-gram})$ 表示提取 NMT 译文和参考译文中 N-gram 出现次数的最小值。

通常 N-gram 中的 N 取值为 1 到 4,该文使用综合评分 BLEU4 作为翻译效果的评价标准,其计算方法如公式 19 和公式 20 所示。

$$\text{BLEU4} = \text{BP} \cdot \exp\left(\sum_{n=1}^{4} w_n \log p_n\right) \tag{19}$$

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leqslant r \end{cases} \tag{20}$$

其中,BP 为长译文惩罚因子,c 表示 NMT 译文的词数,r 表示参考译文的词数,w_n 表示各阶 N-gram 拥有的权重。

3.3 实验参数设置

NMT 模型训练任务在 Facebook 发布的 Fairseq[20]上实现,使用斯坦福大学发布的 CoreNLP[21]自然语言处理工具包作为分词工具。选择 transformer 的轻量级模型 transformer_iwslt_de_en 作为基准算法,对比实验的主要参数跟随 Wu 等[22]的设置。

NMT 数据预处理阶段使用基于子词单元的 BPE[23](Byte Pair Encoding,字节对编码)技术,模型翻译质量评价使用 multi-bleu. Perl 开源脚本。实验中所有模型均在 RTX4070 Ti GPU 上进行训练。

3.4 对比实验设置

为了验证 ESEED 方法的有效性,分别使用以下几种模型进行对比:

(1)FConv[4]:完全基于卷积神经网络,并为每个解码器层配备了单独的注意力模块。

(2)Mixture-Models[24]:使用集成学习的方法,构造基于多个 Transformer 训练任务的混合模型。

(3)Transformer[6]:使用强自注意力机制和交叉自注意力机制进行模型训练的方法。

(4)LightConv[22]:使用基于当前时间步长预测的独立卷积核构建编解码器的方法。

(5)Linformer[25]:使用低秩矩阵来近似表示自注意力机制的方法。

(6)RoBERTa-Transformer:使用 RoBERTa 预训练模型[26]作为 NMT 模型训练的编码器。

(7)RoBERTa-ESEED:使用 RoBERTa 预训练模

型作为编码器,在编/解码器各层融合细粒度情感语义。

(8) Transformer-ESEED:使用 Transformer 作为 NMT 模型训练的编/解码器,在编/解码器的各层均融合细粒度情感语义。

3.5　实验结果分析

该文选择训练过程中 PPL(Perplexity,语言困惑度)最小的模型作为当前方法的最优模型,并使用该模型生成译文并计算 BLEU 评分。实验结果如表 3 所示。从结果可知,RoBERTa-ESEED 模型在 IWSLT14 德英翻译任务中取得了最优的效果,相较基线模型获得了 1.3 的 BLEU 值提升。而 Transformer-ESEED 模型在 IWSLT15 英越翻译任务和 WMT17 英德翻译任务中取得了最优的效果,相较基线模型分别获得了 1.62 和 1.59 的 BLEU 值提升。证明了基于情感语义增强编解码的神经机器翻译模型的有效性。

表 3　不同 NMT 模型在各数据集上的性能

模型	IWSLT14 德-英	IWSLT15 英-越	WMT14 英-法	WMT17 英-德
FConv	31.48	27.51	31.53	22.14
Mixture-Models	31.98	25.26	/	/
Transformer	33.26	26.65	32.03	23.59
LightConv	34.11	28.14	**34.35**	25.09
Linformer	29.50	23.84	/	/
RoBERTa-Transformer	32.81	26.56	32.17	23.36
RoBERTa-ESEED	**34.56**	27.08	34.31	24.89
Transformer-ESEED	34.50	**28.27**	34.22	**25.18**

另一方面,对 NMT 模型在 IWSLT14 德英数据集和 IWSLT15 英越数据集上的 Val_loss 值进行比较,结果如图 2 所示。可以发现,在 Transformer 中融合情感语义可以小幅提升模型训练的收敛速度。

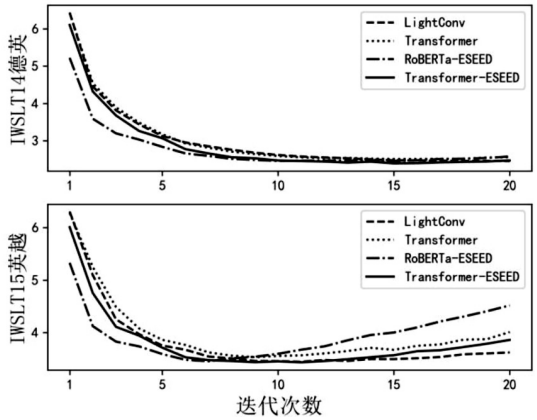

图 2　平行语料验证集损失值变化趋势

相对而言,RoBERTa-Transformer 模型在四个 NMT 任务上的性能表现不佳。这一现象可能源于 RoBERTa 模型的训练方式,即通过动态遮蔽进行语言建模,而未经过平行语料的相关性训练,因而缺乏挖掘源语言和目标语言之间关系的能力。然而,情感语义在各种语言中都普遍存在,源语言中的情感倾向往往也会在目标语言中得到表达。因此,将情感语义融入 RoBERTa 预训练模型,能够显著提升其在收敛速度和模型性能方面的表现。

3.6　不同情感粒度对 NMT 模型性能的影响

该文将训练的 GRU-Doc 分类模型的 softmax 层结果作为情感评分,并根据不同情感粒度对情感评分的范围进行划分。随后,使用 Transformer-ESEED 方法在不同情感粒度的数据集上进行 NMT 模型训练,并对性能进行了对比。具体的实验结果见表 4。

表 4　IWSLT14 德英任务不同情感粒度下的模型性能对比

情感粒度	Transformer-ESEED
2	34.19
5	34.33
7	34.38
19	**34.50**

该文采用不同的情感粒度(2,5,7 和 19)来研究情感评分的分布情况。当情感粒度为 2 时,将情感评分简单地划分为正向情感(大于 0.5)和负向情感(小于 0.5)。然而,在更细粒度的情感粒度下(5,7 和 19),引入了更复杂的划分标准:评分在 0.45 到 0.55 之间被视为中性情感,而其余评分则按照所述的等分操作进行划分。以情感粒度为 19 的情况为例,平行语料被分为了 19 个不同的情感类别,其中包括 9 种积极情绪、1 种中立情绪和 9 种消极情绪。

通过对比实验结果,观察到在细粒度情感下训练的 NMT 模型表现出更为优异的性能。这表明,细粒度情感下的 NMT 模型能够更有效地理解平行语料中复杂的情感信息。

3.7　不同情感强度对 NMT 模型性能的影响

通过在公式 14 中为融入不同编解码层的情感语义添加了情感增强因子 β,定量地分析不同情感强度

对 NMT 模型性能的影响,如表 5 所示。

表 5 IWSLT14 德英任务上不同情感强度的结果

β 值	ESEED
0.4	33.86
0.6	33.93
0.8	34.26
1.0	34.50
1.2	**34.65**
1.4	34.54
1.6	34.09

可以观察到,随着情感语义强度在编解码层的增加,模型性能呈现出逐渐提升,然后迅速下降的趋势,并在 β 约为 1.2 时达到最佳结果。因此,适度增强情感强度有助于提升 NMT 模型的性能,但过度增强情感语义可能会导致整体语义空间过度偏移,从而对模型产生不良影响。

3.8 实例分析

表 6 展示了融合情感语义的 Transformer 方法在 IWSLT14 德英翻译任务中的实例分析,其中情感粒度选择为包含 19 种情绪。通过与参考译文对比,可以发现 Transformer-ESEED 方法能够有效挖掘平行语料中的情感语义,并纠正译文中出现的情感语义偏差。在例句 1 中,平行语料中体现了较为强烈的不满意情感倾向“less satisfied”,然而 Transformer 方法未能发现这一隐藏的情感语义,使用了较为常见的表达“less happy”,而文中方法则准确地体现了原语料中的情感。同样,在例句 2 中,由于文中方法提前获取了语料的情感语义,因此有效地避免了使用“actually incredibly”这样夸张的情感表达方式,从而提升了语义表达的准确性。

表 6 融合情感语义的译文质量对比

示例	源语句	参考译文	Transformer 译文	ESEED 译文
例句 1	gut , wenn es sehr viele alternativen zu bedenken gibt , ist es einfach sich attraktive eigenschaften vorzustellen von alternativen , die sie ausschließen , das macht sie weniger zufrieden mit den alternativen , die sie gewählt haben.	well , when there are lots of alternatives to consider , it is easy to imagine the attractive features of alternatives that you reject , that make you less satisfied with the alternative that you ' ve chosen .	well , if there are many alternatives to think about it , it ' s easy to imagine attractive traits of alternatives that they choose , that makes them less happy with the alternatives that they ' ve chosen .	well , if there are lots of alternatives to consider , it ' s just a attractive traits of alternatives that you exclude , that makes you less satisfied with the alternatives you choose .
例句 2	es war wirklich ungeheuer tiefgreif .	it was really , really quite profound .	it was actually incredibly profound .	it was really , really profound .

4 结束语

为了提升低资源环境下神经机器翻译的性能,提出了一种基于情感语义增强编解码的神经机器翻译方法 Transformer-ESEED。该方法利用少量额外的有标签情感语料,指导平行语料的情感分类和情感语义获取。实验结果表明,该方法显著提升了机器翻译的质量。

然而,将情感语义融入自注意力表征增加了模型训练的计算开销和显存开销。因此,未来的研究会针对自注意力表征进行适当的切分,以缓解语义增强所带来的训练速度降低问题。

参考文献:

[1] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[J]. Advances in Neural Information Processing Systems,2014,27:3104-3112.

[2] ARTETXE M, LABAKA G, AGIRRE E. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings[J]. arXiv:1805.06297,2018.

[3] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[C]// Proceedings of the 6th international conference on learning representations. Vancouver:IEEE,2015:1-15.

[4] GEHRING J, AULI M, GRANGIER D, et al. Convolutional sequence to sequence learning[C]//Proceedings of the 2017 international conference on machine learning. New York: ACM,2017:1243-1252.

[5] 李梦洁,董 峦. 基于 PyTorch 的机器翻译算法的实现[J]. 计算机技术与发展,2018,28(10):160-163.

[6] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 2017 neural information processing systems. Cambridge:MIT Press,2017:5998-6008.

[7] 杨 丹,拥 措,仁青卓玛,等. 基于 mRASP 的藏汉双向神经机器翻译研究[J]. 计算机技术与发展,2023,33(12):200-206.

[8] 亢晓勉,宗成庆. 基于篇章结构多任务学习的神经机器翻译[J]. 软件学报,2022,33(10):3806-3818.

[9] 王振晗,何建雅琳,余正涛,等. 融合句法解析树的汉-越卷积神经机器翻译[J]. 软件学报,2020,31(12):3797-3807.
[10] 薛　媛. 基于 AMR 语义和图神经网络的汉蒙神经机器翻译的研究[D]. 呼和浩特:内蒙古工业大学,2021.
[11] 普浏清,余正涛,文永华,等. 基于依存图网络的汉越神经机器翻译方法[J]. 中文信息学报,2021,35(12):68-75.
[12] 黄　鑫,张家俊,宗成庆. 基于跨模态实体信息融合的神经机器翻译方法[J]. 自动化学报,2023,49(3):1-11.
[13] 朱星浩,胥　备. 基于 GRU 算法的音乐和词语的情感语义匹配算法[J]. 计算机技术与发展,2021,31(11):46-51.
[14] 徐月梅,施灵雨,蔡连侨. 一种基于情感特征表示的跨语言文本情感分析模型[J]. 中文信息学报,2022,36(2):129-141.
[15] CHUNG J,GULCEHRE C,CHO K H,et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[J]. arXiv:1412.3555,2014.
[16] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation,1997,9(8):1735-1780.
[17] DEMSZKY D,MOVSHOVITZ-ATTIAS D,KO J,et al. GoEmotions:a dataset of fine-grained emotions[J]. arXiv:2005.00547,2020.
[18] MIKOLOV T,CHEN K,CORRADO G,et al. Efficient estimation of word representations in vector space[J]. arXiv:1301.3781,2013.
[19] 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. Philadelphia:Association for Computational Linguistics,2002:311-318.
[20] OTT M,EDUNOV S,BAEVSKI A,et al. FAIRSEQ:a fast, extensible toolkit for sequence modeling[J]. arXiv:1904.01038,2019.
[21] MANNING C D, SURDEANU M, BAUER J, et al. The Stanford CoreNLP natural language processing toolkit[C]//Proceedings of 52nd annual meeting of the association for computational linguistics:system demonstrations. Baltimore:Association for Computational Linguistics,2014:55-60.
[22] WU F,FAN A,BAEVSKI A,et al. Pay less attention with lightweight and dynamic convolutions[J]. arXiv:1901.10430,2019.
[23] SENNRICH R, HADDOW B, BIRCH A. Neural machine translation of rare words with subword units[C]//54th annual meeting of the association for computational linguistics. Berlin:Association for Computational Linguistics,2016:1715-1725.
[24] SHEN T,OTT M,AULI M,et al. Mixture models for diverse machine translation:tricks of the trade[C]//International conference on machine learning. Long Beach:PMLR,2019:5719-5728.
[25] WANG S,LI B Z,KHABSA M,et al. Linformer:self-attention with linear complexity[J]. arXiv:2006.04768,2020.
[26] LIU Y,OTT M,GOYAL N,et al. Roberta:a robustly optimized bert pretraining approach[J]. arXiv:1907.11692,2019.