摘　要

机器译文自动评价是指使用计算机技术对机器翻译系统输出译文的质量进行自动评价，它是机器翻译领域的一项重要研究任务，对机器翻译系统的优化起着指导作用。目前机器译文自动评价领域的研究主流为基于神经网络的机器译文自动评价。

最新的神经机器译文自动评价方法使用预训练语境词向量提取深层语义特征，并将它们直接拼接输入多层神经网络预测译文质量，其中直接拼接操作容易导致特征间缺乏深入融合；而逐层抽象进行预测时容易丢失细粒度准确匹配信息。针对以上问题，本文提出基于多元信息融合的神经机器译文自动评价方法，该译文自动评价方法引入中期信息融合方法和后期信息融合方法，使用拥抱融合对不同特征进行交互中期融合，基于细粒度准确匹配的句移距离和句级余弦相似度进行后期融合，实现细粒度准确匹配信息的引入和不同语义特征的高效融合。

另一方面，当前的机器译文自动评价主要通过大规模预训练语言模型直接提取机器译文和参考译文的语义表征后计算表征相似度，然而当前的预训练语言模型可能会将语义相近的句子映射到相距较远的稠密向量空间中。针对该问题，本文提出引入孪生相似特征的神经机器译文自动评价方法，该方法使用孪生网络结构对预训练语言模型进行微调，使其能够将语义相似的句子映射到相近的稠密向量空间中，从而更适用于机器译文自动评价领域。然后使用微调完成的孪生预训练语言模型提取语义相似特征，并将该特征引入神经机器译文自动评价方法中，以提升评价模型性能。

为了验证所提方法的有效性，在 WMT’21 Metrics Task 基准数据集上进行实验，实验结果表明，本文所提方法能有效提高其与人工评价的相关性，达到与参加评测最优系统的可比性能。

关键词：机器翻译；译文自动评价；信息融合；信息表征；拥抱融合

Abstract

Automatic evaluation of machine translation is an important research task in the field of machine translation, which refers to the quality evaluation of machine translation system output by computer technology. The guidance provided by automatic evaluation of machine translation is vital for optimizing machine translation systems. Currently, the mainstream approach to automatic evaluation of machine translation is based on neural network.

The lastest neural automatic evaluation methods of machine translation use pre-trained contextual embeddings to extract different deep semantic features, and then simply concatenate them feed into the multi-layer neural network to predict translation quality. Simply concatenated features results in lack of deep fusion between features. And fine-grained accurate matching information tends to be lost when layer by layer abstraction is used for prediction. To address these limitations, this paper proposes a new neural automatic evaluation method for machine translation based on multiple information fusion. Specifically, we introduce middle fusion and late fusion into machine translation evaluation. We propose to use embrace fusion to interactively fuse different features in the middle stage, and to fuse sentence mover's distance and sentence cosine similarity that are based on fine-grained accurate matching in the late stage.

Moreover, current automatic evaluation methods for machine translation utilize large-scale pre-trained language models to extract the semantic representations of machine translations and reference translations, and then calculate the similarity of these representations. However, current pre-trained language models may map semantically similar sentences into a dense vector space that is far away from each other. In this paper, we propose a new neural automatic evaluation method of machine translation based on siamese similarity feature. We use the siamese network structure to fine-tune the pre-trained language model, so that it can map sentences with similar semantics into a dense vector space with a closer distance, which is more suitable for the task of machine translation evaluation, and then use the fine-tuned siamese pre-trained language model to extract features of semantic similarity, which are incorporated into the neural machine translation evaluation method to improve its performance.

In order to verify the effectiveness of the proposed methods, extensive experiments are carried out on the WMT'21 Metrics Task, experimental results show that our proposed methods can effectively improve the correlation with human judgement, and achieve competitive performance with the best metrics in the evaluation campaign.

Key words: Machine translation; Automatic evaluation of machine translation; Information fusion; Information representation; Embrace fusion

目 录

1 绪论

1.1 研究背景及意义

机器翻译是指使用计算机技术将一种语言的语句转换为另一种语言的语句。机器翻译的技术更迭从早期使用词典匹配的方法到基于规则的方法，进一步发展为基于统计的机器翻译方法，随着人工智能技术的迅猛发展，机器翻译在经历多次技术革新之后实现了从基于统计的机器翻译(Statistical Machine Translation)到基于神经网络的机器翻译(Neural Machine Translation)的跨越，人工智能技术极大提高了机器翻译的性能。目前机器翻译已经被广泛应用于社会生活、生产之中。在对机器翻译系统进行研究改进的过程中，需要了解机器翻译系统的性能，即翻译系统输出的机器译文的质量情况，机器译文自动评价应运而生。

机器译文自动评价(Automatic Evaluation of Machine Translation)是指通过度量机器译文与参考译文的相似程度或偏离程度实现对机器译文质量的评价，进一步实现系统级别的翻译质量评价，机器翻译系统开发人员通过评价结果获知机器译文质量，从而有针对性地对翻译系统进行改进[1-4]。在机器译文自动评价出现之前，主要采用人工评价的方式对机器翻译系统输出的译文进行评价，人工评价尽管比较准确，但评价周期长、费用高且不客观。自 BLEU[5]等机器译文自动评价指标被提出以来，译文自动评价方法因其评价周期短、速度快、成本低等优点被大规模应用于评价机器译文的质量。因此机器译文自动评价对推动机器翻译的发展发挥着重要作用，对机器译文自动评价的研究在理论层面和应用层面都具有重要价值和意义。

1.2 国内外研究现状

早期的译文自动评价方法根据机器译文与参考译文的词形相似程度评价译文质量[5-7]，如基于 *n* 元文法匹配的方法和基于编辑距离的方法。基于 *n* 元文法匹配的方法计算参考译文与机器译文中不同长度词语片段的匹配程度，如 BLEU[5]、NIST[8]和 ROUGE[9]等；基于编辑距离的方法计算将机器译文转换为参考译文所需编辑次数的比例，如单词错误率 WER[10]和翻译错误率 TER[11]等。此

外，一些学者提出基于语言学检测点的方法，该类方法根据构建的语言学检测点对译文相应部分进行打分[12]，如 Woodpecker[13]等。随着人工智能的发展，基于传统机器学习的自动评价方法采用机器学习的“特征工程+任务建模”范式对译文的质量进行评价[14-16]，由人工指定影响译文质量的各类特征，使用支持向量机等传统机器学习算法预测机器译文质量，如 BEER[17]、Blend[18]等。

传统的自动评价方法根据词法、句法和浅层语义知识对译文进行评价。严格将词形进行匹配的方法很难准确评价包含词序变化和一词多义语言现象的译文的质量；而使用句法和浅层语义知识进行匹配的方法需要额外的语言学分析工具或特定的语言资源，这些语言学分析工具和资源与语言种类相关，很难移植到不同语种的译文上，导致其泛化性较差。

近年来，计算机计算性能的提升和可用数据规模的增加促进了神经网络的发展与应用，大规模预训练语言模型可以生成词语或句子的稠密向量表示，这些向量中蕴含丰富的语法、语义信息[1]。因此，基于神经网络的自动评价方法能根据语义评价机器译文的质量，并且该类方法性能表现优良，已成为当前领域内主流的研究方向。基于神经网络的自动评价方法根据评价方式不同分为基于表征匹配的方法和基于端到端神经网络的方法两大类，如图 1-1 所示。

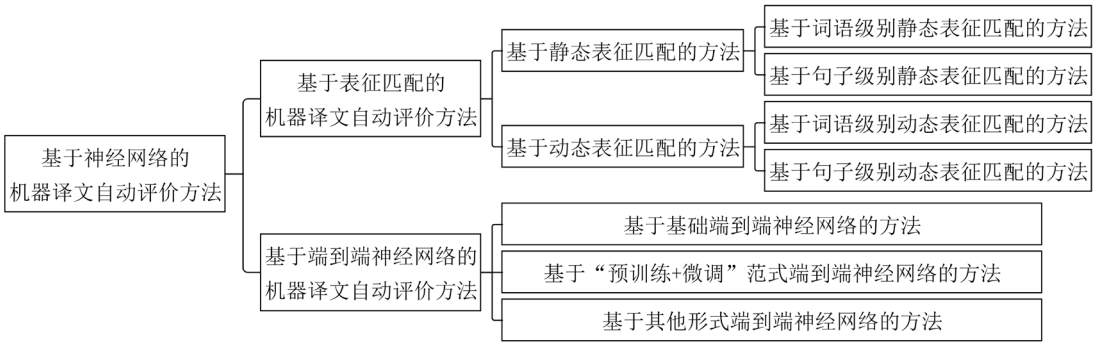

图 1-1 基于神经网络的机器译文自动评价方法分类一览图

其中，基于表征匹配的机器译文自动评价方法将机器译文与参考译文映射到稠密向量空间，将词语级别向量或句子级别向量作为机器译文和参考译文的词或句的表征，并计算二者表征的匹配程度，实现语义层面的匹配度评估。根据表征是否含上下文语境信息将其进一步分为基于静态表征匹配的方法和基于动态表征匹配的方法，基于静态表征匹配的方法使用静态预训练模型获取表征，基于动态表征匹配的方法使用含上下文语境信息的表征。基于端到端神经网络的机器译文自动评价方法使用神经网络提取句子的深层语义信息，将深层语义信息进行回归计算得到质量分数。基于端到端神经网络的方法可进一步分为基于基础端到端神经网络的方法、基于“预训练+微调”范式端到端神经网络的方法和基于其他形式端到端神经网络的方法。

1.2.1 基于表征匹配的机器译文自动评价方法

基于表征匹配的机器译文自动评价方法将词或句映射到高维空间，计算参考译文与机器译文的词语级别表征匹配程度或句子级别表征匹配程度，实现语义层面的质量评价，从而提升评价准确度。根据表征是否含上下文语境信息，将基于表征匹配的机器译文自动评价方法分为基于静态表征匹配的方法和基于动态表征匹配的方法。

(1)基于静态表征匹配的方法

基于静态表征匹配的方法使用静态预训练的词向量 GloVe 或 Word2Vec 等获取词表征，计算机器译文和参考译文中词表征的匹配相似度或偏离程度，或将词表征加工为句级表征后计算其匹配程度。根据用于匹配的表征粒度不同将基于静态表征匹配的方法分为基于词语级别静态表征匹配的方法和基于句子级别静态表征匹配的方法。

1)基于词语级别静态表征匹配的方法

基于词语级别静态表征匹配的方法使用静态预训练词表征生成模型获取机器译文和参考译文的词表征，然后计算二者的匹配程度。贪心匹配法[19]计算机器译文中所有词表征与参考译文中词表征的最大匹配余弦相似度、参考译文中所有词表征与机器译文中词表征的最大匹配余弦相似度，取二者均值作为评价分数，如式(1-1)~式(1-3)所示。

$$\mathrm{G}(t,r)=\frac{\sum_{w\in r}\max_{\hat{w}\in t}\mathrm{cosine}(\vec{w},\vec{\hat{w}})}{|r|} \tag{1-1}$$

$$\mathrm{G}(r,t)=\frac{\sum_{\hat{w}\in t}\max_{w\in r}\mathrm{cosine}(\vec{\hat{w}},\vec{w})}{|t|} \tag{1-2}$$

$$\mathrm{GM}=\frac{\mathrm{G}(t,r)+\mathrm{G}(r,t)}{2} \tag{1-3}$$

其中，$\mathrm{G}(t,r)$ 为参考译文中所有词表征的最大匹配余弦相似度的平均值，$\mathrm{G}(r,t)$ 为机器译文中所有词表征的最大匹配余弦相似度的平均值，t 为机器译文，r 为参考译文，w 为参考译文的词表征，$\hat{w}$ 为机器译文的词表征，$|r|$ 为参考译文中的词表征总数，$|t|$ 为机器译文中的词表征总数，评价分数 GM 为两个匹配方向的平均值。该方法在语义层面对机器译文和参考译文的词表征作映射匹配。

为了将浅层语义分析与语义匹配相结合，MEANT[20]使用语义角色标注给词或片段标注其在句子中的角色标签，通过测量机器译文和参考译文的语义框架与角色填充物的相似度评估翻译的充分度。MEANT2.0[21]在 MEANT 的工作基础上引入词频加权，赋予实词比功能词更高的权重，并通过计算 n 元词表征匹配相似度实现在评价时关注词序信息。MEE[22]分别对机器译文和参考译文进行精准词

形匹配(Exact Match)、根匹配(Root Match)和近义匹配(Synonym Match)，其中精准词形匹配为机器译文和参考译文的词形匹配数，根匹配和近义匹配设定匹配阈值，计算机器译文和参考译文的 FastText 词表征匹配相似度，FastText 词表征是指 Facebook 于 2016 年开源的词向量计算工具生成的词表征。根据匹配相似度所在的阈值空间判定其所属匹配类型。最终将以上三个匹配模块的 F 值加权平均为评价分数。不同于上述基于机器译文和参考译文的相似程度的质量评价方法，基于偏离程度的方法如词移距离 WMD[23]计算机器译文与参考译文词表征的最小匹配欧几里得距离。

2)基于句子级别静态表征匹配的方法

基于句子级别静态表征匹配的方法将机器译文和参考译文的词表征使用平均池化或其他处理方式加工为句子级别表征，然后计算句子级别表征间的相似程度。

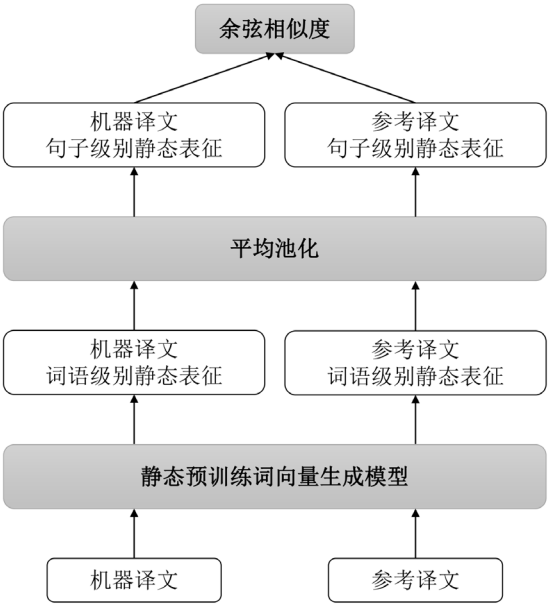

图 1-2 平均词向量自动评价方法图

如图 1-2 所示，平均词向量自动评价方法(Embedding Average Metric)[24]分别将机器译文和参考译文中的词表征通过平均池化加工为句子级别表征，计算句子级别表征的余弦相似度。为了增强句子级别向量的表征能力，极值向量法(Vector Extrem)[25]沿维度取所有词表征的最大值或最小值作为句子级别表征的各维度值。陈博兴等人[26]提出分别基于独热表征、分布式词表征、RAE 句子表征或上述三种表征的组合的译文自动评价方法，并在此基础上提出将句子级别的评分加权求和为篇章级别的评分[27]。其中，RAE 句子表征是使用贪心无监督递归自编码器策略(Recursive Auto-Encoder, RAE)生成的分布式句子表征。

相比仅根据词形进行评价的基于 *n* 元文法匹配的方法，基于静态表征匹配的方法在一定程度上实现了根据语义进行评价，但静态表征独立于上下文，无法获

取上下文语境信息，故基于静态表征匹配的方法存在无法结合语境信息进行译文质量评价的不足。

(2)基于动态表征匹配的方法

针对基于静态表征匹配的方法中静态表征无法获知语境信息这一问题，基于动态表征匹配的自动评价方法使用基于上下文语境的词表征获取语境信息。根据所采用的表征的粒度不同将基于动态表征匹配的方法分为基于词语级别动态表征匹配的方法和基于句子级别动态表征匹配的方法。

1)基于词语级别动态表征匹配的方法

基于词语级别动态表征匹配的自动评价方法计算机器译文和参考译文含语境信息的词向量的匹配相似度。

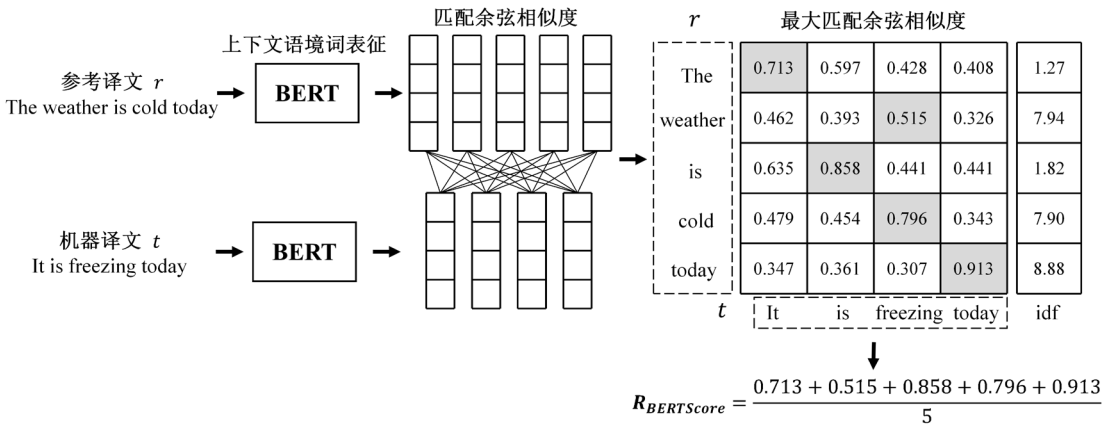

图 1-3 BERTScore 机器译文自动评价方法示意图(图片引自文献[28])

如图 1-3 所示，BERTScore[28]用 BERT 模型生成上下文语境词表征，计算参考译文中词表征 r 与机器译文中词表征 t 的最大匹配余弦相似度，计算召回率 $R_{BERTScore}$ 和准确率 $P_{BERTScore}$，进一步计算 F 值 $F_{BERTScore}$ 作为评价分数，如式(1-4)~式(1-6)所示。

$$R_{BERTScore} = \frac{1}{|r|}\sum_{r_i \in r} \max_{t_j \in t} r_i^T t_j \tag{1-4}$$

$$P_{BERTScore} = \frac{1}{|t|}\sum_{t_j \in t} \max_{r_i \in r} r_i^T t_j \tag{1-5}$$

$$F_{BERTScore} = 2 \cdot \frac{P_{BERTScore} \cdot R_{BERTScore}}{P_{BERTScore} + R_{BERTScore}} \tag{1-6}$$

Mathur 等人提出的 BERTr[29]与 BERTScore 类似，但仅使用召回率作为评价分数，方法简单有效。BERTScore 采用词表征间一对一的匹配余弦相似度，然而句子对中的词还存在一对多的关系，出于对该语言现象的考虑，Zhao 等人提出的 MoverScore[81]计算 n 元词组上下文语境词表征的欧几里得距离。由于对不同翻译难度的句子的翻译能力反映了翻译系统的质量情况，Zhan 等人提出的

DA-BERTScore[30]将翻译难度引入 BERTScore，赋予更难翻译的词以更高的评价权重，增加其对评价结果的影响。评判翻译难度的方法为机器译文与参考译文的词表征最大匹配余弦相似度越低，则翻译该词的难度越大，并赋予该词更高的难度系数。最后将难度系数作为最大匹配余弦相似度的权重参与到 F 值的计算中，该方法能有效对性能相近的优秀翻译系统进行质量排名。Vernikos 等人提出的 Doc-BERTScore[31]将 BERTScore 扩展为篇章级别自动评价，该方法将译文与该条译文的上下文一起输入 BERT 模型进行编码，使译文表征获得篇章级别上下文信息，然后以单条句子为单位进行评分，评分方法与 BERTScore 的评分方法相同。

2)基于句子级别动态表征匹配的方法

基于句子级别动态表征匹配的方法计算机器译文与参考译文含语境信息的句子表征的匹配程度。Wieting 等人提出的 SIMILE[32]使用经过训练的含软注意力机制的编码器[33]生成机器译文和参考译文的句子表征，计算二者余弦相似度，并引入长度惩罚因子以惩罚机器译文与参考译文长度相差过大的场景。长度惩罚因子 LP 计算方式如式(1-7)所示，其中，$|t|$指机器译文的长度，$|r|$指参考译文的长度。

$$\mathrm{LP}(r,t)=e^{1-\frac{\max(|r|,|t|)}{\min(|r|,|t|)}} \tag{1-7}$$

目前，世界上只有英德、英汉等少数语言对有丰富的语料资源，大多数语言对的语料资源匮乏。YiSi 系列评价指标[34]根据可获得的语料资源规模不同设计对应的自动评价指标。其中，YiSi-0 适用于低资源语言，计算机器译文和参考译文的最长公共字符子串；YiSi-1 计算使用 BERT 生成的上下文词表征的匹配余弦相似度，可自由选择是否使用语义角色标注获取浅层语义结构信息；YiSi-2 适用于无参考译文的评价场景，该方法使用跨语种词表征生成模型获取源语言句子和机器译文的跨语种词表征，然后计算二者余弦相似度，可自由选择是否使用语义角色标注。

近年，跨语种表征生成模型技术取得长足进步，一些学者使用 XLM[35]、LaBSE[36]等跨语种表征生成模型获取源语言句子和机器译文在同一语义空间内词语级别或句子级别的表征，对比源语言句子和机器译文在同一高维空间内的语义相似度。基于跨语种预训练表征生成模型的 LaBSE 文本相似度分数虽然性能优良，但所需的 GPU 等硬件资源开销大且模型复杂，Han 等人提出的 cushLEPOR[37]模型使用知识蒸馏学习 LaBSE 模型内部映射方式，用较低的资源开销实现接近 LaBSE 模型的性能。

基于表征匹配的机器译文自动评价方法计算机器译文与参考译文的表征匹配程度，一定程度实现语义层面的评价，该类方法依托预训练表征生成模型，随

着跨语种预训练表征生成模型技术的成熟，基于表征匹配的方法展现了较强的鲁棒性与易用性。

1.2.2 基于端到端神经网络的机器译文自动评价方法

基于端到端神经网络的机器译文自动评价方法使用神经网络提取深层语义信息，根据深层语义信息预测译文质量，根据神经网络架构不同将基于端到端神经网络的机器译文自动评价方法分为基于基础端到端神经网络的方法、基于“预训练+微调”范式端到端神经网络的方法和基于其他形式端到端神经网络的方法。

(1)基于基础端到端神经网络的方法

基于基础端到端神经网络的自动评价方法构建神经网络提取译文的深层语义信息后预测译文质量分数。

Shimanaka 等人提出的 RUSE[38]自动评价方法的结构如图 1-4 所示。

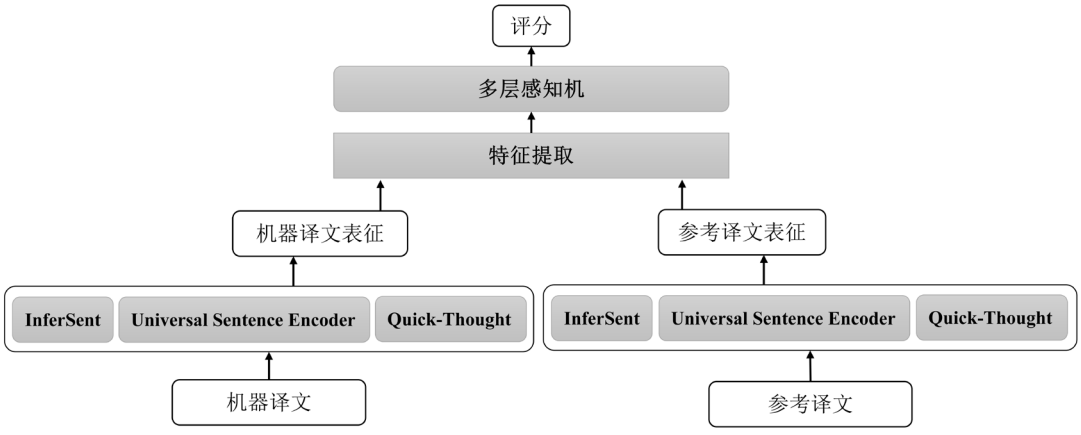

图 1-4 RUSE 评价方法结构图

RUSE 分别使用 InferSent、Universal Sentence Encoder 和 Quick-Thought 三种预训练句子表征生成模型生成机器译文和参考译文的句子级别表征，用启发式方法将句子表征组合后输入多层感知机(MLP)进行回归计算评分，如式(1-8)、式(1-9)所示。

$$\vec{s} = \text{InferSent}(s) \oplus \text{Quick-Thought}(s) \oplus \text{UniversalSentenceEncoder}(s) \quad (1\text{-}8)$$

$$\text{RUSE} = \text{MLP}(\vec{t};\vec{r};|\vec{t}-\vec{r}|;\vec{t}\odot\vec{r}) \quad (1\text{-}9)$$

其中，符号“$\oplus$”表示向量间的拼接操作，符号“$\odot$”表示向量间的逐元素相乘操作。$\vec{s}$ 表示由三种句子表征拼接构成的表征向量，$\vec{t}$ 表示机器译文的表征向量，$\vec{r}$ 表示参考译文的表征向量。

Mathur 等人提出的$(\text{BiLSTM+attention})_{\text{BERT}}$[29]方法将词向量输入双向长短时记忆网络 Bi-LSTM 获取上下文语境信息，使用跨句注意力机制获取机器译文和参考译文的交互信息。此外，Mathur 等人提出的$(\text{ESIM})_{\text{BERT}}$[29]方法使用自然语言

推理领域中的增强序列推理模型 ESIM[39]对机器译文和参考译文进行编码，使用跨句注意力机制对表征进行加权，并依次通过双向长短时记忆网络 Bi-LSTM 和池化层获取局部序列信息与特征信息提取，最后将加工完成的信息表征输入前馈神经网络预测译文质量分数，如式(1-10)、式(1-11)所示。

$$x = v_{r,avg} \oplus v_{r,max} \oplus v_{t,avg} \oplus v_{t,max} \tag{1-10}$$

$$(\mathrm{ESIM})_{\mathrm{BERT}} = \mathrm{ReLU}(x \cdot w + b) \cdot U + b' \tag{1-11}$$

其中，x 表示拼接完成后的句子增强表征，r 表示参考译文，t 表示机器译文，$v_{r,avg}$、$v_{r,max}$ 分别指参考译文的平均池化表征和最大池化表征，U、w、b 和 b' 为通过训练得到的参数。罗琪等人[40]在 Mathur 的工作基础上引入源端信息，使用质量估计模型从源语言句子和机器译文中提取句子级别质量向量，将句子级别质量向量与$(\mathrm{ESIM})_{\mathrm{BERT}}$ 模型中的增强表征拼接后输入前馈神经网络中预测译文评价分数。Hu 等人[41]在罗琪的工作基础上引入差异特征，差异特征的生成方式为将参考译文、源语言句子和机器译文两两组成三组句子对，然后使用跨语种预训练模型将三组句子对映射到同一语义空间，对比机器译文和源语言句子与参考译文的语义差异。

Rei 等人提出的 COMET[42][43]含两类评价模型，第一类为分数预测模型(Estimator Model)，该类模型对译文的质量进行评分；第二类为排名模型(Translation Ranking Model)，该类模型对译文质量进行排名，选出相对优质的译文。COMET 方法首先使用跨语种预训练语言模型 XLM-RoBERTa 分别对参考译文、机器译文和源语言句子进行编码。由于 Tenney 等人[44]实验表明预训练语言模型中的不同层会捕获不同类型的语义信息，且只依据模型最后一层的输出评判译文质量的效果不佳，故 COMET 使用分层注意力机制综合各层生成的不同类型的语义信息，使用平均池化将词语级别表征进一步处理为句子级别表征[45]，并在模型训练过程中采用层级 dropout[46]提高句子级别表征能力。

对于 COMET 中的分数预测模型(Estimator Model)，Rei 等人使用上述跨语种编码器分别对机器译文和参考译文进行编码，并采用类似 RUSE 中的方式对句子级别表征进行组合，如式(1-12)所示。

$$x = [t; r; t \odot r; t \odot s; |t - r|; |t - s|] \tag{1-12}$$

其中，t 表示机器译文，r 表示参考译文，符号“⊙”表示向量间的逐元素相乘操作，符号“–”表示向量间的逐元素相减操作，最后将组合完成的信息表征 x 输入前馈神经网络进行回归评分。

对于 COMET 中的排序模型(Translation Ranking Model)，Rei 等人将源语言句子 s、参考译文 r、相对优质的机器译文 $t+$、相对劣质的机器译文 $t-$ 的句子四元组$\{s, t+, t-, r\}$输入跨语种编码器，然后通过池化层生成四元组的句子级别信

息表征，使用三元组损失函数(Triplet Loss)优化语义空间中句子表征之间的相对距离，该损失函数的目标是优化模型使其在最终表征空间内相对优质的机器译文和黄金参考(参考译文与源语言句子)的距离更近、相对劣质机器译文和黄金参考的距离更远。除了分数预测模型和排序模型两个主要模型，Rei 等人还提出了直接对比源语言句子和机器译文的相似度、无需参考译文的 Reference-free COMET，轻量级的 COMET 模型 COMETINHO[47]。Vernikos 等人提出的 Doc-COMET[31]将译文与译文的上下文拼接后输入编码器，将 COMET 扩展为篇章级别的译文评价方法 Doc-COMET。

上述方法均为将含深层语义信息的向量作为神经网络的输入，另一类方法为将译文的各类特征分值作为神经网络的输入。REGEMT[48] 方法集成分别基于词形、句法和语义特征的自动评价指标，以此提升仅基于单种类型的自动评价指标性能。该方法集成的自动评价指标包括软余弦相似度、词移距离和词性标注转换距离，使用神经网络对各个特征分值进行回归预测分数。Rony 等人提出的 RoMe[49]将译文的语法、句法和语义三个方面的质量得分组合为向量输入神经网络回归计算评分，其中语义分数采用融入了词对齐和词序差异惩罚的基于语义相似度的 EMD 距离(Earth Mover's Distance)，EMD 距离可以计算机器译文和参考译文的偏离程度；句法分数采用经过改进的语义增强树编辑距离算法(Tree Edit Distance)[50]，计算机器译文和参考译文的句法结构差异；语法分数采用在 CoLA 语料库上训练的二分类器，判定译文语法是否在可接受范围内。

(2)基于“预训练+微调”范式端到端神经网络的方法

目前基于“预训练+微调”范式的深度学习模型被广泛应用于自然语言处理的各个任务，根据具体评价场景对包含大量可重用知识的预训练模型进行微调的机器译文自动评价模型展现出优异的性能。

Sellam 等人提出的 BLEURT[51]使用随机扰动后的维基百科句子和一组词汇级和语义级的监督信号对评价模型进行预训练，预训练监督信息包括：1) BLEU、ROUGE 和 BERTScore 自动评价指标评价结果；2)回译似然值；3)判断原句和扰动句的三类文本关系：蕴含、矛盾、中立；4)标注扰动句是否为原句回译生成的回译标志。拼接机器译文和参考译文输入预训练完成的 BERT 模型中，取特殊标志“[*CLS*]”位置的向量作为句子表征输入前馈神经网络预测译文质量分数。Wan 等人提出的 ROBLEURT[52]在 BLEURT 的工作基础上作三处优化提升模型的鲁棒性：1)根据源语言句子的资源可获得程度设计不同评价方式，在源语言句子资源匮乏的情况下仅拼接机器译文和参考译文作为模型的输入，在源语言句子资源充沛的情况下拼接参考译文、源语言句子和机器译文作为模型的输入，使模型能够在评价时综合考虑机器译文与黄金参考(参考译文和源语言句子)的语义一致性；2)使用大规模人工合成数据对模型进行持续性预训练；3)使用降噪后的数据对模

型进行微调。该自动评价方法结合单语模型和多语模型，使用“预训练+微调”范式进行训练，引入迁移学习，性能较 BLEURT 有进一步提升。

不同于基于基础端到端神经网络的RUSE方法和$(ESIM)_{BERT}$方法中将机器译文和参考译文分别输入 BERT 模型，BERT for MTE[53]将机器译文和参考译文拼接后输入 BERT 进行编码，将特殊位置“[*CLS*]”的向量输入多层感知机预测译文质量，并通过微调提升模型性能，如图 1-5 所示。其中，“[*SEP*]”为句子间的分割符号，“[*CLS*]”为每对输入的标识符。

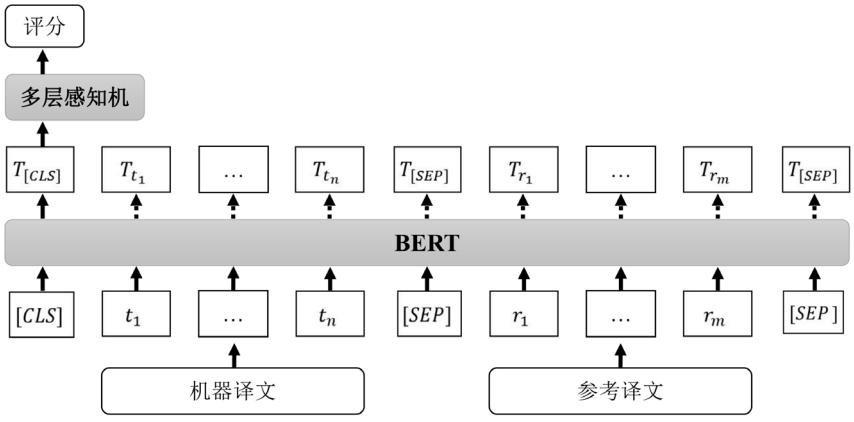

图 1-5 BERT for MTE 自动评价方法结构图(图片引自文献[53])

Kane 等人提出的 NUBIA[54]利用大规模预训练语言模型提取译文深层语义特征，并在提取特征时遵循 “预训练+微调”学习策略，该方法的评价过程分为三个步骤：第一步，分别用 RoBERTa STS、RoBERTa MNLI 和 GPT-2 模型抽取句子间的语义相似度、逻辑一致程度和语法正确性三类特征。具体来说，使用 STS-B-benchmark 数据集对 RoBERTa 预训练模型进行微调，提取机器译文和参考译文的语义相似度；用 RoBERTa 模型在 GLUE 的 MNLI 任务上微调，捕获机器译文和参考译文的逻辑一致程度；依托 GPT-2 模型计算困惑度，评判机器译文的语法正确性。第二步，将第一步抽取的特征输入线性回归模型，预测译文质量分数。第三步，将译文质量分数进行归一化。

为了减少硬件资源开销、提升模型效率，Eddine 等人的 FrugalScore[55]使用知识蒸馏构建轻量版 BERTScore 或 MoverScore。该自动评价模型先让轻量级预训练语言模型学习高开销模型的内部映射方式，然后在合成数据集上继续训练该轻量级预训练语言模型，最后在人工标注的语料上微调微缩模型。

(3)基于其他形式端到端神经网络的方法

以上方法均为构建神经网络提取深层语义信息，使用监督学习方式训练评价模型，通过回归方式预测机器译文质量。近年，一些新形式的自动评价模型被陆续提出，如 Thompson 和 Post 提出的 Prism[56]使用端到端释义模型预测机器译文在对应参考译文下出现的概率，概率值越大，则机器译文的质量越高。Vernikos

等人提出的 Doc-Prism[31]为篇章级别 Prism，该方法将参考译文与其上下文拼接输入端到端释义模型。Krubiński 等人提出的 MTEQA[57]是首个基于问答框架的机器译文自动评价指标，该指标的评价过程分为两个步骤：第一步，从参考译文中抽取信息作为答案，并生成相应的问题；第二步，使用问答系统根据机器译文生成上一步骤中问题的答案，用字符串比较法计算依据机器译文而得的答案和依据参考译文而得的答案的相似度，对于同一语段，取所有问题答案对相似度的平均值作为最终质量评分。

在易用性方面，基于端到端神经网络的机器译文自动评价方法在使用时需要根据模型的需求进行环境配置，虽然相关研究人员对基于端到端神经网络展开了大量研究，但当前可直接使用的基于端到端神经网络的自动评价模型较少，故相比其他方法，该类方法易用性较差，未来应当对性能优良的端到端神经网络评价模型的易用性提升进行深入研究。

1.2.3 自动评价方法的评测(元评测)

机器译文自动评价方法的评测(元评测)是指对机器译文自动评价方法性能进行评测。目前机器译文自动评价评测活动主要为 WMT 机器译文自动评价任务。国内的全国机器翻译大会 CCMT 组织过多次机器翻译相关任务评测，包括无需参考译文的机器译文质量估计评测活动。WMT 机器译文自动评价任务于 2008 年开始，用于评测机器译文自动评价方法的性能表现，任务涵盖中英、德英、中俄等被广泛使用的语言对和部分低资源语言对[58-64]。机器译文自动评价评测活动为不同自动评价指标提供公平比较的平台，它发布公开的数据集、基准方法，促进了机器译文自动评价的研究与发展。

评测活动中，为了比较参与评测的不同自动评价方法的优劣，一般使用肯德尔相关系数度量自动评价方法打分在句子级别与人工评价的相关性，使用皮尔逊相关系数度量自动评价方法打分在系统级别与人工评价的相关性，有时使用成对精确度度量在系统级别自动评价打分与人工评价的相关性，相关性越高，表示对应方法越可靠。

(1)肯德尔相关系数τ (Kendall Correlation)：通过度量自动评价与人工评价对译文质量高低排序一致程度衡量自动评价方法与人工评价的相关性，计算方法如下所示：

$$\tau = \frac{Concordant - Discordant}{Concordant + Discordant} \tag{1-13}$$

其中，*Concordant* 指自动评价方法给人工评价打分较高的机器译文以较高的分数，自动评价与人工评价打分一致；*Discordant* 指给人工评价打分较低的机器译文以较高的分数，自动评价与人工评价打分不一致。

(2)皮尔逊相关系数r_{xy} (Pearson Correlation)：通常用于衡量系统级别自动评价与人工评价的相关性。该相关系数的计算方式为首先按式(1-14)、式(1-15)方式分别计算自动评价x和人工评价y的均值$\bar{x}$和$\bar{y}$及方差s_x和s_y，然后进一步计算皮尔逊相关系数r_{xy}，如式(1-16)所示。

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i \tag{1-14}$$

$$s_x^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \tag{1-15}$$

$$r_{xy}=\frac{\sum_i(x_i-\bar{x})(y_i-\bar{y})}{(n-1)s_x s_y} \tag{1-16}$$

(3)成对精确度(Pairwise Accuracy)：用于衡量自动评价与人工评价的系统级别相关性，计算方式如式(1-17)所示。

$$\text{Pairwise Accuracy}=\frac{\left|\text{sign}(metric\Delta)=\text{sign}(human\Delta)\right|}{\left|All\ System\ Pairs\right|} \tag{1-17}$$

其中，自动评价(*metric*)和人工评价(*human*)分别对多个系统进行打分，对于其中任意两个系统，*metric*Δ指自动评价的评分差值，*human*Δ指人工评价的评分差值，|*All System Pairs*|指系统对的总数，通过比较评分差值是否一致衡量自动评价与人工评价的相关性。

元评测通过计算自动评价指标评分与人工评价评分的相关性度量自动评价指标的性能，故人工评价分数的可靠性直接决定了元评测是否有效，许多学者对元评测中的人工评价评分机制进行研究与探索，以期得到更可靠的人工评分，目前主要的人工评价方式为以下四种：

(1)传统 DA 人工评价(Direct Assessment)：该评价机制采用众包的方式对机器译文直接进行评分，由于其成本较低，2020 年及之前历届 WMT 自动评价任务均采用该人工评价方式。但近年研究发现，众包评分者由于缺乏专业翻译知识，存在对翻译中的错误过于包容、众包评分与专家评分相关性较低[65]等问题，故 2021 年 WMT 自动评价任务提出采用 MQM 评价机制作为人工评价分数的评测子任务。

(2)HTER(Human-mediated Translation Edit Rate)[66]：HTER 在翻译编辑率(TER)的基础上引入人工注解，让精通目标语言的人工译员结合机器译文和参考译文给出一个新的参考译文，使用 TER 算法计算机器译文和新参考译文的编辑率。其中，翻译编辑率(TER)计算从机器译文转换到参考译文所需的删除、单词替换、插入和词组平移的编辑次数的比例。

(3)多维度质量评价机制 MQM(Multidimensional Quality Metric)[67]：Freitag

等人的研究显示[64]传统众包 DA 人工评价对高质量机器翻译译文的评价不可靠。为了得到更可靠的人工评分，MQM 评价机制将翻译错误分为不同类型，综合错误的次数及其相应权重对译文进行评分，该方法比直接为译文评定一个分数更可靠，2021 年 WMT 自动评价任务开始采用 MQM 评价机制作为参考评分。MQM 评价机制将译文错误分为微小错误(*minor*)、主要错误(*major*)和严重错误(*crit*)，并赋予不同程度的错误不同的权重，按式(1-18)方式计算译文评分，其中，*SentenceLength* 表示句子长度，I_{minor}、I_{major} 和 I_{crit} 分别表示微小错误次数、主要错误次数和严重错误次数。

$$\mathrm{MQM}=100-\frac{I_{minor}+5\times I_{major}+10\times I_{crit}}{SentenceLength\times 100} \tag{1-18}$$

(4)分级质量度量指标 SQM(The Scalar Quality Metric)[68]：Freitag 等人受 MQM 启发，将机器译文质量分为六个等次，评价者在评分过程中可以看到句子的上下文。其中，质量分数为 6 分时指语法与语义完全正确；4 分为语义基本转述完成，语法错误较少；2 分为未表达源语言句子的主要语义；0 分为译文没有表达任何源语言句子的信息。

2019 年以来，每届的 WMT 自动评价任务含不同子任务，如 2019 年和 2020 年发布篇章级自动评价任务、2021 年新增专家多维度质量评价机制 MQM 作为人工评价的子任务，帮助自动评价研究人员准确了解自动评价模型性能、对比不同评价模型性能。

历届 WMT 自动评价任务的评测结果均整理成文并发表，研究人员可以通过每年的评测报告了解各个自动评价方法在该年评测任务中的表现及自动评价领域最新趋势。为了解近年评测任务中表现优良的自动评价方法的共同特点，WMT'21 部分自动评价子任务上获最优性能的评价方法汇总如表 1-1 所示。

表 1-1 WMT'21 metrics task 上获最优性能的自动评价方法汇总(表格引自文献[64])

自动评价指标	获最优性能的次数总计	语言对			粒度		评测数据集		
		英语-德语	英语-俄语	中文-英语	系统级别	句子级别	news w/o HT	news w/ HT	TED
C-SPECpn	**11**	4	3	4	6	5	3	5	3
bleurt-20	**10**	4	5	1	4	6	4	3	3
COMET-MQM_2021	**10**	3	3	4	3	7	3	2	5
tgt-regEMT	4	1	1	2	3	1	2	1	1
RoBLEURT*	3	-	-	3	1	2	1	-	2
cushLEPOR(LM)	2	1	-	1	2	-	1	-	1
BERTScore	2	1	1	-	2	-	1	-	1
Prism	2	-	2	-	2	-	1	-	1

续表

自动评价指标	获最优性能的次数总计	语言对			粒度		评测数据集		
		英语-德语	英语-俄语	中文-英语	系统级别	句子级别	news w/o HT	news w/ HT	TED
YiSi-1	2	-	2	-	2	-	1	-	1
MEE2	2	2	-	-	2	-	1	-	1
BLEU	1	1	-	-	1	-	1	-	-
hLEPOR	1	-	1	-	1	-	-	-	1
MTEQA	1	-	-	-	1	-	-	-	1
TER	1	-	-	-	1	-	-	-	1
chrF	1	-	-	-	1	-	-	-	1

其中，部分自动评价指标名称旁标注的符号“*”表示该方法未参与所有语言对上的评测，符号“-”表示该方法在该类任务上未取得最优性能。结果表明，显著优于其他自动评价方法的 C-SPECpn[69]、bleurt-20 和 COMET-MQM_2021 均为遵循“大规模预训练+微调”范式的端到端神经网络自动评价模型，这表明遵循“大规模预训练+微调”学习范式能显著提升评价性能。

在国内机器译文自动评价研究方面，澳门大学的 NLP2CT 实验室与阿里巴巴达摩研究院共同提出的 RoBLEURT 在 WMT’21 的自动评价任务中取得多项第一的优良成绩。中国科学院的马青松团队提出的 Blend、DPMFCOMB[70]和基于融合策略的机器翻译自动评价方法[71]性能优良，其中 Blend 在 WMT’17 自动评价任务的德英、俄英等多个语言对任务上取得第一名，DPMFCOMB 在 WMT’16 自动评价任务的法语至英语、土耳其语至英语句子级别直接评价任务中排名第一。北京大学的研究团队在 2020 年提出引入语义加权句子相似度的自动评价方法 SWSS[72]，该方法能效提升基于词形匹配的机器译文自动评价指标的性能。北京大学计算语言学重点实验室提出的 Meteor++[73]与 Meteor++ 2.0[74]对经典自动评价指标 Meteor 作改进，其中 Meteor++ 2.0 在 WMT’15 至 WMT’17 自动评价任务数据集上与人工评价的相关性超过了当时所有版本的 Meteor。苏州大学的李良友、贡正仙等人提出的融合文档信息的机器翻译自动评价[75]以语言学短语为基本评价单位，研究了文档信息在评价方法中的应用。江西师范大学的研究团队[76-78]提出的 MPEDA 在 WMT’16 自动评价系统级别任务的法语至英语和芬兰语至英语语言对上排名第二。

1.3 本文的研究内容与贡献

本文首先介绍了近年来国内外基于神经网络的机器译文自动评价的相关研

究方法与机器译文自动评价的评测平台和相关评测指标，并提出基于多元信息融合的神经机器译文自动评价方法与引入孪生相似特征的神经机器译文自动评价方法。

本文提出的基于多元信息融合的神经机器译文自动评价方法是为了缓解当前神经机器译文自动评价方法中直接拼接语义特征导致的特征间缺乏深入融合的问题和使用神经网络对语义特征进行抽象过程中丢失细粒度准确匹配信息的问题。该方法将中期信息融合和后期信息融合引入神经机器译文自动评价。其中，中期融合采用拥抱融合[79]高效融合不同信息表征；后期融合将基于细粒度准确匹配的句级余弦相似度[80]与句移距离[81]和中期融合后的神经网络模型评分结果融合，实现评价过程中同时关注深层语义信息和细粒度准确匹配信息，使自动评价模型的性能得到进一步提高。

本文提出的引入孪生相似特征的神经机器译文自动评价方法针对当前的预训练语言模型可能会将语义相似的句子映射到相距较远的稠密向量空间中的问题，采用孪生网络结构对预训练语言模型进行微调，使预训练语言模型能将语义相似的句子映射到相近的稠密向量空间中，从而让预训练语言模型更适用于机器译文自动评价任务，然后使用微调完成的孪生预训练模型提取孪生相似特征，将孪生相似特征引入神经机器译文自动评价方法中，使自动评价模型的性能得到一定提升。

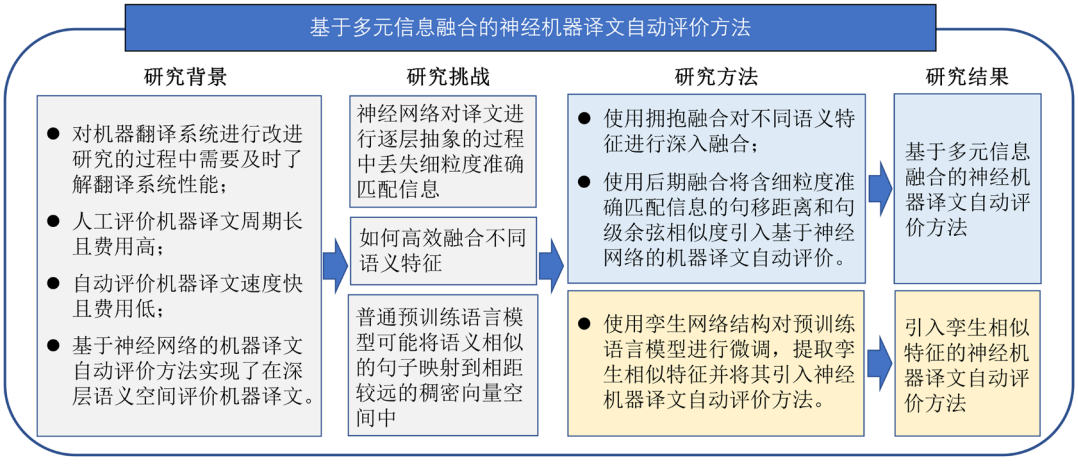

图 1-6 基于多元信息融合的神经机器译文自动评价方法的研究框架

本文的贡献主要有以下几个方面：

(1)本文对国内外基于神经网络的机器译文自动评价方法进行综述，将基于神经网络的机器译文自动评价方法分为基于表征匹配的方法和基于端到端神经网络的方法两大类，并将基于表征匹配的方法进一步分为基于静态表征匹配的方法和基于动态表征匹配的方法，将基于端到端神经网络的方法进一步分为基于基础端到端神经网络的方法、基于“预训练+微调”范式端到端神经网络的方法和

基于其他形式端到端神经网络的方法。此外，本文对自动评价方法的相关评测活动与评测指标进行了介绍。

(2)本文提出了基于多元信息融合的神经机器译文自动评价方法，该方法将中期融合和后期融合引入神经机器译文自动评价。其中，中期采用拥抱融合对神经网络提取的信息表征进行融合，使不同的信息表征能够深入融合；后期将句级余弦相似度和句移距离与神经网络自动评价模型评价结果进行融合，使模型能够依据细粒度准确匹配信息与深层语义信息进行评分。本文将所提方法与参与评测的其他方法进行比较，并进行消融实验和质量分析实验来对所提模型的性能进行验证。

(3)本文提出了引入孪生相似特征的神经机器译文自动评价方法，该方法使用孪生网络结构对预训练语言模型进行微调，使预训练语言模型能将语义相似的句子映射在距离相近的稠密向量空间中，从而更适用于基于机器译文和参考译文相似程度的机器译文自动评价任务，并使用该微调完成的孪生模型提取孪生相似特征，将孪生相似特征引入神经机器译文自动评价方法中，实现评价模型的性能提升。最后，本文将所提方法于其他参与评测的方法进行比较。

(4)本文对主要研究内容进行了总结，并从自动评价指标的易用性与鲁棒性、参考译文对自动评价方法研究的影响、篇章级别自动评价的研究与发展三个方面展望基于神经网络的机器译文自动评价的未来发展趋势。

1.4 本文的组织结构

本文的组织结构安排如下：

第一章，首先介绍了机器译文自动评价的研究背景与研究意义，其次介绍了国内外基于神经网络的机器译文自动评价的研究现状与机器译文自动评价的评测平台和评测方法，最后详细地说明了本文的研究内容以及本文作出的贡献。

第二章，介绍了本文所提方法的背景知识与相关理论基础，其中包括基于动态词向量的译文自动评价方法、引入源端信息的机器译文自动评价方法和融合XLM 词语表示的神经机器译文自动评价方法。

第三章，介绍了本文提出两个自动评价方法，分别为基于多元信息融合的神经机器译文自动评价方法和引入孪生相似特征的神经机器译文自动评价方法。对基于多元信息融合的神经机器译文自动评价方法中的中期融合和后期融合作详细阐述，对引入孪生相似特征的神经机器译文自动评价方法中的孪生相似特征的提取与引入方法作详细阐述。

第四章，介绍了模型的实验细节，如参数设置、采用的数据集、评价方式等

细节，然后给出实验结果并对实验结果进行分析。此外，通过消融实验对基于多元信息融合的神经机器译文自动评价方法实验有效性进行分析与证明，最后给出评价实例证明该方法的有效性。

第五章，总结本文的主要研究工作，并就当前领域内方法存在的不足展开讨论，最后展望机器译文自动评价领域未来趋势。

2 相关模型

2.1 基于动态词向量的机器译文自动评价方法

在机器译文自动评价领域中，传统的基于词形匹配的方法无法获知上下文信息，针对该问题，Mathur 等人[29]提出基于动态词向量的机器译文自动评价方法，将参考译文与机器译文映射在含语境信息的高维语义空间，使用注意力机制获取机器译文与参考译文的交互信息，并引入自然语言推理领域的启发式方法与增强序列推理模型 ESIM[39]分别构建基于 BERT 的双向长短时记忆网络与注意力机制的机器译文自动评价方法 $(\text{Bi-LSTM+attention})_{\text{BERT}}$ 与基于 BERT 的 ESIM 机器译文自动评价方法 $(\text{ESIM})_{\text{BERT}}$。实验结果显示，上述方法在 WMT'17 机器译文自动评价数据集上表现优良。

2.1.1 基于 BERT 的双向长短时记忆网络和注意力机制的机器译文自动评价方法

在基于 BERT 的双向长短时记忆网络和注意力机制的机器译文自动评价方法 $(\text{Bi-LSTM+attention})_{\text{BERT}}$ 中，Mathur 等人首先通过预训练语言模型 BERT 分别生成参考译文 r 与机器译文 t 的动态词向量，然后使用双向长短时记忆网络 Bi-LSTM 分别将参考译文 r 与机器译文 t 的动态词向量处理为语境词向量 h_r 和 h_t。为了获取参考译文语境词向量 h_r 与机器译文语境词向量 h_t 的交互信息，按式(2-1)方式通过点积运算构建机器译文与机器译文的相似度矩阵 A。$a_{i,j}$ 为相似度矩阵 A 中的元素。$i=1,2,\ldots,l_r$，l_r 为参考译文的长度。$j=1,2,\ldots,l_t$，l_t 为机器译文的长度。

$$a_{i,j}=h_{r_i}^{T}\cdot h_{t_j} \tag{2-1}$$

然后将相似度矩阵 A 与语境词向量 h_r、h_t 按式(2-2)、式(2-3)方式计算得到参考译文的相互表示 $\tilde{h}_r$、机器译文的相互表示 $\tilde{h}_t$。

$$\tilde{h}_r=\sum_{j=1}^{l_t}\frac{\exp(a_{i,j})}{\sum_i\exp(a_{i,j})}\cdot h_t \tag{2-2}$$

$$\tilde{h}_t=\sum_{i=1}^{l_r}\frac{\exp(a_{i,j})}{\sum_j\exp(a_{i,j})}\cdot h_r \tag{2-3}$$

为了缓解参考译文的相互表示 $\tilde{h}_r$ 与机器译文的相互表示 $\tilde{h}_t$ 的序列长度敏感问题，对相互表示 $\tilde{h}_r$ 、$\tilde{h}_t$ 分别进行最大池化和平均池化，如式(2-4)~式(2-7)所示。

$$\tilde{h}_{r,max} = \max_{k=1}^{l_r} \tilde{h}_{r,k} \tag{2-4}$$

$$\tilde{h}_{t,max} = \max_{k=1}^{l_t} \tilde{h}_{t,k} \tag{2-5}$$

$$\tilde{h}_{r,avg} = \sum_{k=1}^{l_r} \frac{\tilde{h}_{r,k}}{l_r} \tag{2-6}$$

$$\tilde{h}_{t,avg} = \sum_{k=1}^{l_t} \frac{\tilde{h}_{t,k}}{l_t} \tag{2-7}$$

其中，$\tilde{h}_{r,max}$ 为参考译文的最大池化后的相互表示，$\tilde{h}_{t,max}$ 为机器译文的最大池化后的相互表示，$\tilde{h}_{r,avg}$ 为参考译文的平均池化后的相互表示，$\tilde{h}_{t,avg}$ 为机器译文的平均池化后的相互表示。

Mathur 等人将池化结果按式(2-8)、式(2-9)方式分别进行拼接，然后使用启发式方法对拼接后的向量进行处理，如式(2-10)所示。

$$c_r = [\tilde{h}_{r,max}, \tilde{h}_{r,avg}] \tag{2-8}$$

$$c_t = [\tilde{h}_{t,max}, \tilde{h}_{t,avg}] \tag{2-9}$$

$$m_{att} = [c_t \oplus c_r \oplus (c_t \odot c_r) \oplus (c_t - c_r)] \tag{2-10}$$

其中，符号“⊕”表示向量之间的拼接操作，符号“⊙”表示两个向量逐元素相乘操作，符号“–”表示两个向量逐元素相减操作。

最后将向量 m_{att} 输入前馈神经网络预测得到译文质量分数，如式(2-11)所示。

$$y = \mathrm{ReLU}(m_{att} \cdot W + b) \cdot w + b' \tag{2-11}$$

其中 W、w、b、b' 为前馈神经网络中的权值，ReLU 为隐藏层激活函数，y 为预测得到的质量分数。

2.1.2 基于 BERT 的 ESIM 机器译文自动评价方法

在基于 BERT 的 ESIM 机器译文自动评价方法 $(\mathrm{ESIM})_{\mathrm{BERT}}$ 中，Mathur 等人将自然语言推理领域中的 ESIM 模型引入机器译文自动评价，使用 ESIM 模型生成参考译文与机器译文的增强表征 m_r 与 m_t ，如式(2-12)、式(2-13)所示。其中，h_r 和 $\tilde{h}_r$ 分别为参考译文的语境词向量和相互表示，h_t 和 $\tilde{h}_t$ 分别为机器译文的语境词向量和相互表示。

$$m_r = [h_r \oplus \tilde{h}_r \oplus (h_r \odot \tilde{h}_r) \oplus (h_r - \tilde{h}_r)] \tag{2-12}$$

$$m_t = [h_t \oplus \tilde{h}_t \oplus (h_t \odot \tilde{h}_t) \oplus (h_t - \tilde{h}_t)] \tag{2-13}$$

然后将增强表征 m_r 与 m_t 按照式(2-4)~式(2-7)方式生成机器译文最大池化增强向量 $v_{t,max}$、参考译文最大池化增强向量 $v_{r,max}$、机器译文平均池化增强向量 $v_{t,avg}$ 和参考译文平均池化增强向量 $v_{r,avg}$。将上述池化增强向量按式(2-14)方式进行拼接，拼接后的向量为 m_{ESIM}。最后将向量 m_{ESIM} 输入前馈神经网络中，预测得到译文质量分数 y，如式(2-15)所示。

$$m_{ESIM} = [v_{r,avg} \oplus v_{r,max} \oplus v_{t,avg} \oplus v_{t,max}] \tag{2-14}$$

$$y = \mathrm{ReLU}(m_{ESIM} \cdot W + b) \cdot w + b' \tag{2-15}$$

其中 W、w、b、b' 为前馈神经网络中的权值，ReLU 为隐藏层激活函数。

2.2 引入源端信息的机器译文自动评价方法

Mathur 等人的基于动态词向量的机器译文自动评价方法实现了在高维语义空间对比机器译文与人工参考译文的相似程度，使机器译文自动评价性能得到了一定提升，但由于语言中存在一词多义现象，该类方法仅仅将机器译文与单一的参考译文进行对比进而评价译文质量，导致机器译文自动评价方法在评价译文质量时所参考的信息可能不全面，在评价过程中忽略了源语言句子包含的信息。罗琪等人[40]针对该问题提出将源端信息引入机器译文自动评价，引入方法为使用训练完成的联合神经网络模型(Unified Neural Network for Quality Estimation, UNQE)[82][83]从机器译文和源语言句子中提取出质量向量，将含有源端信息的译文质量向量引入基于动态词向量的机器译文自动评价方法中。

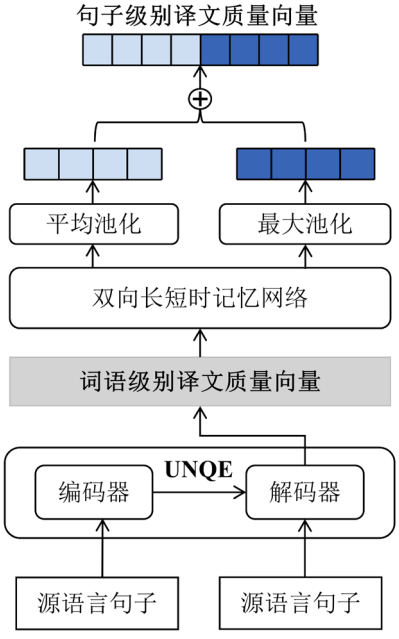

图 2-1 译文质量向量提取过程示意图

罗琪等人使用联合神经网络模型(UNQE)从源语言句子和参考译文中提取译文质量向量，其中的译文质量向量是指译文质量估计中可以描述译文质量的特征向量，译文质量估计是无需参考译文、仅以源语言句子作为参考依据对机器译文的质量进行评价的方法。提取译文质量向量仅使用到联合神经网络模型中的特征提取模块，该模块基于编码器-解码器(Encoder-Decoder)架构[84]。编码器端使用双向循环网络，作用为提取源语言句子的特征；解码器端采用 GRU 神经网络作为基础架构，将编码器端输出的特征矩阵与机器译文的独热编码表示输入解码器，得到词语级别的译文质量向量，再进一步处理为句子级别译文质量向量，译文质量向量提取过程如图 2-1 所示。

在编码器端(Encoder)，首先将源语言句子(*Source, s*)输入双向循环神经网络(Bi-RNN)提取源语言句子的特征 h_s，如式(2-16)~式(2-18)所示，其中 $\overrightarrow{\text{RNN}}$ 和 $\overleftarrow{\text{RNN}}$ 分别为前向、后向循环神经网络，$\vec{h}_s$ 和 $\overleftarrow{h}_s$ 分别为源语言句子对应的前向、后向循环神经网络的隐藏层状态。

$$\vec{h}_s = \overrightarrow{\text{RNN}}(Source) \tag{2-16}$$

$$\overleftarrow{h}_s = \overleftarrow{\text{RNN}}(Source) \tag{2-17}$$

$$h_s = [\vec{h}_s \oplus \overleftarrow{h}_s] \tag{2-18}$$

在解码器端(Decoder)，首先将机器译文(*Machine Translation, t*)使用独热编码表示为 $h_t = \{h_t^1, h_t^2, \ldots, h_t^{l_t}\}$，其中，$l_t$ 为机器译文的长度。将机器译文的独热编码表示 h_t 与源语言句子的特征 h_s 输入以 GRU 模型为主体的解码器得到词语级别的译文质量向量 $q_i, i = 1,2,\ldots,l_t$，如式(2-19)所示。

$$q_i = \text{Decoder}(h_s, h_t) \tag{2-19}$$

然后使用双向长短时记忆网络 Bi-LSTM 将词语级别质量向量 $q_i, i = 1,2,\ldots,l_t$ 处理为初始句子级别质量向量 $v_{qe}^{'}$，进一步通过平均池化和最大池化突出质量向量特征信息，得到池化后的初始句子级别质量向量 $v_{qe,max}^{'}$ 和 $v_{qe,avg}^{'}$，最后将池化后的结果拼接得到最终的句子级别机器译文质量向量 v_{qe}，如式(2-20)所示。

$$v_{qe} = [v_{qe,max}^{'} \oplus v_{qe,avg}^{'}] \tag{2-20}$$

为了将源语言句子中的语义信息引入机器译文自动评价，罗琪等人[40]将译文质量向量引入基于动态词向量的机器译文自动评价方法 $(\text{Bi-LSTM+attention})_{\text{BERT}}$ 方法和 $(\text{ESIM})_{\text{BERT}}$ 方法，构建 $(\text{Bi-LSTM+attention})_{\text{BERT+QE}}$ 方法和 $(\text{ESIM})_{\text{BERT+QE}}$ 方法。

在 $(\text{Bi-LSTM+attention})_{\text{BERT+QE}}$ 方法中，将译文质量向量 v_{qe} 与 $(\text{Bi-LSTM+attention})_{\text{BERT}}$ 方法中式(2-10)的向量 m_{att} 拼接，将拼接后的向量 m_{att+QE} 输入前馈神经网络预测译文质量得分，如式(2-21)、式(2-22)所示。

$$m_{att+QE} = [v_{qe} \oplus m_{att}] \tag{2-21}$$

$$y = \mathrm{ReLU}(m_{att+QE} \cdot W_1 + b_1) \cdot W_2 + b_2 \tag{2-22}$$

在 $(\mathrm{ESIM})_{\mathrm{BERT+QE}}$ 方法中，将译文质量向量 v_{qe} 与 $(\mathrm{ESIM})_{\mathrm{BERT}}$ 方法中式(2-14)的向量 m_{ESIM} 拼接，将拼接后的向量 $m_{ESIM+QE}$ 输入前馈神经网络预测译文质量得分，如式(2-23)、式(2-24)所示。

$$m_{ESIM+QE} = [v_{qe} \oplus m_{ESIM}] \tag{2-23}$$

$$y = \mathrm{ReLU}(m_{ESIM+QE} \cdot W_1 + b_1) \cdot W_2 + b_2 \tag{2-24}$$

其中，m_{att+QE} 为 $(\text{Bi-LSTM+attention})_{\mathrm{BERT+QE}}$ 方法中用于预测译文质量分数的向量，$m_{ESIM+QE}$ 为 $(\mathrm{ESIM})_{\mathrm{BERT+QE}}$ 方法中用于预测译文质量分数的向量，W_1、b_1、W_2、b_2 为前馈神经网络中的权重和偏置，y 为预测得到的质量分数。

2.3 融合 XLM 词语表示的神经机器译文自动评价方法

现有的基于神经网络的机器译文自动评价方法虽然能够使用深度神经网络捕获句子的深层语义信息，也能在一定程度上考虑到源端信息，但未能在同一语义空间中对比机器译文与源语言句子、参考译文的相似程度。针对上述不足，Hu 等人[41]提出将 XLM 词语表示引入神经机器译文自动评价方法。该方法使用跨语种预训练模型 XLM[35]获取源语言句子、机器译文与参考译文三者在同一语义空间中的表征，并使用分层注意力机制[42]和内部注意力机制[41]对其进行增强表示，得到机器译文与黄金参考(参考译文和源语言句子)的差异特征，将差异特征引入自动评价方法中能有效提升评价模型性能。

Hu 等人将源语言句子 s、参考译文 r、机器译文 t 组合拼接为三组句对，分别为源语言句子与机器译文的组合句对“s+t”、参考译文与源语言句子的组合句对“r+s”、参考译文与机器译文的组合句对“r+t”。将三组句对输入 XLM 模型，并以参考译文与源语言句子的组合句对“r+s”作为评判机器译文质量的黄金参考。

差异特征生成方式如图 2-2 所示。首先将字节对编码、位置编码与语言编码相加输入 XLM 模型，得到 XLM 模型的各层输出，如式(2-25)~式(2-27)所示。

$$H_{r+s} = \mathrm{XLM}(X_{r+s}) \tag{2-25}$$

$$H_{s+t} = \mathrm{XLM}(X_{s+t}) \tag{2-26}$$

$$H_{r+t} = \mathrm{XLM}(X_{r+t}) \tag{2-27}$$

X_{r+s}、X_{s+t}、X_{r+t} 分别为句子对“r+s”、“s+t”、“r+t”的 XLM 模型输

入；H_{r+s}、H_{s+t}、H_{r+t}分别为句子对“r+s”、“s+t”、“r+t”的 XLM 编码器各层输出。

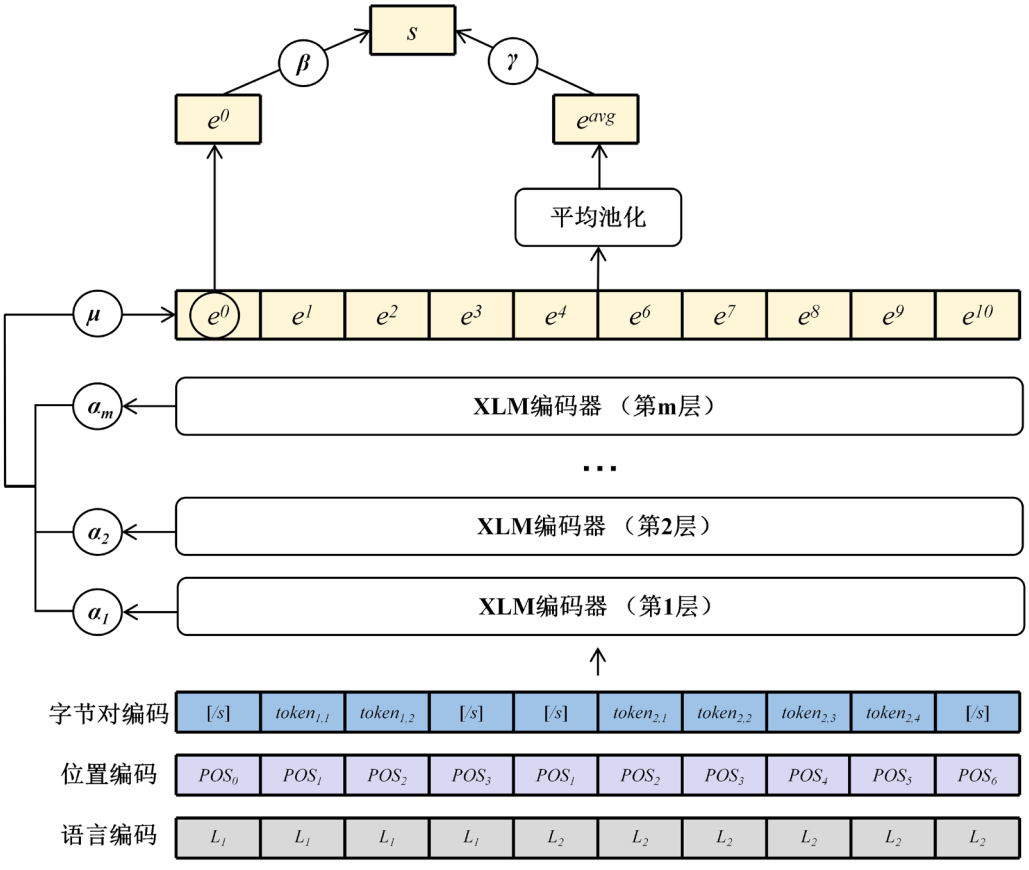

图 2-2 差异向量生成模块示意图

Zhang 等人[28]的研究显示仅用最后一层编码器的输出向量作为评判译文质量的依据容易导致自动评价模型性能的下降。因此，Hu 等人[41]使用分层注意力机制融合编码器各层向量的语言特征。具体方式为抽取出 XLM 模型的每一层隐藏层向量，然后将其进行加权求和，如式(2-28)、式(2-29)所示。

$$\alpha = \text{Softmax}([\alpha_1, \alpha_2, ..., \alpha_m]) \tag{2-28}$$

$$e^j = \mu(H_j^{\mathrm{T}} \cdot \alpha) \tag{2-29}$$

其中，$\alpha_1, \alpha_2, ..., \alpha_m$为 XLM 模型各层输出向量对应的权重，该权重通过学习得到。m 为 XLM 模型的层数。e^j 为 XLM 模型第 j 个位置的输出向量，由于该输出向量使用分层注意力机制综合了 XLM 模型各层的向量信息，故其蕴含 XLM 模型各层输出向量的语言学特征。

为了进一步突出特征，Hu 等人将 XLM 模型输出向量e_{r+s}、e_{s+t}、e_{r+t}通过平均池化层得到向量e_{r+s}^{avg}、e_{s+t}^{avg}、e_{r+t}^{avg}。然后使用内部注意力机制分别将向量e_{r+s}^{avg}、e_{s+t}^{avg}、e_{r+t}^{avg}与其对应的首个位置的 XLM 模型输出向量e_{r+s}^{0}、e_{s+t}^{0}、e_{r+t}^{0}进行加权求和得到句子对表征向量z_{r+s}、z_{s+t}、z_{r+t}，如式(2-30)~式(2-32)所示。

$$z_{r+s} = \beta_{r+s} e^{0}_{r+s} + \gamma_{r+s} e^{avg}_{r+s} \tag{2-30}$$

$$z_{s+t} = \beta_{s+t} e^{0}_{s+t} + \gamma_{s+t} e^{avg}_{s+t} \tag{2-31}$$

$$z_{r+t} = \beta_{r+t} e^{0}_{r+t} + \gamma_{r+t} e^{avg}_{r+t} \tag{2-32}$$

其中，β_{r+s}、γ_{r+s}、β_{s+t}、γ_{s+t}、β_{r+t}、γ_{r+t}为通过学习得到的内部注意力权重。为了获取句子对表征向量的差异信息，Hu 等人按式(2-33)方式将句子对表征向量拼接为差异向量e_{dv}。

$$e_{dv} = [z_{s+t} \oplus z_{r+s} \oplus z_{r+t} \oplus (z_{s+t} \odot z_{r+s}) \oplus (z_{s+t} - z_{r+s})] \tag{2-33}$$

最后，融合 XLM 词语表示的神经机器译文自动评价方法将差异向量e_{dv}分别与引入源端信息的机器译文自动评价方法$(\text{Bi-LSTM+attention})_{\text{BERT+QE}}$方法中的向量$m_{att+QE}$或$(\text{ESIM})_{\text{BERT+QE}}$方法中的向量$m_{ESIM+QE}$拼接，将拼接后的向量$m_{att+QE+DV}$或$m_{ESIM+QE+DV}$输入前馈神经网络预测译文质量分数。

$$m_{att+QE+DV} = [m_{att+QE} \oplus e_{dv}] \tag{2-34}$$

$$y = \text{ReLU}(m_{att+QE+DV} \cdot W_1 + b_1) \cdot W_2 + b_2 \tag{2-35}$$

或

$$m_{ESIM+QE+DV} = [m_{ESIM+QE} \oplus e_{dv}] \tag{2-36}$$

$$y = \text{ReLU}(m_{ESIM+QE+DV} \cdot W_1 + b_1) \cdot W_2 + b_2 \tag{2-37}$$

其中，W_1、b_1、W_2、b_2 为前馈神经网络中的权重矩阵，y 为前馈神经网络输出的机器译文的质量分数。

2.4 本章小结

本章介绍了基于多元信息融合的神经机器译文自动评价方法与引入孪生相似特征的神经机器译文自动评价方法涉及到的基本理论知识和相关模型。首先介绍了基于动态词向量的机器译文自动评价方法，相比传统基于词形的评价方法，该方法可以在语义层面对机器译文进行评价；然后介绍了在其基础上延伸的引入源端信息的机器译文自动评价方法，该方法引入质量估计中描述译文质量的译文质量向量，进一步提升评价性能；最后介绍了融合 XLM 词语表示的神经机器译文自动评价方法，该方法将源语言句子、参考译文和机器译文在同一语义空间内进行比较，表现出优良的性能。

3 方法

本文提出两种神经机器译文自动评价方法，分别为基于多元信息融合的神经机器译文自动评价方法和引入孪生相似特征的神经机器译文自动评价方法。

基于多元信息融合的神经机器译文自动评价方法从领域内以往研究的两处不足为出发点进行研究。第一处不足为当前基于神经网络的机器译文自动评价方法从预训练语言模型中提取译文的各类语言特征，然后将其直接拼接后输入神经网络中预测机器译文的质量分数，其中的直接拼接操作容易造成信息冗余，且各个特征未能实现深度融合。针对该处不足，本文提出使用拥抱融合对各个特征向量进行融合，使得各类特征向量能够深入融合，并且使得神经网络能够学习到各个特征之间的相关信息。第二处不足为当前基于神经网络的机器译文自动评价方法在使用神经网络对译文进行逐层抽象的过程中容易丢失细粒度准确匹配信息，针对该处不足，本文提出将含有细粒度准确匹配信息的句移距离和句级余弦相似度与神经网络模型评分进行后期融合，实现细粒度准确匹配信息的补充，从而提高评价模型性能。

引入孪生相似特征的神经机器译文自动评价方法在基于多元信息融合的神经机器译文自动评价方法的基础上展开进一步研究。该方法引入使用孪生微调后的预训练语言模型提取的孪生相似特征来提升评价模型的性能。由于当前的预训练语言模型可能会将语义相近的句子映射到相距较远的稠密向量空间中，使用孪生网络结构对预训练语言模型进行微调，使其能将语义相近的句子映射到更近的稠密向量空间中，通过该微调完成的孪生预训练模型提取孪生相似特征，并将该特征引入神经机器译文自动评价方法。

3.1 基于多元信息融合的神经机器译文自动评价方法

当前最新的机器译文自动评价方法从预训练语境词向量中提取不同的深层语义特征，将它们直接拼接后构建多层神经网络预测译文质量。然而，该类方法在对译文进行逐层抽象时容易丢失细粒度准确匹配信息，且采用直接拼接的方式融合不同语义特征不仅使得神经网络无法学习各类特征之间的相关信息，还易造成信息冗余、表征向量维度过高、浪费训练资源等问题。因此，如何高效融合不同信息是提升信息表征能力的关键。