自动化学报
ACTA AUTOMATICA SINICA

基于跨模态实体信息融合的神经机器翻译方法

黄鑫 张家俊 宗成庆

Neural Machine Translation Method Based on Cross-modal Entity Information Fusion

HUANG Xin, ZHANG Jia-Jun, ZONG Cheng-Qing

在线阅读 View online: https://doi.org/10.16383/j.aas.c220230

基于跨模态实体信息融合的神经机器翻译方法

黄 鑫 [1, 2] 张家俊 [1, 2] 宗成庆 [1, 2]

摘 要 现有多模态机器翻译 (Multi-modal machine translation, MMT) 方法将图片与待翻译文本进行句子级别的语义融合. 这些方法存在视觉信息作用不明确和模型对视觉信息不敏感等问题, 并进一步造成了视觉信息与文本信息无法在翻译模型中充分融合语义的问题. 针对这些问题, 提出了一种跨模态实体重构 (Cross-modal entity reconstruction, CER) 方法. 区别于将完整的图片输入到翻译模型中, 该方法显式对齐文本与图像中的实体, 通过文本上下文与一种模态的实体的组合来重构另一种模态的实体, 最终达到实体级的跨模态语义融合的目的, 通过多任务学习方法将 CER 模型与翻译模型结合, 达到提升翻译质量的目的. 该方法在多模态翻译数据集的两个语言对上取得了最佳的翻译准确率. 进一步的分析实验表明, 该方法能够有效提升模型在翻译过程中对源端文本实体的忠实度.

关键词 实体重构, 跨模态学习, 多任务学习, 多模态机器翻译

引用格式 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. 自动化学报, 2023, **49**(6): 1170−1180

DOI 10.16383/j.aas.c220230

Neural Machine Translation Method Based on Cross-modal Entity Information Fusion

HUANG Xin[1, 2] ZHANG Jia-Jun[1, 2] ZONG Cheng-Qing[1, 2]

Abstract Existing multi-modal machine translation (MMT) methods perform the sentence-level semantic fusion of images and text to be translated. These methods have problems such as the unclear role of visual information played in the translation procedure and the insensitivity of the model to visual information, and further cause the problem that visual information and text information cannot be fully semantically integrated into the translation models. To solve these problems, a cross-modal entity reconstruction (CER) method has been proposed. Different from incorporating the complete image into the translation model, this method explicitly aligns the entities in the text and the image, reconstructs the entity of one modality through the combination of the text context and the entity of the other modality, and finally achieves the purpose of entity-level cross-modal semantic fusion. Through the multi-task learning method, the CER model is combined with the translation model to improve the translation quality. The method achieves the best translation accuracy on the two language pairs of the multi-modal translation dataset. Further analysis experiments show that this method can effectively improve the fidelity to the source-end textual entities in the translation procedure.

Key words Entity reconstruction, cross-modal learning, multi-task learning, multi-modal machine translation (MMT)

Citation Huang Xin, Zhang Jia-Jun, Zong Cheng-Qing. Neural machine translation method based on cross-modal entity information fusion. *Acta Automatica Sinica*, 2023, **49**(6): 1170−1180

神经网络方法在计算机视觉和自然语言处理等领域均取得了很好的效果. 因此, 通过图像和文本跨模态信息融合的方式来提高机器翻译的质量也成为了可能. 多模态机器翻译 (Multi-modal machine translation, MMT) 就是一种通过在文本中融入视觉信息来提升翻译质量的机器翻译方法[1−2]. 目前的相关研究主要针对图像描述的翻译. 相比于描述图像的文本, 图像自身包含了更完整的信息. 因此, 将视觉信息作为文本以外的补充知识提供给翻译模型能够获得更加准确的译文, 是多模态机器翻译的一个基本假设[3].

然而, 如何有效地将视觉信息融入到翻译中成为了研究者们所面临的挑战. 早期的相关研究[4−8] 尝试在神经机器翻译 (Neural machine translation, NMT) 模型中输入经过卷积神经网络 (Convolutional neural network, CNN) 提取得到的图像全局特征来达到跨模态信息融合的目的. 另有研究工作尝试利用注意力机制使翻译模型能够动态地关注图片内部的局部信息[9−12]. 为了输入图像内部与文本

收稿日期 2022-03-27 录用日期 2022-07-21
Manuscript received March 27, 2022; accepted July 21, 2022
国家自然科学基金 (U1836221) 资助
Supported by National Natural Science Foundation of China (U1836221)
本文责任编委 张民
Recommended by Associate Editor ZHANG Min
1. 中国科学院自动化研究所模式识别国家重点实验室 北京 100190 2. 中国科学院大学人工智能学院 北京 100049
1. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190 2. School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049

内容相关的信息, 文献 [13] 尝试在 NMT 模型中输入经过提取的图片内部视觉目标. 然而以上方法在设计上并没有考虑视觉信息如何明确地作用到翻译任务中, 使得视觉信息在翻译过程中的具体贡献不明确. 文献 [14−16] 为了探究视觉信息是否在翻译的过程中有所帮助, 将 MMT 模型的输入图片替换为与文本内容不相关的图片, 并观察到模型的翻译性能没有显著下降. 该实验结果表明翻译模型对视觉信息不敏感. 这是因为 MMT 模型在生成译文时可以很容易地从与参考译文有良好对齐关系的原文中寻找有用信息, 并且原文本已经包含了大部分翻译所需信息, 所以多数情况下模型没有必要从图片中寻找补充信息.

针对以上问题, 本文提出了跨模态实体重构 (Cross-modal entity reconstruction, CER) 方法, 用于帮助 NMT 模型提升译文质量. 区别于其他方法将图片中的视觉信息与整个句子进行语义融合, CER 以更加明确的方式针对实体进行跨模态信息的融合. 这样做能够保证具有相同语义的文本实体和视觉实体产生直接的相互作用. CER 模型主要负责文本实体和视觉实体的重构. 文本实体的重构主要依赖于文本上下文和视觉实体所提供的信息. 文献 [17] 的研究表明, 文本中实体缺失时 MMT 模型开始关注图片中的信息. 为了确保 CER 模型在重构文本实体时主要从视觉实体中获得信息, 本文对输入的文本上下文采用退化操作, 即删除文本中的文本实体. 文献 [18] 中采用了类似的方法来生成整个源端或目标端文本, 证明了退化文本与视觉实体的结合方式的有效性. 本文首次采用了视觉实体重构方法. 视觉实体重构主要依赖于文本实体与文本上下文所提供的信息. 文献 [19] 曾尝试利用文本生成整张图片的方法, 但这种方法更难对齐两个模态的语义信息. 本文所提视觉实体重构则以更细粒度且更精确的方式进行文本到图像的生成任务. 本文还在 CER 模型训练的过程中加入了少量的非实体文本的重构, 以保证实体与非实体之间语义关系的建立. CER 的混合重构方法能够保证实体级的视觉信息与文本信息的充分融合. 最后, 通过多任务学习方法将 CER 模型与 NMT 模型进行部分参数的共享, 达到提升翻译性能的目的. 实验结果表明, CER 能够很好地帮助 NMT 模型提升译文质量. 在进一步的分析实验中发现, CER 能够帮助 NMT 模型提升在翻译过程中对源端文本实体的忠实度, 使 NMT 模型更准确地从源端实体词中获得信息.

本文主要贡献如下: 1) 提出了跨模态实体重构方法, 在 MMT 常用数据集 Multi30K 上验证了 CER 能够帮助翻译获得很好的性能提升; 2) 验证了采用明确的方式融合跨模态信息的方法的有效性, 从方法上规避了 MMT 模型在训练阶段对视觉信息不敏感的问题; 3) 实现了双向的实体级跨模态信息的融合, 使视觉信息更充分地融入到文本中; 4) 验证了实体级信息融合具有更强的可解释性, 跨模态的实体信息融合可以帮助 NMT 模型在生成目标端实体时更忠实于源端实体所提供的信息.

1 相关工作

近年来有很多在自然语言处理任务中融入多模态信息的工作. 例如结合视觉信息以及语音信息的词表示[20−21], 结合图片与问题产生相关答案的视觉问答任务[22], 以及通过融入主题相关的图片来辅助生成文本摘要的多模态摘要任务[23]. 这些融入视觉信息的多模态任务意图利用视觉信息来纠正文本中存在的歧义, 或是直接指导目标文本的生成. 机器翻译任务同样适用于这些应用场景. 目前主流的多模态机器翻译相关研究集中于将视觉信息与句子级别的语义相融合, 以此来达到提高翻译准确率的目的. 本节将介绍多模态机器翻译方法的相关工作, 以及作为本文所提方法的基础模型 Transformer.

1.1 多模态机器翻译

MMT 方法的相关研究一直紧随着 NMT 方法的发展而更新. 从最开始基于循环神经网络编解码结构的神经机器翻译模型, 到后来在模型的编码器和解码器之间增加注意力机制, 多模态机器翻译也一直在尝试将视觉特征融入到这些翻译模型中. 例如, 文献 [4] 中将图片经过 CNN 提取得到的全局特征作为完整的视觉语义单元拼接到输入文本序列中, 或是用全局特征初始化编码器与解码器. 文献 [7−8] 则通过在翻译模型中加入变分编码器的方式来融合视觉信息. 有方法尝试将包含图片中空间信息的未经全局池化 (Global pooling) 的局部特征输入到翻译模型中. 例如文献 [9−10] 分别提出了利用注意力机制同时关注文本序列和图像局部特征的方法. 基于多头自注意力机制的 Transformer 能够更高效地对信息进行编码[24]. 因此也有工作尝试将在基于循环神经网络 (Recurrent neural network, RNN) 上的 MMT 方法移植到基于 Transformer 的模型上[11]. 这些方法普遍采用了拼接、连接以及求和等方式融合文本和视觉两种模态的向量表示, 并认为神经网络方法能够通过学习利用到跨模态信息. 因难以明确视觉信息在句子的哪些细粒度的语义单元上起作用, 本文称此类方法为句子级融合方法.

句子级融合方法饱受对视觉信息不敏感的诟病. 有研究认为这类方法中的图片在输入到模型后主要起到正则作用, 从而得到提升了翻译性能的结果[15]. 融入视觉实体的方法则尝试为模型输入更细粒度的视觉语义信息. Huang 等[13] 首先尝试了将图片的视觉目标提取出来形成一个序列再输入到翻译模型中. 该工作尝试了多种将视觉目标序列与文本序列在编码端进行先拼接再编码的方案. 另有 Wang 等[25] 通过设计损失函数的方式使模型关注与文本内容相关的视觉目标. Yin 等[26] 则对输入序列进行了更细致的建模. 该工作视源语言句子与视觉目标的关系为一个多模态图结构. 其中词和视觉目标作为节点, 词与词的关系以及词与视觉目标的关系作为边. 最后设计了一个多模态图结构编码器对图进行编码. 以上方法均需要在训练和测试时输入图片到模型中. 另有一些方法只在训练时加入图片, 得到的最终模型是利用视觉信息增强后的 NMT 模型. 其中最具有代表性的是文献 [5], 该方法尝试利用源语言文本的编码结果“想象”到图片的向量表示. 文献 [18] 则为了让模型对视觉信息敏感采用了退化文本与视觉目标组合的方式生成源端或目标端的句子. 而文献 [19] 则尝试利用生成对抗网络生成完整的图片.

本文所提 CER 方法属于将细粒度的视觉语义信息融入到细粒度的文本中, 最终用于增强 NMT 模型的翻译性能. 相比于其他方法, CER 的作用目标更明确, 并以双向跨模态信息融合的方式吸取了已有方法的优点.

1.2 Transformer

NMT 模型一般采用端到端序列生成框架, 包含源语言编码器和目标语言的解码器两部分. 其中编码器负责将源语言文本序列编码为语义表征向量, 解码器则根据编码端生成的表征向量和已经生成的目标端历史序列预测生成下一个目标端词. Vaswani 等[24] 提出的 Transformer 是目前 NMT 模型中最广泛使用的基础框架. 本文提出的 CER-NMT 就是基于 Transformer 结构实现的.

Transformer 通过多头自注意力 (Multi-head self-attention) 机制可以建立输入序列中任意两个单词之间的语义关系. 这使得 Transformer 具有强大的编码序列信息的能力. 具体地, 模型的输入为 $\boldsymbol{X} = \{x_1, x_2, \cdots, x_N\}$, 经过位置编码模块得到词向量 (Embedding) 序列

$$\boldsymbol{X}_{\rm PE} = PosEmb(\boldsymbol{X}) \tag{1}$$

将 $\boldsymbol{X}_{\rm PE}$ 经过 3 个不同的线性变换可得到多头自注意力模块的 3 个输入 $\boldsymbol{Q}$, $\boldsymbol{K}$ 和 $\boldsymbol{V}$. 自注意力机制的计算方式为

$$Att(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = softmax\left(\frac{\boldsymbol{Q}K^{\rm T}}{\sqrt{d_k}}\boldsymbol{V}\right) \tag{2}$$

其中, d_k 表示 K 的维度. $\boldsymbol{Q}$ 与 $\boldsymbol{K}$ 的点积操作可以建立输入序列中任意两个位置的关联. 若 $\boldsymbol{Q}$ 来自目标端, 则点积的结果代表目标端序列与源端序列之间的关联, 该关联值可以代表目标端单词在解码时受源端单词的影响程度. 本文将在后面的分析中利用该关联值测量模型对源端文本实体的忠实度. 在经过残差连接 (Residual connection)、层归一化 (Layer normalization)、前馈网络 (Feed forward network) 以及 L 层以上编码操作后, 输入序列就完成了编码过程, 并得到输入序列的隐层表示 (Hidden states) $\boldsymbol{H}_X^L$, 该过程简化为

$$\boldsymbol{H}_X^L = MHSAtt(\boldsymbol{X}, \boldsymbol{X}, \boldsymbol{X}) \tag{3}$$

其中, $\boldsymbol{H}_X^L = \{\boldsymbol{h}_1^l, \boldsymbol{h}_2^l, \cdots, \boldsymbol{h}_N^l\}$. 从以上描述可知, $\boldsymbol{h}_i^l$ 不仅代表 x_i 独有的隐层表示, 还结合了整个输入序列的信息. 这使得 Transformer 具有为序列中每个位置融合上下文信息的能力. 本文将借助 Transformer 这一优点, 为每个实体生成融合了上下文信息的向量表示, 从而重构出跨模态的实体.

2 方法描述

本节将首先描述文本实体和视觉实体的显式对齐方法, 然后详细介绍 CER 模型如何为实体和上下文信息编码, 以及如何进行显式地跨模态实体重构, 最后阐述如何将 CER 模型与 NMT 模型相融合.

2.1 显式实体对齐

MMT 方法主要针对的是静态图片文本描述的翻译. 图片在输入到翻译模型之前, 一般会利用计算机视觉领域已有的经过深度验证的方法进行预处理, 例如, 使用经过图像分类任务进行预训练的 Resnet-50[27] 模型提取图像特征. 本文所提方法主要针对文本实体 $\boldsymbol{T} = \{t_a, t_b, \cdots\}$ 与视觉实体 $\boldsymbol{E} = \{e_a, e_b, \cdots\}$ 的信息融合, 因此需要预先提取出文本 $\boldsymbol{X} = \{x_1, x_2, \cdots, x_N\}$ 与对应图片中的实体, 其中 $\boldsymbol{T} \subseteq \boldsymbol{X}$. 图 1 为提取两个模态实体的方法示例.

在本文中, 文本实体就是在所描述图片中有对应视觉目标的名词, 例如图 1 中的“man”对应了图片中的 e_1. 如图 1(a) 所示, 提取文本实体时, 首先需要利用文本分析工具提取出文本中的名词短语“A man”和“a hat”, 然后保留其中的名词作为预选文本实体 t_1 = “man” 和 t_5 = “hat”. 为了提取视觉

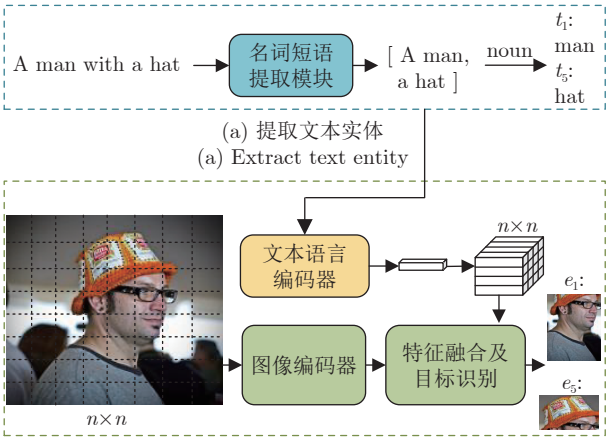

图 1　显式实体对齐示例

Fig. 1　An example of the explicit way to align cross-modal entities

实体, 本文采用了 Yang 等[28] 所提供的文本−视觉定位 (Visual grounding) 方法, 图 1(b) 为简化示意图. 该方法首先将图片输入视觉编码器 (Image encoder) 得到一个 $n \times n$ 的特征表示, 每个位置的视觉特征对应图片中的一个矩形子区域. 然后将短语的向量表示复制为 $n \times n$ 份, 每份对应了图片中的各子区域, 如图 1(b) 中白色立方体所示. 最后将两个模态的向量表示融合后就可以计算出输入图片中与输入短语相关的子区域, 从而合并得到相应的视觉目标 e_1 和 e_5. 将以上可以定位到视觉目标的预选文本实体选定为文本实体 t_1 和 t_5. 通过以上操作即可完成显式对齐文本实体与视觉实体. 提取出的 $\{e_1, e_5\}$ 两个视觉目标再经过 CNN 模型提取出视觉特征 (Visual feature) 向量, 然后利用前馈神经网络 (Feed forward network, FFN) 映射到所需的维度, 即可得到视觉实体的向量表示 $\{\boldsymbol{e}_1, \boldsymbol{e}_5\}$ 用于后续任务的使用.

2.2　跨模态编码

图 2 展示了 CER 模型与 NMT 模型相融合的简化过程, 其中包含了跨模态编码、跨模态实体重构以及 NMT 等 3 个主要任务. 跨模态编码负责在进行实体重构之前将视觉信息与文本信息编码融合. 所编码的信息包含实体重构主要依赖的 3 个部分:

1) 跨模态实体. 其中实体重构主要依赖于跨模态实体所提供的具体信息. 如图 2 所示, 文本实体 t_1 的重构主要依赖于视觉实体 e_1, 视觉实体 e_1 的重构也主要依赖于文本实体 t_1.

2) 文本上下文. 本文所用文本上下文为经过退化的文本 $\tilde{\boldsymbol{X}}$. 例如图 2 中进行实体重构时, 文本上下文为“A ⟨M⟩ with a ⟨M⟩”, 其中, ⟨M⟩代表将实体词所对应的位置删除并保留该位置空缺. 文本上下文与跨模态实体共同组成文本的完整语义. 在进行视觉实体重构时, 模型的输入由文本上下文和文本实体共同组成, 即原文本 $\boldsymbol{X}$. 在进行文本实体

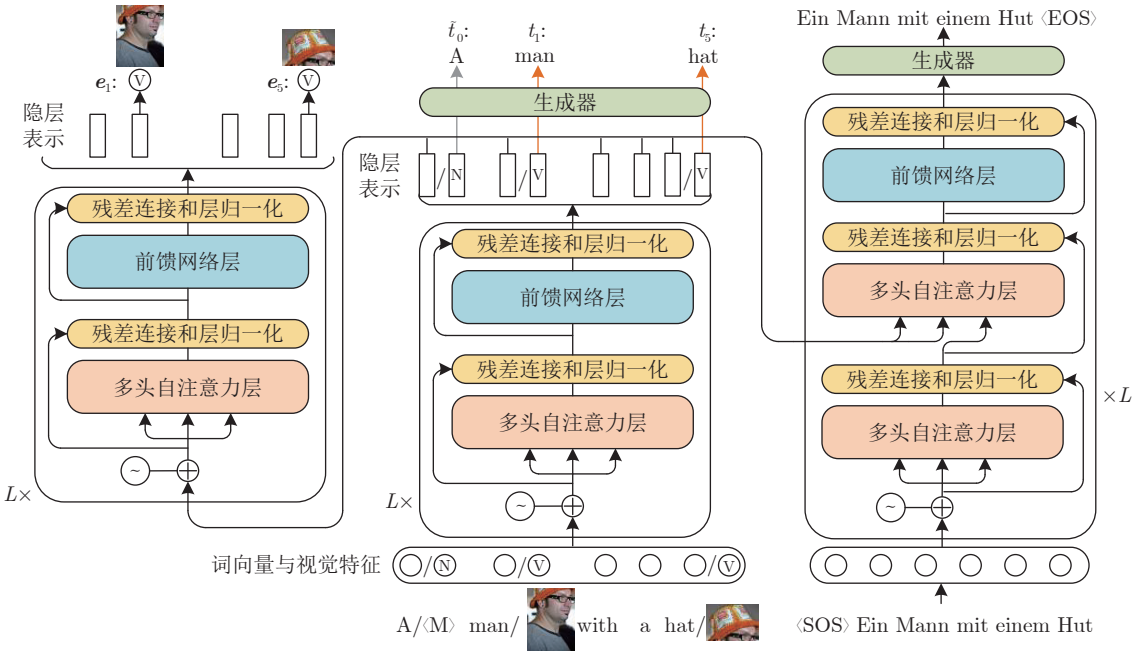

图 2　结合跨模态实体重构方法的神经机器翻译模型图

Fig. 2　NMT model framework combined with CER

重构时, 模型的输入是文本上下文和视觉实体所组成的多模态混合序列 $\boldsymbol{Z}$. 在进行文本非实体的重构时, 模型的输入同样是多模态混合序列, 但是需要将待重构的非实体词利用掩码词"⟨M⟩"替换掉, 形成 $\tilde{\boldsymbol{Z}}$. 跨模态实体与文本上下文组合重构实体的方式, 既能保证重构实体时具备充足的语义信息, 又能充分利用到自注意力机制的特性使实体与文本上下文建立良好的语义关系.

3) 实体对齐关系. 为了避免让模型完成较为困难的跨模态实体关系对齐, 本文采用直接替换对应位置向量表示的方式. 例如图 2 中重构文本实体 t_1 时, 将输入序列中的对应位置直接替换为视觉实体的向量表示 $\boldsymbol{e}_1$.

2.3 跨模态实体重构

2.3.1 视觉实体重构

视觉实体重构 (Visual entity reconstruction, VER) 依赖于原文本 $\boldsymbol{X}$. 例如图 2 中, "A man with a hat"为重构视觉实体时模型的输入序列, 此时图中输入到模型中的是 5 个圆形空心词向量. 该句子经过 L 层的编码器进行文本编码, 再利用 L 层视觉实体解码器解码出与文本实体 $\{t_1, t_5\}$ 相对应的视觉实体 $\{\boldsymbol{e}_1, \boldsymbol{e}_5\}$. 由于文本无法完整地描述图片, 只是针对图片中的个别属性的描述, 由文本直接还原像素级的图片或图片中的视觉实体几乎是不可能的. 因此本文仅尝试生成与视觉实体最相近的向量表示 $\{\boldsymbol{e}_1, \boldsymbol{e}_5\}$. 其中, $\boldsymbol{X}$ 经过如下编码过程

$$\boldsymbol{H}_X^L = MHSAtt_{\text{enc}}(\boldsymbol{X}, \boldsymbol{X}, \boldsymbol{X}) \tag{4}$$

得到编码后的隐层表示 $\boldsymbol{H}_X^L = \{\boldsymbol{h}_1^l, \boldsymbol{h}_2^l, \cdots, \boldsymbol{h}_N^l\}$, 其中 L 表示编码器层数, $MHSAtt_{\text{enc}}(\cdot)$ 为 L 层多头自注意力编码器. 然后经过如下解码过程

$$\boldsymbol{H}_X^{2L} = MHSAtt_{\text{dec}}(\boldsymbol{H}_X^L, \boldsymbol{H}_X^L, \boldsymbol{H}_X^L) \tag{5}$$

得到 $\boldsymbol{H}_X^{2L} = \{\boldsymbol{h}_1^{2l}, \boldsymbol{h}_2^{2l}, \cdots, \boldsymbol{h}_N^{2l}\}$, $MHSAtt_{\text{dec}}(\cdot)$ 为与 $MHSAtt_{\text{enc}}(\cdot)$ 具有相同结构的视觉实体解码器. 然后利用 $\boldsymbol{H}_X^{2L}$ 生成与 $\{\boldsymbol{e}_a, \boldsymbol{e}_b, \cdots\}$ 接近的向量表示, 文本采用与文献 [5] 相近的方式实现该过程

$$\mathcal{L}_{\text{VER}}(\theta, \varphi) = \frac{1}{|\boldsymbol{E}|}\sum_{j\neq i}\max\Big\{0, \varepsilon - \cos\left(\boldsymbol{h}_i^{2l}, \boldsymbol{e}_j\right) + \cos\left(\boldsymbol{h}_i^{2l}, \boldsymbol{e}_i\right)\Big\} \tag{6}$$

其中, $i, j \in \{a, b, \cdots\}$, θ 为文本编码器参数, φ 为视觉实体解码器的参数, $\cos(\cdot, \cdot)$ 计算向量间的余弦相似度, ε 为最小边缘常量. 显然, 优化 $\mathcal{L}_{\text{VER}}(\theta, \varphi)$ 能够缩减 $\boldsymbol{h}_i^{2l}$ 与正样本 $\boldsymbol{e}_i$ 之间的余弦距离, 增加与负样本 $\boldsymbol{e}_j$ 之间的余弦距离.

2.3.2 文本实体重构

文本实体重构 (Textual entity reconstruction, TER) 依赖于由文本上下文 $\tilde{\boldsymbol{X}}$ 和视觉实体 $\boldsymbol{E}$ 组合成的多模态混合序列

$$\boldsymbol{Z} = \text{Combine}\left(\tilde{\boldsymbol{X}}, \boldsymbol{E}\right) \tag{7}$$

如图 2 所示, 输入的文本上下文"A ⟨M⟩ with a ⟨M⟩"与视觉实体 $\{\boldsymbol{e}_1, \boldsymbol{e}_5\}$ 组合成多模态混合序列 $\boldsymbol{Z}$ = "A $\boldsymbol{e}_1$ with a $\boldsymbol{e}_5$", 此时输入到模型中的是三个圆形空心词向量和两个标记"V"的圆形视觉特征. 多模态混合序列经过编码器进行跨模态编码后, 得到序列的跨模态表示

$$\boldsymbol{H}_Z^L = MHSAtt_{\text{enc}}(\boldsymbol{Z}, \boldsymbol{Z}, \boldsymbol{Z}) \tag{8}$$

其中, $\boldsymbol{H}_Z^L = \{\boldsymbol{h}_1^l, \boldsymbol{h}_2^l, \cdots, \boldsymbol{h}_N^l\}$. 然后依据对应位置的隐层表示生成文本实体, 在图 2 中为两个标记"V"矩形隐层表示 $\boldsymbol{h}_2^l$ 和 $\boldsymbol{h}_5^l$, 该过程需要优化以下损失函数

$$\mathcal{L}_{\text{TER}}(\theta, \vartheta) = \frac{1}{|\boldsymbol{T}|}\sum_{x_i\in\boldsymbol{T}} OH\left(x_i\right)\ln p\left(\boldsymbol{Z}|\theta, \vartheta\right) \tag{9}$$

$$\ln p\left(\boldsymbol{Z}|\theta, \vartheta\right) = g\left(\boldsymbol{h}_i^l|\vartheta\right) \tag{10}$$

其中, $OH\left(x_i\right)$ 代表 x_i 的独热编码 (One-hot, OH) 表示, ϑ 为生成器 (Generator) 参数, 生成器 $g\left(\cdot\right)$ 将位置 i 处的隐层向量 $\boldsymbol{h}_i^l$ 映射并得到词表中每个词的预测概率 p.

2.3.3 文本非实体重构

文本非实体重构 (Textual none-entity reconstruction, TNER) 方法与文本实体重构方法相似, 区别在于非实体重构的目的是使视觉实体与文本上下文进行充分的句子级语义融合, 而实体的重构方法更侧重实体在两个模态之间的实体级语义融合. 文本非实体的输入同样是文本上下文 $\tilde{\boldsymbol{X}}$ 和视觉实体 $\boldsymbol{E}$ 组合成的多模态混合序列, 但是在 $\tilde{\boldsymbol{X}}$ 中除对实体的位置进行退化还要将待重构的非实体词退化. 例如图 2 中, 当选中对非实体词"A" 进行重构时, 模型所输入的多模态混合序列为 $\boldsymbol{Z}$ = "⟨M⟩ $\boldsymbol{e}_1$ with a $\boldsymbol{e}_5$", 此时输入到模型中的是一个标记"N"的圆形词向量代表"⟨M⟩"、两个圆形空心词向量和两个标记"V" 的圆形视觉特征. 损失函数为

$$\mathcal{L}_{\text{TNER}}(\theta, \vartheta) = \frac{1}{|\tilde{\boldsymbol{T}}|}\sum_{x_i\in\tilde{\boldsymbol{T}}} OH\left(x_i\right)\ln p\left(\tilde{\boldsymbol{Z}}|\theta, \vartheta\right) \tag{11}$$

其中, $\tilde{\boldsymbol{T}}$ 代表待重构的非实体词集, 并有 $\tilde{\boldsymbol{T}} \subseteq \boldsymbol{X} - \boldsymbol{T}$. 在图 2 中用于预测非实体词"A" 的隐层为一个标记"N" 的矩形隐层表示 $\boldsymbol{h}_1^l$.

本文采用随机选取的方式选择待重构的非实体词. 在执行 TNER 任务时, 每个文本中的非实体词有 30% 的概率会被选择到. 该概率与所采用数据集中实体词占 32.6% 的比例相近.

2.3.4　CER 联合训练

联合以上所有重构方法的目标函数, 可得到本文所提 CER 方法的联合损失函数为

$$\mathcal{L}_{\mathrm{CER}}(\theta, \varphi, \vartheta) = \alpha\mathcal{L}_{\mathrm{VER}}(\theta, \varphi) + \beta\mathcal{L}_{\mathrm{TER}}(\theta, \vartheta) + \gamma\mathcal{L}_{\mathrm{TNER}}(\theta, \vartheta) \tag{12}$$

其中, α, β, γ 为控制 3 种重构方法训练比例的超参数. 值得注意的是, 本文没有设置利用纯文本上下文 $\tilde{\boldsymbol{X}}$ 重构文本词的方案. 这是因为该方案与一般的预训练方法相似, 用于建立纯文本的语言模型. 而机器翻译模型的编码端本身就是一个很好的纯文本语言模型, 因此不必再设置纯文本的重构方案.

2.4　与机器翻译的融合

构建 CER 模型的目的是辅助翻译模型提升翻译质量. 本文采用多任务学习方法将 CER 模型中的跨模态编码器与 NMT 模型的文本编码器进行参数共享, 从而实现将跨模态信息融合到 NMT 模型的目的. NMT 任务所要优化的目标函数为

$$\mathcal{L}_{\mathrm{NMT}}(\theta, \psi, \vartheta) = -\sum_{j}^{M} \ln P(y_j | y_{<j}, \boldsymbol{X}|\theta, \psi, \vartheta) \tag{13}$$

其中, $\boldsymbol{Y} = \{y_1, y_2, \cdots, y_M\}$ 为目标语言的文本序列. 将式 (13) 与 CER 模型的损失函数式 (12) 相结合, 得到联合目标函数为

$$\mathcal{L}(\theta, \psi, \varphi, \vartheta) = \omega\mathcal{L}_{\mathrm{NMT}}(\theta, \psi, \vartheta) + (1-\omega)\mathcal{L}_{\mathrm{CER}}(\theta, \varphi, \vartheta) \tag{14}$$

其中, 用超参数 ω 调节 NMT 模型与 CER 模型的训练比例.

3　实验设置

3.1　数据

本文使用多模态机器翻译任务的常用数据集 Multi30K[3] 来测试本文所提方法. 该数据集每幅图片配有一个英文句子和其对应的德语和法语的译文. 该数据集的训练集包含 29 000 个平行句对. 其验证集和测试集 Test2016 的大小分别为 1 014 和 1 000. 本文还在更新的 Multi30K Test2017 以及 Ambiguous MSCOCO 上进行了测试, 其大小分别为 1 000 和 461.

本文采用 Moses SMT[29] 工具包对文本数据进行分词 (Tokenization) 和归一化 (Normalization) 处理. 为了防止对视觉实体与文本实体对应关系的破坏, 本文并没有采用双字节编码 (Byte pair encoding, BPE) 进行分词处理[30]. 在数据的预处理阶段, 如图 1 所示, 本文采用 spaCy[1] 为提取名词短语的工具. 视觉目标检测工具[28] 与文献 [26] 采用了相同的单步文本−视觉定位方案. 最后将视觉实体输入 Resnet-50[27] 提取出 2 048 维的全局图像特征, 该特征向量在输入到 CER 模型的时候会通过一个全连接层映射到模型所需的 128 维.

3.2　模型参数

本文在基于 Transformer[24] 结构的模型上进行实验. 由于数据集 Multi30K 的规模相对较小, 很容易在 Transformer-big 或 Transformer-base 上过拟合, 因此本文采用了更小规模的参数设置. 本文所设置的参数与文献 [26] 基本保持一致, 其词嵌入层设置为 128 维, 前馈内层为 256 维. 编码器、视觉实体解码器和文本解码器各 4 层 ($L = 4$), 多头自注意力的头数为 4. 词表采用源语言与目标语言共享的方式, 合并后英译德的共享词表大小为 27 226, 英译法为 19 393. 利用 Adam[31] 优化器优化整个模型参数, 并在优化 80 000 步后停止训练得到最终模型参数. 训练时, 批数据的大小设置为 2 000 个词. 测试时, 采用了搜索空间 $b = 4$ 的柱搜索算法.

多任务学习方法的训练方式为随机选择一个任务在当前批数据下对模型进行优化. 本文所提 CER 方法一共包含 3 个子任务: VER、TER 和 TNER. 这 3 个任务根据超参数 α, β, γ 控制训练的比例. 本文将 3 个子任务的训练比例设置为 $\alpha = 40\%$, $\beta = 40\%$, $\gamma = 20\%$. 这样保证了 CER 在训练中以 80% 的比例用于跨模态实体重构, 20% 的比例用于视觉实体与非实体上下文的语义融合. 本文将 NMT 设置为主任务, 并设置超参数 ω 控制 NMT 与 CER 之间的训练比重. 例如当 $\omega = 0.5$ 时, NMT 与 CER 的训练比例分别为 50%, 其中 VER、TER 和 TNER 3 个子任务的训练比重依据上文调整为 $\omega \times \alpha = 20\%$, $\omega \times \beta = 20\%$, $\omega \times \gamma = 10\%$.

本文所有的模型结果都是在 3 次随机初始化的条件下训练得到的平均值. 最后, 通过 Bleu4[32] 和 Meteor[33] 来测试翻译质量, 在后面的实验中分别简记为“B”和“M”.

3.3　基准模型

本文将 MMT 方法分为句子级融合、视觉实体

[1] https://spacy.io

融合以及增强 NMT 三大类. 句子级融合方法主要尝试将视觉信息与整个文本句子进行语义融合. 视觉实体融合方法则预先提取出图片中的视觉目标后再输入到所设计的模型中. 增强 NMT 方法一般在训练中融合视觉信息, 而在测试时不需要输入图片. 后两类方法中的多数模型在本质上也属于句子级融合方法. 各模型简介如下:

1) Base 是标准句子级 Transformer 翻译模型[24]. 具体参数见第 3.2 节.

2) IMGD (Image for decoder)[4] 是一种基于 GRU (Gate recurrent unit)[34] 的翻译模型, 利用图像全局特征初始化解码器.

3) VMMT_{C} (Variational MMT with conditional Gaussian latent model)[8] 是一种基于 LSTM (Long short-term memory)[35] 的翻译模型, 在翻译过程中利用视觉信息与文本利用变分编码器融合跨模态语义. VMMT_{F} (Variational MMT with fixed Gaussian prior) 在 VMMT_{C} 的基础上没有使用图像.

4) SerAttTrans (Serial attentions for Transformer)[11] 在解码端串行使用两个交叉注意力模块分别用于关注源端文本信息和图像局部特征.

5) GumAttTrans (Gumbel-Sigmoid Attention for Transformer)[36] 使用 Gumbel-Sigmoid 改造注意力机制, 帮助翻译模型关注到图片中与文本内容更相关的区域.

6) Parallel RCNNs (Region CNNs)[13] 将每个提取出的视觉实体分别与源语言文本进行先拼接再编码的过程, 得到的并行的源端编码序列将用于目标端的解码.

7) GMMT (Graph-based MMT)[26] 模型视源语言句子与对应图片中的视觉目标的关系为一个多模态图结构, 然后利用专门的基于图的跨模态编码器进行编码, 最终解码出目标端句子[26].

8) DelMMT (Deliberation MMT)[37] 模型提出使用推敲网络进行多模态二次解码. 在第 2 次解码中融合源语言、目标语言以及视觉信息.

9) Imagination[5] 是一种基于 GRU 的翻译模型, 并采用多任务的方式在附属任务中利用源语言文本的编码结果生成完整图片的全局特征. 同本文所提方法相似, 其在测试过程中不需要输入图片.

10) EMMT (Entity-level MMT)[18] 采用退化文本与视觉目标的组合重建源端或目标端的句子, 同样在测试阶段不需要输入图片. 因为该方法输入的图片为视觉目标, 因此也可以归类为视觉实体融合类方法.

4 实验结果与分析

4.1 超参数 ω 对翻译的影响

对于本文所提的多任务学习方法, 超参数 ω 是一个影响最终翻译性能的重要参数. 因此, 本文首先依据该参数在英译德验证集上的 Bleu 值结果来选取一个合适的值.

如图 3 所示实验结果, 横轴代表 ω 从 0.4 以 0.05 为间隔增加至 0.9, 纵轴为 NMT 模型在英译德验证集上的 Bleu 值. 该图反映出, 当 $\omega = 0.7$ 时翻译模型的性能最佳.

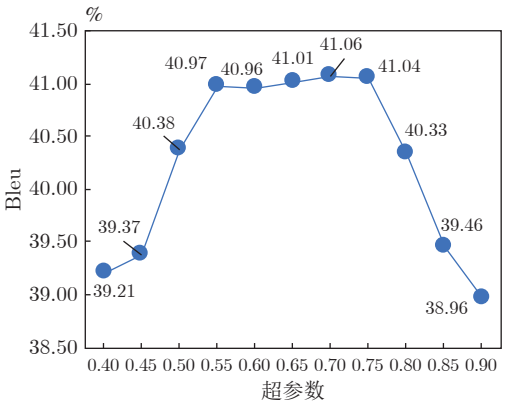

图 3 超参数 ω 对 CER-NMT 翻译性能的影响

Fig. 3 Effect of hyperparameter ω on translation performance of CER-NMT

4.2 翻译结果

表 1 显示了采用 CER 方法的 NMT 系统的翻译结果. 其中, ω 按照第 4.1 节的结论取值为 0.7, 对应的 VER、TER 和 TNER 三个子任务的训练比重分别为 12%、12% 和 6%.

表 1 中加粗项表示整列中的最佳结果. 根据以上英译德和英译法的实验结果可以得到以下结论:

1) 本文所提方法在 Multi30K Test2016 测试集的两种语言对上取得了 Bleu 和 Meteor 的最佳性能, 并且在 Test2017 测试集上的结果超过了多数模型的结果, 并与其他模型的最佳结果相比差距很小. 这说明 CER 方法能够有效提升 NMT 模型的翻译质量.

2) CER-NMT 在歧义词较多的 Ambiguous MSCOCO 上的表现落后于 GMMT 和 EMMT 两个模型. 这是因为 CER 方法主要帮助 NMT 在训练阶段融合视觉信息, 在测试阶段因为不需要输入图像使得模型无法借助视觉信息解决歧义词的问题.

3) GMMT、EMMT 和 CER-NMT 同为实体

表 1　MMT 模型在 Multi30K 以及 Ambiguous MSCOCO 上的英译德和英译法的翻译结果

Table 1　Results of MMT models on the English-German Multi30K and English-French Ambiguous MSCOCO

模型	英译德						英译法	
	Test2016		Test2017		MSCOCO		Test2016	
	B	M	B	M	B	M	B	M
句子级融合方法								
IMG_D	37.3	55.1	—	—	—	—	—	—
VMMT_C	37.5	55.7	26.1	45.4	21.8	41.2	—	—
SerAttTrans	38.7	57.2	—	—	—	—	60.8	75.1
GumAttTrans	39.2	**57.8**	31.4	51.2	26.9	46.0	—	—
视觉实体融合方法								
Parallel RCNNs	36.5	54.1	—	—	—	—	—	—
DelMMT	38.0	55.6	—	—	—	—	59.8	74.4
GMMT	39.8	57.6	32.2	51.9	28.7	**47.6**	60.9	74.9
增强 NMT 方法								
Imagination	36.8	55.8	—	—	—	—	—	—
VMMT_F	37.7	56.0	30.1	49.9	25.5	44.8	—	—
EMMT	39.7	57.5	**32.9**	51.7	**29.1**	47.5	61.1	75.8
本文方法								
Base	38.5	57.5	31.0	51.9	27.5	47.4	60.5	75.6
CER-NMT	**40.2**	**57.8**	32.5	**52.0**	28.3	47.1	**61.6**	**76.1**

级方法, 结果均优于传统方法. 这说明 NMT 模型融入更细粒度的视觉信息时效果更好, 传统句子级方法对视觉信息利用效率较差.

综合以上的实验结果, 本文所提的跨模态实体重构方法能够有效提升机器翻译的质量.

4.3　消融实验

为了探究 VER、TER 和 TNER 三个子任务对 CER-NMT 模型的影响, 本节设置了 12 组消融实验. 其中序号 0 代表第 4.2 节中 CER-NMT 的结果. 第 1~3 组各去掉一个子任务, 并保持剩余子任务的训练权重; 第 4 组~6 组各去掉一个子任务, 保持 NMT 的权重; 第 7~9 组各保留一个子任务, 保持子任务的权重; 第 10~12 组各保留一个子任务, 保持 NMT 的权重.

实验结果如表 2 所示, 其中, "—"代表去掉所对应的子任务. 从表 2 中可以得出:

1) 由第 1~6 组与第 7~12 组的对比可以看出, 子任务组合的方式要优于仅使用单一子任务. 其中, VER 和 TER 的组合已经可以使 NMT 达到很好的结果. TNER 对 NMT 的影响最小, 但是依旧可以为 NMT 模型带来小幅度的提升.

2) $\omega > 0.8$ 的实验组结果与第 4.1 节中的实验结果均说明减少跨模态任务至一定比重后, 模型的性能将逐渐趋近于 NMT.

3) 与第 4.2 节中 CER-NMT 的结果相比可以说明, VER、TER 和 TNER 三个子任务共同配合可以使 NMT 的性能达到最佳.

表 2　在 Multi30K Test2016 英译德翻译任务上的消融实验

Table 2　Ablation study on the English-German Multi30K Test2016

序号	NMT	VER	TER	TNER	B
	ω	$(1-\omega)\times\alpha$	$(1-\omega)\times\beta$	$(1-\omega)\times\gamma$	
0	0.70	0.12	0.12	0.06	40.2
1	0.76	0.12	0.12	—	40.0
2	0.82	0.12	—	0.06	39.5
3	0.82	—	0.12	0.06	39.6
4	0.70	0.15	0.15	—	39.9
5	0.70	0.20	—	0.10	39.2
6	0.70	—	0.20	0.10	39.3
7	0.88	0.12	—	—	38.8
8	0.88	—	0.12	—	38.8
9	0.94	—	—	0.06	39.0
10	0.70	0.30	—	—	39.2
11	0.70	—	0.30	—	39.4
12	0.70	—	—	0.30	39.0

4.4　文本实体忠实度

本文所提方法主要将视觉信息与文本实体相融合, 这使得视觉信息具有更明确的作用方向. 因此

检验视觉信息是否对文本实体的翻译产生影响成为一个必要的环节. 本文尝试测量在解码生成目标单词时对文本实体的忠实度来反映模型的行为变化.

Transformer 的解码器采用的是交叉注意力机制, 与一般的注意力机制类似的是, 在解码过程中通过给源端的词不同的“权重” 来达到“关注” 或“忽视” 的作用. 该“权重” 体现了当前要解码的目标端单词对源端单词所提供信息的需求程度. 因此, 本文选择 Transformer 解码器最后一层交叉注意力权重的多头平均值为生成目标端词时对源端词的注意力“权重”, 并定义该“权重” 为忠实度 (Fidelity), 用于量化生成目标端文本实体时对源端文本实体的忠实程度. 第 2.1 节中提到, 源端的文本实体是通过文本分析工具 spaCy 提取得到. 本节中所要确定的目标端文本实体是通过 fast-align[38] 对齐工具对齐源端与目标端单词得到的. 为了得到一个较好的对齐结果, 笔者将测试集与训练集拼接后训练对齐模型.

实验结果如图 4 所示, 图中横轴代表测试集中源端文本实体以某种方式的排序, 纵轴为范围从 0 到 1 的忠实度. 每个小像素点代表一个源端文本实体在一个翻译句子样本中对应目标端文本实体的注意力权重值, 即实体词忠实度. 大圆点代表每个实体的平均忠实度. 图 4(a) 为纯文本 Transformer 的测试结果, 横轴为测试集中的 1 110 个源端文本实体按照平均忠实度由小到大排序. 图 4(b) 和图 4(c) 为 CER-NMT 的结果, 其中图 4(b) 的横轴排序与图 4(a) 保持一致. 图 4(d) 为第 4.3 节第 12 组仅设置 TNER 的模型. 图中“avg”代表平均值. 从 4 个图的对比中可以得到:

1) 图中忠实度为 0 的横线部分代表对齐模型无法在目标端句子中找到对应的实体词, 因此无法确定其忠实度.

2) 图 4(b) 相比于图 4(a) 存在更多靠近 1.0 的小像素点. 从图 4(c) 中的均值结果 (avg) 可以看到, 小像素点的均值从 0.4238 提升至 0.4498, 大圆点的均值从 0.4024 提升至 0.4478, 均具有较明显的提升. 该数值结果表明 CER 方法能够明显地提升 NMT 在翻译过程中对源端文本实体的忠实度.

3) 图 4(c) 经过重排序实体的顺序后可以更直

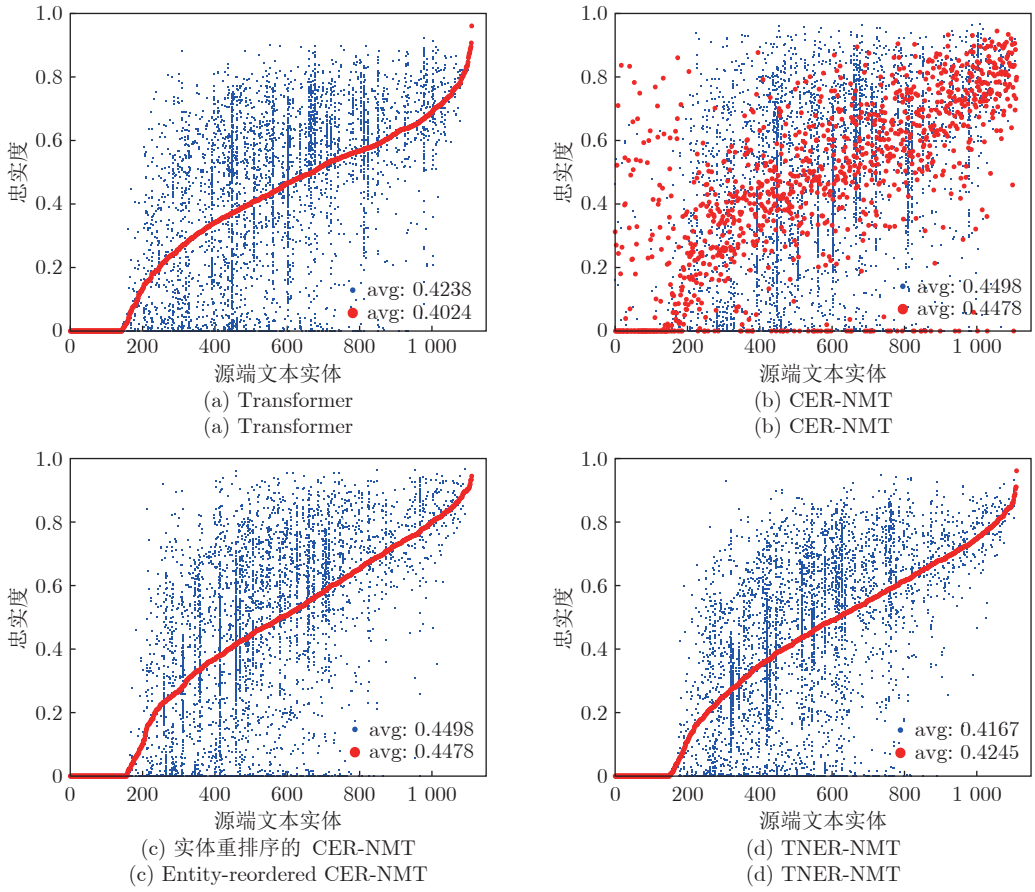

图 4 文本实体在不同模型下的忠实度

Fig. 4 The fidelity of textual entities on different models

观地从曲线的趋势看出, 采用 CER 方法所带来的忠实度的提升.

4) TNER 对文本上下文和视觉实体进行句子级的语义融合, 可以观察到图 4(d) 相对于图 4(a) 有提升也有降低, 这说明 VER 和 TER 才是帮助提升文本实体忠实度的主要因素, 非实体重构方法所带来的翻译性能提升无法带来显著的实体忠实度的提升.

以上结果表明, CER 通过融入视觉信息的方式, 增加了翻译模型在翻译过程中对源端文本实体的忠实度, 从而使得翻译结果得到了进一步的提升. 该实验同样表明本文所提的显式融合视觉信息的方式是有效可行的, 该显式方法使得模型更具备可解释性, 视觉信息的作用方式更有迹可循.

5 结束语

本文提出了一种跨模态实体重构方法用于探究以显式方式融合视觉信息与文本信息的可行性. 在翻译性能方面, 实验结果表明, CER-NMT 能够在英译德和英译法两个数据集上达到更高的翻译准确率. 在消融实验中发现, 视觉实体重构、文本实体重构以及文本非实体重构三种重构方法组合后 NMT 模型从视觉信息中获益最大. 最后, 本文尝试验证该显式方法的可解释性, 实验结果表明跨模态实体重构方法显著地增加了模型对源端文本实体的忠实度, 从而带来翻译质量的提升.

References

1 Barrault L, Bougares F, Specia L, Lala C, Elliott D, Frank S. Findings of the third shared task on multimodal machine translation. In: Proceedings of the 3rd Conference on Machine Translation: Shared Task Papers. Brussels, Belgium: Association for Computational Linguistics, 2018. 304–323

2 Elliott D, Frank S, Barrault L, Bougares F, Specia L. Findings of the second shared task on multimodal machine translation and multilingual image description. In: Proceedings of the 2nd Conference on Machine Translation. Copenhagen, Denmark: Association for Computational Linguistics, 2017. 215–233

3 Elliott D, Frank S, Sima′an K, Specia L. Multi30K: Multilingual English-German image descriptions. In: Proceedings of the 5th Workshop on Vision and Language. Berlin, Germany: Association for Computational Linguistics, 2016. 70–74

4 Calixto I, Liu Q. Incorporating global visual features into attention-based neural machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: Association for Computational Linguistics, 2017. 992–1003

5 Elliott D, Kádár Á. Imagination improves multimodal translation. In: Proceedings of the 8th International Joint Conference on Natural Language Processing. Taipei, China: Asian Federation of Natural Language Processing, 2017. 130–141

6 Zhou M Y, Cheng R X, Lee Y J, Yu Z. A visual attention grounding neural model for multimodal machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, 2018. 3643–3653

7 Toyama J, Misono M, Suzuki M, Nakayama K, Matsuo Y. Neural machine translation with latent semantic of image and text [Online], available: https://arxiv.org/pdf/1611.08459.pdf, November 25, 2016

8 Calixto I, Rios M, Aziz W. Latent variable model for multimodal translation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, 2019. 6392–6405

9 Calixto I, Liu Q, Campbell N. Doubly-attentive decoder for multi-modal neural machine translation. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada: Association for Computational Linguistics, 2017. 1913–1924

10 Libovický J, Helcl J. Attention strategies for multi-source sequence-to-sequence learning. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada: Association for Computational Linguistics, 2017. 196–202

11 Libovický J, Helcl J, Mareček D. Input combination strategies for multi-source transformer decoder. In: Proceedings of the 3rd Conference on Machine Translation: Research Papers. Brussels, Belgium: Association for Computational Linguistics, 2018. 253–260

12 Yao S W, Wan X J. Multimodal transformer for multimodal machine translation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Seattle, USA: Association for Computational Linguistics, 2020. 4346–4350

13 Huang P Y, Liu F, Shiang S R, Oh J, Dyer C. Attention-based multimodal neural machine translation. In: Proceedings of the 1st Conference on Machine Translation. Berlin, Germany: Association for Computational Linguistics, 2016. 639–645

14 Elliott D. Adversarial evaluation of multimodal machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, 2018. 2974–2978

15 Wu Z Y, Kong L P, Bi W, Li X, Kao B. Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Association for Computational Linguistics, 2021. 6153–6166

16 Li J D, Ataman D, Sennrich R. Vision matters when it should: Sanity checking multimodal machine translation models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021. 8556–8562

17 Caglayan O, Madhyastha P, Specia L, Barrault L. Probing the need for visual context in multimodal machine translation. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, Minnesota, USA: Association for Computational Linguistics, 2019. 4159–4170

18 Huang X, Zhang J J, Zong C Q. Entity-level cross-modal learning improves multi-modal machine translation. In: Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021. Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021. 1067–1080

19 Long Q Y, Wang M X, Li L. Generative imagination elevates machine translation. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021. 5738–5748

20 Wang S N, Zhang J J, Zong C Q. Associative multichannel autoencoder for multimodal word representation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational

Linguistics, 2018. 115−124

21 Wang S N, Zhang J J, Zong C Q. Learning multimodal word representation via dynamic fusion methods. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence and the 30th Innovative Applications of Artificial Intelligence Conference and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence. New Orleans, USA: AAAI, 2018. Article No. 733

22 Agrawal A, Lu J S, Antol S, Mitchell M, Zitnick C L, Parikh D, et al. VQA: Visual question answering. *International Journal of Computer Vision*, 2017, **123**(1): 4−31

23 Li H R, Zhu J N, Ma C, Zhang J J, Zong C Q. Multi-modal summarization for asynchronous collection of text, image, audio and video. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: Association for Computational Linguistics, 2017. 1092−1102

24 Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc., 2017. 6000−6010

25 Wang D X, Xiong D Y. Efficient object-level visual context modeling for multimodal machine translation: Masking irrelevant objects helps grounding. In: Proceedings of the 35th AAAI Conference on Artificial Intelligence, the 33rd Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, the 11th Symposium on Educational Advances in Artificial Intelligence. AAAI, 2021. 2720−2728

26 Yin Y J, Meng F D, Su J S, Zhou C L, Yang Z Y, Zhou J, et al. A novel graph-based multi-modal fusion encoder for neural machine translation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Seattle, USA: Association for Computational Linguistics, 2020. 3025−3035

27 He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 770−778

28 Yang Z Y, Gong B Q, Wang L W, Huang W B, Yu D, Luo J B. A fast and accurate one-stage approach to visual grounding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, South Korea: IEEE, 2019. 4682−4692

29 Koehn P, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, et al. Moses: Open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions. Prague, Czech Republic: Association for Computational Linguistics, 2007. 177−180

30 Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany: Association for Computational Linguistics, 2016. 1715−1725

31 Kingma D P, Ba J. Adam: A method for stochastic optimization [Online], available: https://arxiv.org/pdf/1412.6980.pdf, July 23, 2015

32 Papineni K, Roukos S, Ward T, Zhu W J. Bleu: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, USA: Association for Computational Linguistics, 2002. 311−318

33 Denkowski M, Lavie A. Meteor universal: Language specific translation evaluation for any target language. In: Proceedings of the 9th Workshop on Statistical Machine Translation. Baltimore, USA: Association for Computational Linguistics, 2014. 376−380

34 Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, 2014. 1724−1734

35 Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Computation*, 1997, **9**(8): 1735−1780

36 Liu P B, Cao H L, Zhao T J. Gumbel-attention for multi-modal machine translation [Online], available: https://arxiv.org/pdf/2103.08862.pdf, March 16, 2021

37 Ive J, Madhyastha P, Specia L. Distilling translations with visual awareness. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, 2019. 6525−6538

38 Dyer C, Chahuneau V, Smith N A. A simple, fast, and effective reparameterization of IBM model 2. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Atlanta, Georgia, USA: Association for Computational Linguistics, 2013. 644−648

黄 鑫 中国科学院自动化研究所模式识别国家重点实验室博士研究生. 主要研究方向为多模态机器翻译. E-mail: xin.huang@nlpr.ia.ac.cn

(**HUANG Xin** Ph.D. candidate at the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His main research interest is multi-modal machine translation.)

张家俊 中国科学院自动化研究所研究员, 中国科学院大学岗位教授. 主要研究方向为机器翻译和自然语言处理. E-mail: jjzhang@nlpr.ia.ac.cn

(**ZHANG Jia-Jun** Professor at the Institute of Automation, Chinese Academy of Sciences, and professor at University of Chinese Academy of Sciences. His research interest covers machine translation and natural language processing.)

宗成庆 中国科学院自动化研究所研究员, 中国科学院大学岗位教授, 中国计算机学会会士, 中国人工智能学会会士. 主要研究方向为自然语言处理, 机器翻译. 本文通信作者. E-mail: cqzong@nlpr.ia.ac.cn

(**ZONG Cheng-Qing** Professor at the Institute of Automation, Chinese Academy of Sciences, and an adjunct professor at University of Chinese Academy of Sciences. He is CCF Fellow and CAAI Fellow. His research interest covers natural language processing and machine translation. Corresponding author of this paper.)