文章编号：1003-0077(2022)03-0045-09

基于多粒度特征的文本生成评价方法

赖 华[1,2]，高玉梦[1,2]，黄于欣[1,2]，余正涛[1,2]，张勇丙[1,2]

（1. 昆明理工大学 信息工程与自动化学院，云南 昆明 650504；
2. 昆明理工大学 云南省人工智能重点实验室，云南 昆明 650504）

摘 要：近年来，基于预训练语言模型的文本生成评价方法得到了广泛关注，其通过计算两个句子间子词粒度的相似度来评价生成文本的质量。但是对于越南语、泰语等存在大量黏着语素的语言，单个音节或子词不能独立成词表达语义，仅基于子词粒度匹配的方法并不能够完整表征两个句子间的语义相似关系。基于此，该文提出一种基于子词、音节、词组等多粒度特征的文本生成评价方法。首先基于 MBERT 模型生成文本的表示，然后引入音节、词组等粗粒度语义单元之间的相似性来增强子词粒度的相似度评价模型。在机器翻译、跨语言摘要、跨语言数据筛选等任务上的实验结果表明，该文提出的多粒度特征评价方法相比 ROUGE、BLEU 等基于统计的评价方法以及 Bertscore 等基于语义相似度的评价方法都取得了更好的性能，与人工评价结果相关性更高。

关键词：文本生成；评价方法；黏着语素；多粒度特征；MBERT

中图分类号：TP391 **文献标识码**：A

Evaluation Method of Text Generation Based on Multi-granularity Features

LAI Hua[1,2], GAO Yumeng[1,2], HUANG Yuxin[1,2], YU Zhengtao[1,2], ZHANG Yongbing[1,2]

(1. Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan 650504, China;
2. Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming, Yunnan 650504, China)

Abstract: Recently, the evaluation method of text generation based on pre-trained language model has gained attention, which evaluates the quality of generated text by computing the granularity similarity of sub-words of two sentences. However, for languages that contain many adhesive morphemes, such as Vietnamese and Thai, a single syllable or sub-word cannot form the semantic integrity, which means that the sub-word granularity matching method cannot fully represent the semantic relationship between two sentences. Therefore, we propose a text generation evaluation method with multi-granularity features of sub-words, syllables, and phrases. After the representation of text is obtained by MBERT, the semantic similarity of syllables and phrases is introduced to enhance the evaluation model of sub-words. Experimental results on such tasks as cross-language summarization, machine translation, and data screening show that, compared with ROUGE, BLEU based on statistical evaluation and Bertscore based on deep semantic matching, the proposed metric correlates better with human judgments.

Keywords: text generation; evaluation method; adhesive morphemes; multi-granularity feature; MBERT

收稿日期：2021-08-25 **定稿日期**：2021-11-01

基金项目：国家自然科学基金(61732005, 61972186, 61762056, 61761026)；云南省重大科技专项计划项目(202002AD080001-5)；云南省重大科技专项计划项目(202103AA080015)；云南省高新技术产业专项(201606)；云南省基础研究计划项目(202001AT070047, 2018FB104)

0 引言

目前，机器翻译、文本摘要等生成任务取得了很好的性能，但生成文本的质量评价仍是一个难点问题[1-3]，后者可以看作一个计算生成句和标准参考句之间相似度的任务。根据匹配方法的不同，可将其分为基于统计的方法和基于语义相似度的方法两类。

基于统计的方法一般是通过计算生成句和参考句之间的不同粒度语义单元的共现程度来评价模型性能，根据计算方式不同可以分为基于编辑距离的方法和基于词重叠率的方法。基于编辑距离的方法是通过计算生成句改写到参考句所需的编辑操作次数来衡量两个句子之间的相似性。编辑操作次数越多则表示一个句子改写成另一个句子越困难，句子之间的相似性越低。Snover 等人[4]将生成句通过增加、删除、替换改写到参考句所需最少操作次数定义为编辑距离，使用参考句的长度对编辑距离归一化得到错误率。与基于编辑距离的方法不同，基于词重叠率的评价是当前的主流评价方法，通过计算生成句与标准参考句之间的字词重叠率来评价模型性能。Papineni 等人提出的 BLEU[5]，以及随后 Lin 等人提出的 ROUGE[6]，是机器翻译与文本摘要模型评估中最常用的评价指标。基于此，研究者相继提出了 METEOR[7-9] 以及 ROUGE 变体[10-11] 等改进的评价方法，其核心均是通过融入外部知识（如 WordNet）来解决同义词匹配的问题。综上所述，虽然基于统计的方法能够很好地表征句子之间的重叠度，但是缺乏对句子深层语义的建模，因此可能出现两个句子语义相近，但是重叠词很少的情况下评分较低的现象，如图 1(a)所示，两个句子均表示大雪覆盖的含义，语义较为相近，但是重叠的词较少，这在一定程度上限制了生成句的多样性。

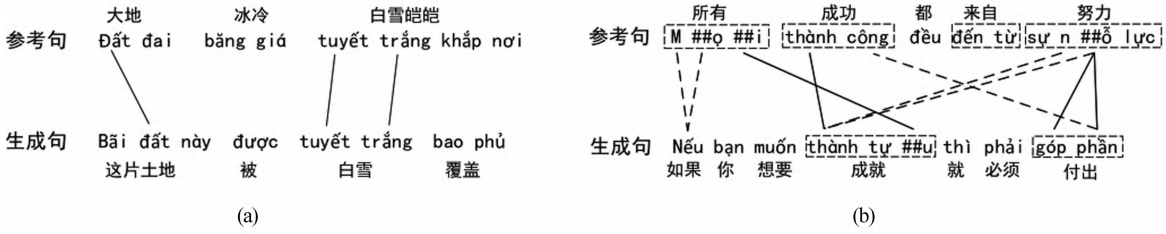


图 1 评价示意图

为了应对以上问题，近年来基于预训练语言模型（Pre-trained Language Models，PLMs）的语义相似度评价方法得到了广泛关注[12-13]，其基本思想是利用 PLMs 提取语义特征，基于此实现语义相似度匹配。如 Zhang 等人提出的 BERTScore[14] 方法，首先基于 BERT（Bidirectional Encoder Representations from Transformers，BERT）将参考句与生成句进行子词切分，然后得到子词语义表征向量，最后采用贪婪匹配方法计算参考句与生成句之间子词粒度表征向量的最大余弦距离，从而表征两个句子的语义相似度，取得了较好的效果。但针对越南语、泰语等存在大量黏着语素的语言，多数子词并不具有完整的语义，基于子词粒度的对应关系不能很好地反映语义之间的相似关系。如图 1(b)所示，本文给出了基于 MBERT（Multilingual BERT，MBERT）[15] 的子词切分及语义匹配结果，图中虚线连接表示子词语义匹配不合理的情况。在图 1(b)中，词组“sự nỗ lực”，“thành tựu”，“góp phần”中的子词均为黏着语素，即“sự nỗ lực”三个音节表示努力的意思，如果将“sự nỗ lực”拆开就类似将“努力”拆成“努”与“力”二字，无论单独使用哪一个都不能完整表达其语义。另外参考句中“sự nỗ lực”被切分为子词序列“sự，n，##ỗ，lực”，其中，“n”与生成句“thành”匹配，生成句中与子词“##ỗ”匹配的有三个子词（三个子词匹配相似度值相等且最大），分别为“thành”“góp”“phần”。若简单计算上述子词匹配组合之间的相似度，并无明显的实际意义。相对于子词，音节与词组是更粗粒度的语义单元，如图 1(b)中，将子词“n，##ỗ”合并为更完整表达语义的音节“nỗ”或词组“sự nỗ lực”，与生成句中的“góp phần（付出）”进行匹配，能更好地反映参考句与生成句之间的语义相关关系。由此，本文提出一种多粒度特征的文本生成评价方法（Multi-granularity feature score，Mgfscore）。在 MBERT 进行子词语义表征的基础上，对音节、词组等更完整表达语义的单位进行语义表征，采用多粒度特征匹配来评价生成句与参考句之间的相似性，进一步增强基于子词语义特征匹配的自动评价方法。

1　基于多粒度特征的文本生成评价方法

基于多粒度特征的文本生成评价方法主要包括两个部分：多粒度语义表征部分与多粒度特征匹配部分，如图 2 所示。

多粒度语义表征部分包括子词语义表征模块、音节语义表征模块以及词组语义表征模块。首先，为了获取不同粒度的语义特征表示，采用 MBERT 对参考句与生成句的子词序列进行向量表示。基于此，结合构成音节与词组的子词组合，得到音节语义表征向量以及词组语义表征向量。其次，分别对不同粒度下的参考句与生成句的语义特征向量进行最大余弦相似度匹配，使模型在多个粒度上考虑两个句子间的语义相关关系。

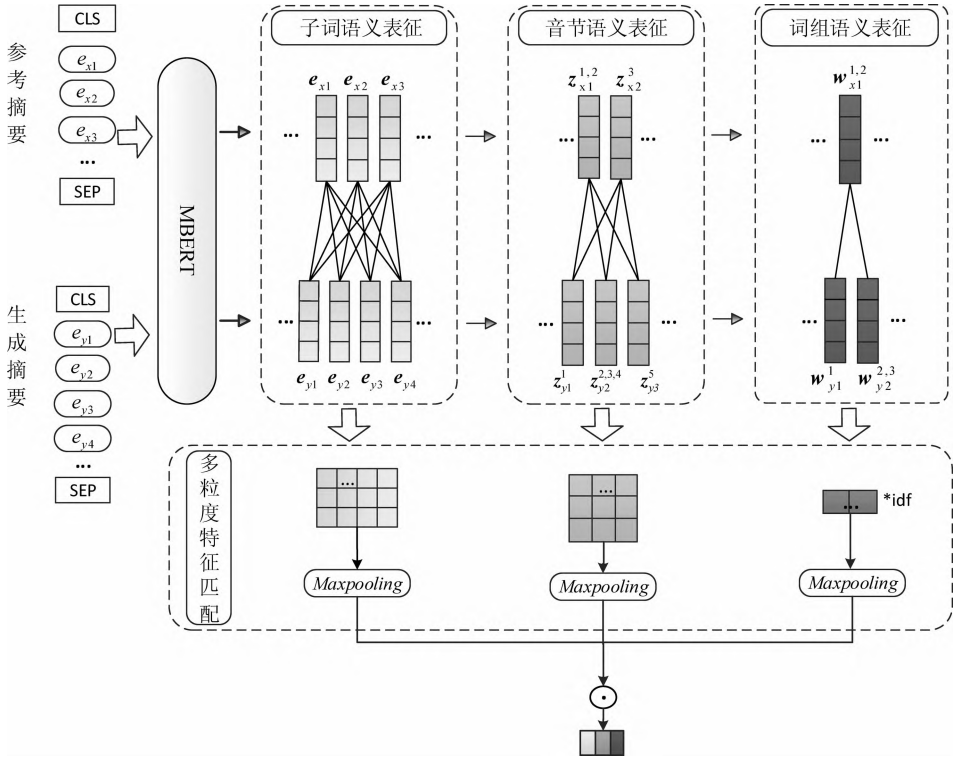


图 2　多粒度特征评价方法模型

1.1　多粒度语义表征部分

1.1.1　子词语义表征模块

给定一个参考句 X 和一个生成句 Y，首先使用 MBERT 对 X 与 Y 进行子词粒度切分得到参考句子词序列 $\{e_{x1},e_{x2},e_{x3},\cdots,e_{xn}\}$ 与生成句子词序列 $\{e_{y1},e_{y2},e_{y3},\cdots,e_{ym}\}$。如式(1)、式(2)将得到的子词序列输入 MBERT 上下文嵌入，得到参考句子词表征向量 $\boldsymbol{E}_X=\{\boldsymbol{e}_{x\mathrm{cls}},\boldsymbol{e}_{x1},\boldsymbol{e}_{x2},\boldsymbol{e}_{x3},\cdots,\boldsymbol{e}_{xn},\boldsymbol{e}_{x\mathrm{sep}}\}$ 与生成句子词表征向量 $\boldsymbol{E}_Y=\{\boldsymbol{e}_{y\mathrm{cls}},\boldsymbol{e}_{y1},\boldsymbol{e}_{y2},\boldsymbol{e}_{y3},\cdots,\boldsymbol{e}_{ym},\boldsymbol{e}_{y\mathrm{sep}}\}$，其中，$e_{xn}$ 表示参考句 X 中第 n 个子词。

$$\{\boldsymbol{e}_{x\mathrm{cls}},\boldsymbol{e}_{x1},\boldsymbol{e}_{x2},\boldsymbol{e}_{x3},\cdots,\boldsymbol{e}_{xn},\boldsymbol{e}_{x\mathrm{sep}}\}=\mathrm{MBERT}\{[\mathrm{cls}],e_{x1},e_{x2},e_{x3},\cdots,e_{xn},[\mathrm{sep}]\}\quad(1)$$

$$\{\boldsymbol{e}_{y\mathrm{cls}},\boldsymbol{e}_{y1},\boldsymbol{e}_{y2},\boldsymbol{e}_{y3},\cdots,\boldsymbol{e}_{ym},\boldsymbol{e}_{y\mathrm{sep}}\}=\mathrm{MBERT}\{[\mathrm{cls}],e_{y1},e_{y2},e_{y3},\cdots,e_{ym},[\mathrm{sep}]\}\quad(2)$$

其中，MBERT 表示 Multilingual BERT 词嵌入。

1.1.2　音节语义表征模块

音节是发音的最小语音单元，相比于子词它是一个更粗的粒度。在越南语与泰语中，音节由元音、辅音、声调构成。越南语书写时，音节与音节之间用空格隔开，因此，可利用空格对越南语进行音节切分。而泰语书写时，音节间无空格，因此，在泰语中需要借助 PyThaiNLP[16] 工具对泰语进行音节切分。针对中文，大多情况下，子词与音节相同，均由一个汉字构成。因此，本文在中文上不进行音节粒度语义表征。

为了得到音节语义表征，使用对应语言的音节切分工具或方法对原始输入序列 X 和 Y 进行切分，得到音节序列 $Z_X=\{z_{x1}^{1,2},z_{x2}^{3},\cdots,z_{xi}^{n-1,n}\}$ 和 $Z_Y=\{z_{y1}^{1},z_{y2}^{2,3,4},\cdots,z_{yj}^{m}\}$，其中，$z_{xi}^{n-1,n}$ 表示参考句 X 中

的第 i 音节由第 $n-1$ 与第 n 个子词构成。基于子词语义表征模块得到的子词序列表征向量，如式(3)、式(4)所示，分别得到参考句与生成句的音节序列语义向量表征 $\boldsymbol{Z}_X$ 与 $\boldsymbol{Z}_Y$。

$$\{\boldsymbol{z}_{x1}^{1,2},\boldsymbol{z}_{x2}^{3},\cdots,\boldsymbol{z}_{xi}^{n-1,n}\}=\{\text{pooling}(\boldsymbol{e}_{x1},\boldsymbol{e}_{x2}),\boldsymbol{e}_{x3},\cdots,\text{pooling}(\boldsymbol{e}_{x(n-1)},\boldsymbol{e}_{xn})\} \tag{3}$$

$$\{\boldsymbol{z}_{y1}^{1},\boldsymbol{z}_{y2}^{2,3,4},\cdots,\boldsymbol{z}_{yj}^{m}\}=\{\boldsymbol{e}_{y1},\text{pooling}(\boldsymbol{e}_{y2},\boldsymbol{e}_{y3},\boldsymbol{e}_{y4}),\cdots,\boldsymbol{e}_{ym}\} \tag{4}$$

其中，“pooling”表示 average pooling。

1.1.3 词组语义表征模块

词组是比音节更粗的粒度，它是语言中能够独立运用的最小语义单位。词组是由单个或多个语素构成的，语素又可分为自由语素与黏着语素，不能够独立成词的语素称为黏着语素。在越南语、泰语等语言中存在大量的黏着语素，音节可构成语素，进而构成词。由此，可以简单理解为词组由音节构成。在各语言中，需采用不同的工具对其进行分词，中文使用结巴分词，越南语使用 VnCoreNLP[17] 分词，泰语使用 PyThaiNLP[16] 分词。

在不同语种下，对原始输入序列 X 和 Y 使用工具分词得到词组序列 $W_X=\{w_{x1}^{1,2},w_{x2}^{3},\cdots,w_{xl}^{i}\}$ 与 $W_Y=\{w_{x1}^{1},w_{x2}^{2,3},\cdots,w_{yp}^{j-2,j-1,j}\}$，其中，$w_{yp}^{j-2,j-1,j}$ 表示生成句 Y 中第 p 个词组由第 $j-2$，$j-1$，j 个音节构成。基于音节语义表征向量，如式(5)、式(6)所示，对参考句词组序列 W_X 和生成句词组序列 W_Y 进行词组语义向量表征。

$$\{\boldsymbol{w}_{x1}^{1,2},\boldsymbol{w}_{x2}^{3},\cdots,\boldsymbol{w}_{xl}^{i}\}=\{\text{pooling}(\boldsymbol{z}_{x1}^{1,2},\boldsymbol{z}_{x2}^{3}),\boldsymbol{z}_{x3}^{4,5},\cdots,\boldsymbol{z}_{xi}^{n-1,n}\} \tag{5}$$

$$\{\boldsymbol{w}_{y1}^{1},\boldsymbol{w}_{y2}^{2,3},\cdots,\boldsymbol{w}_{yp}^{j-2,j-1,j}\}=\{\boldsymbol{z}_{y1}^{1},\text{pooling}(\boldsymbol{z}_{y2}^{2,3,4},\boldsymbol{z}_{y3}^{5}),\cdots,\text{pooling}(\boldsymbol{z}_{y(j-2)}^{m-3},\boldsymbol{z}_{y(j-1)}^{m-2,m-1},\boldsymbol{z}_{yj}^{m})\} \tag{6}$$

1.2 多粒度特征匹配部分

利用多粒度语义表征部分对参考句与生成句的各粒度序列进行语义特征向量后，在不同粒度下，分别计算召回率与精确率与 F 值。最后，结合多粒度特征确定两个句子间的语义相关关系。

基于式(3)与式(4)得到的参考句音节语义向量表征序列 $\boldsymbol{Z}_X$ 和生成句音节序列语义向量表征 $\boldsymbol{Z}_Y$ 之后，按式(7)分别计算 $\boldsymbol{Z}_X$ 中与 $\boldsymbol{Z}_Y$ 中的每个音节语义向量表征的最大余弦相似度匹配结果，按式(7)计算音节粒度下的召回率 R_Z，按式(8)计算 $\{\boldsymbol{z}_{y1}^{1},\boldsymbol{z}_{y2}^{2,3,4},\cdots,\boldsymbol{z}_{yj}^{m}\}$ 分别与 $\boldsymbol{Z}_X$ 中每一个音节语义向量表征 $\boldsymbol{z}_x$ 匹配计算音节粒度下的精确率 P_Z，按式(9)计算 F_Z 值。

$$R_Z=\frac{1}{|z_x|}\sum_{\boldsymbol{z}_{xi}\in\boldsymbol{z}_x}\max_{\boldsymbol{z}_{yj}\in\boldsymbol{z}_y}\frac{\boldsymbol{z}_{xi}^{\mathrm{T}}\boldsymbol{z}_{yj}}{\|\boldsymbol{z}_{xi}\|\,\|\boldsymbol{z}_{yj}\|} \tag{7}$$

$$P_Z=\frac{1}{|z_y|}\sum_{\boldsymbol{z}_{yj}\in\boldsymbol{z}_y}\max_{\boldsymbol{z}_{xi}\in\boldsymbol{z}_x}\frac{\boldsymbol{z}_{xi}^{\mathrm{T}}\boldsymbol{z}_{yj}}{\|\boldsymbol{z}_{xi}\|\,\|\boldsymbol{z}_{yj}\|} \tag{8}$$

$$F_Z=2\frac{P_ZR_Z}{P_Z+R_Z} \tag{9}$$

其中，$|z_x|$ 为 z_x 序列的长度，$\frac{\boldsymbol{z}_{xi}^{\mathrm{T}}\boldsymbol{z}_{yj}}{\|\boldsymbol{z}_{xi}\|\,\|\boldsymbol{z}_{yj}\|}$ 表示为 $\boldsymbol{z}_{xi}$ 与 $\boldsymbol{z}_{yj}$ 余弦相似度计算。

参考式(7)～式(9)可得到在子词粒度下的 R_E，P_E，F_E，以及词组粒度下的 R_W，P_W，F_W。

idf 常用来衡量字或某个词在一个文件集内的重要程度，如式(10)所示。如果文件集中包含词组 w 的文档越少，idf 越大，说明该词在突出文章主题上发挥了关键作用。词组是能够独立运用的最小语义单位，因此，本文认为基于词组计算 idf 能够更大化地凸显句子中的重要内容，本文可选择 idf 对词组粒度进行加权，如式(11)所示。

$$\text{idf}(w)=\log\frac{N}{n_w} \tag{10}$$

$$R(\text{idf}_w)=\frac{\sum_{w_{xl}\in w_x}\text{idf}(w_{xl})\max_{w_{yp}\in w_y}\frac{\boldsymbol{w}_{xl}^{\mathrm{T}}\boldsymbol{w}_{yp}}{\|\boldsymbol{w}_{xl}\|\,\|\boldsymbol{w}_{yp}\|}}{\sum_{w_{xl}\in w_x}\text{idf}(w_{xl})} \tag{11}$$

其中，N 为总文档数，n_w 为包括词组的文档数。

最后，本文在多个粒度特征上进行观察，按式(12)计算生成句的精确率 P，并计算多粒度特征下的召回率 R 以及 F 值。

$$P=\frac{P_E+P_Z+P_W}{3} \tag{12}$$

2 实验

2.1 实验设置

为了验证本文提出的多粒度特征评价指标的效果，本文在机器翻译、跨语言摘要与跨语言数据集筛选三个任务上进行测评实验。另外，为了比较不同自动评价指标的性能，本文采用与前人相同的研究方法[1,13,18]，通过皮尔逊相关系数 r、斯皮尔曼等级系数 ρ 以及肯德尔等级系数 τ 来判定不同的自动评价指标与人工评价的相关性，r、ρ、τ 取值越接近 1，表示相关性越好。其中，人工评价由 12 名精通中文、越南

语、泰语的志愿者对文本生成测评数据进行人工打分得到，打分范围为 1 到 5 分(1 最差，5 最好)。

本文中文本生成测评实验数据由两部分构成，一部分是本文在网络上爬取并整理的越南语、泰语句子，将该数据集称为 VTSTS(Vietnamese and Thai Dataset)；另一部分是公开的中文短文本摘要数据集 LCSTS[19]。构建测评数据具体过程如下：

(1) 针对机器翻译任务(MT)，本文从 LCSTS 和 VTSTS 中分别抽取中文、越南语、泰语各 200 个句子作为目标语言参考句。利用谷歌翻译模型进行回译(将 A 语言的正向翻译成 B，再将 B 反向翻译成 A)，反向翻译得到目标语言生成句，构建越中、中越、中泰机器翻译任务测评所需数据。

(2) 针对跨语言摘要任务，在 LCSTS 中抽取 200 个中文摘要对，用传统的基线方法先摘要后翻译(**ST**)和先翻译后摘要(**TS**)构建中-越摘要对。翻译时采用谷歌翻译模型；摘要时，考虑到中-越跨语言摘要缺乏大规模监督数据，采用 LexRank[20] 这种无监督的抽取式摘要模型，构建中越跨语言摘要任务测评所需数据。

(3) 针对跨语言数据集筛选任务(STS)，从 VTSTS 数据集中抽取 200 个越南语单句，利用谷歌翻译模型，回译构建越-中-越伪平行句。

2.2　基线方法

本文选取以下四种评价方法作为基线方法。在同一文本生成任务中，所有的基线方法测评对象均相同。

BLEU：基于统计的评价方法，在机器翻译评价任务中使用得最为广泛，将生成句与参考句之间的长短关系作为惩罚因子，基于此，结合参考句与生成句间的 N-gram 重叠率进行评分，在本文计算中，N 最大为 4。

ROUGE：基于统计的评价方法，在文本摘要评价任务中使用得最为广泛，通过对比参考句与生成句间的 N-gram 重叠率进行评分，本文分别计算 ROUGE-1、ROUGE-2、ROUGE-L 的 $F1$ 值。

Bertscore：基于深度语义匹配的评价方法，用于文本生成任务的评价，通过计算参考句与生成句的子词粒度语义相似度进行评分，本文计算在该评价方法下的三个指标为 R 召回率、P 精确率、F 值。

Mgfscore(Ours)：本文提出的评价方法，可用于文本生成任务的评价，综合子词、音节、词组语义特征信息对生成句进行评分，本文计算在该评价方法下的三个指标为召回率 R、精确率 P 及 F 值。

3　实验结果

3.1　机器翻译

为了探究本文提出的评价方法在评估不同语言中的有效性，本文在越中、中越、中泰机器翻译任务上进行实验。不同评价指标与人工评估的相关系数 r、ρ、τ 结果如表 1 所示。其中，MT(vi→zh)表示越中机器翻译测评任务，MT(zh→vi)表示中越机器翻译测评任务，MT(zh→th)表示中泰机器翻译测评任务，F_{idf} 表示进行 idf 加权后 F 值得分。

表 1　机器翻译测评任务中各指标下的相关系数

		MT(vi→zh)			MT(zh→vi)			MT(zh→th)		
		r	ρ	τ	r	ρ	τ	r	ρ	τ
BLEU		0.400	0.455	0.318	0.326	0.322	0.228	0.374	0.377	0.259
Bertscore	R	0.425	0.457	0.312	0.441	**0.456**	0.334	0.424	0.436	0.304
	P	**0.471**	**0.522**	**0.362**	0.379	0.390	0.281	0.394	0.406	0.284
	F	0.457	0.500	0.344	0.419	0.435	0.317	0.426	0.433	0.303
	F_{idf}	0.442	0.500	0.341	—	—	—	—	—	—
Mgfscore (**Ours**)	R	0.424	0.469	0.318	**0.451**	**0.456**	**0.336**	**0.461**	**0.464**	**0.323**
	P	0.461	0.514	0.355	0.386	0.393	0.281	0.415	0.426	0.301
	F	0.451	0.493	0.337	0.429	0.439	0.320	0.452	0.455	0.319
	F_{idf}	0.461	0.500	0.343	0.431	0.441	0.322	0.438	0.453	0.319

从表1可以看出，在越中测评任务上，无idf对重要词组进行加权时，本文提出的方法Mgfscore与人工评价相关性低于Bertscore。这是由于中文上子词属于较粗粒度(中文子词与音节基本相同)，子词嵌入表征向量存在误差，导致了多粒度语义特征提取时出现错误传播现象较为严重。

另外，在中越及中泰两个语言对的机器翻译测评任务上，Mgfscore与人工评价的相关性明显高于基线BLEU和Bertscore，证明了本文提出的多粒度特征评价方法的有效性。其中，Mgfscore在中泰测评任务中提升最大，与Bertscore最优结果相比，在相关性系数r、ρ、τ分别得到了3.5%，2.8%，1.9%的提升。

针对idf加权来说，在越中测评任务中使用idf加权时，Mgfscore-F_{idf}相关性高于Bertscore-F_{idf}、Mgfscore-F，证明了本文提出在词组粒度进行idf加权的策略的优越性。相比越中任务，在中越及中泰两个语言对任务上，Mgfscore-F_{idf}与Mgfscore-F相比，性能提升并不明显，甚至出现性能下降，是由于越南语与泰语的分词工具更倾向于将句子切分为音节，且部分词组分词不准确导致的。

3.2 跨语言摘要

为了验证本文提出的评价方法在不同文本生成测评任务上的有效性，本实验在中越跨语言摘要任务上进行验证，对比不同评价指标的好坏。表2展示了TS以及ST作为跨语言摘要模型时，在中越跨语言摘要测评任务上不同评价指标与人工评估的相关系数r、ρ、τ。

从表2可以看出，在中越跨语言摘要测评任务中，相比于ROUGE与Bertscore，本文评价方法Mgfscore取得了最优结果。其中，Bertscore与Mgfscore均是基于MBERT表征进行评价，二者与人工评价的相关性明显高于ROUGE，说明了基于语义相似度评价的方法比基于统计的评价方法更具有优势。在TS模型中，Mgfscore与Bertscore最优结果相比，在相关系数r、ρ、τ上分别提升了1.6%，2.0%，1.9%；类似地，在ST模型中，Mgfscore与Bertscore最优结果相比，在相关系数r、ρ、τ上分别得到1.3%，0.8%，0.2%的提升，证明了本文提出的多粒度特征评价方法与人工评价相关性更高。

表2 跨语言摘要测评任务中各指标下的相关系数

		TS (zh→vi)			ST(zh→vi)		
		r	ρ	τ	r	ρ	τ
ROUGE	1	0.702	0.723	0.549	0.613	0.638	0.471
	2	0.611	0.656	0.509	0.517	0.560	0.420
	L	0.679	0.712	0.538	0.599	0.649	0.480
Bertscore	R	0.604	0.610	0.447	0.519	0.538	0.391
	P	0.716	0.737	0.564	0.661	0.693	0.521
	F	0.705	0.721	0.548	0.630	0.662	0.493
Mgfscore (**Ours**)	R	0.615	0.614	0.451	0.516	0.531	0.387
	P	**0.732**	**0.757**	**0.583**	**0.674**	**0.701**	**0.523**
	F	0.719	0.738	0.561	0.635	0.663	0.493
	F_{idf}	0.713	0730	0.553	0.633	0.656	0.490

3.3 跨语言数据集筛选

表3展示了跨语言数据集筛选测评任务中，不同评价指标与人工评估的相关系数r、ρ、τ。

表3 跨语言数据集筛选中各指标下的相关系数

		STS(vi↔vi)		
		r	ρ	τ
ROUGE	1	0.292	0.269	0.189
	2	0.276	0.270	0.190
	L	0.322	0.308	0.216
Bertscore	R	0.357	0.341	0.241
	P	0.312	0.304	0.213
	F	0.342	0.330	0.233
Mgfscore (**Ours**)	R	**0.370**	**0.347**	**0.246**
	P	0.317	0.302	0.211
	F	0.353	0.338	0.237
	F_{idf}	0.353	0.340	0.238

从表3可以看出，在中越跨语言数据集构建问题上，本文提出的评价方法Mgfscore仍是与人工评价相关性最高的方法，使用本文的评价方法能够更准确地筛选出高质量数据。

表2与表3均以中-越人工打分作为基准，与跨语言摘要任务相比，STS任务上相关系数r、ρ、τ均出现大幅降低的现象。对此，将两个任务的测评样本进行对比分析，得出与Chaganty等人[21]类似的结论：

当评价对象是低质量生成文本(人工评价中得分低的文本)时，使用词汇级别的评价方法测评与人工评价的相关系数更高。但在本文三个不同的文本生成测评实验中，可以看出，无论是在测评机器翻译生成的高质量文本，或是测评跨语言摘要生成的低质量文本，乃至测评跨语言数据集筛选任务中的错误传播严重的文本中，本文提出的多粒度特征评价方法 Mgfscore 均是与人工评价相关性最高的方法。

4　性能分析

为了探究本文提出的基于多粒度特征文本生成评价方法的性能，本节首先进行消融实验，对比了单一粒度特征与结合多粒度特征下的评价方法与人工评价的相关性。其次，本节探究了不同预训练语言模型及层数对多粒度特征评价方法的性能影响。最后，本节对 Mgfscore 评价方法进行实例分析，探讨本文提出的多粒度评价方法的优缺点。

4.1　消融实验

以中越与中泰两个语言对的测评任务为例探究在不同粒度特征匹配下 F 值与人工评价的相关系数 r、ρ、τ，实验结果如图 3 所示。

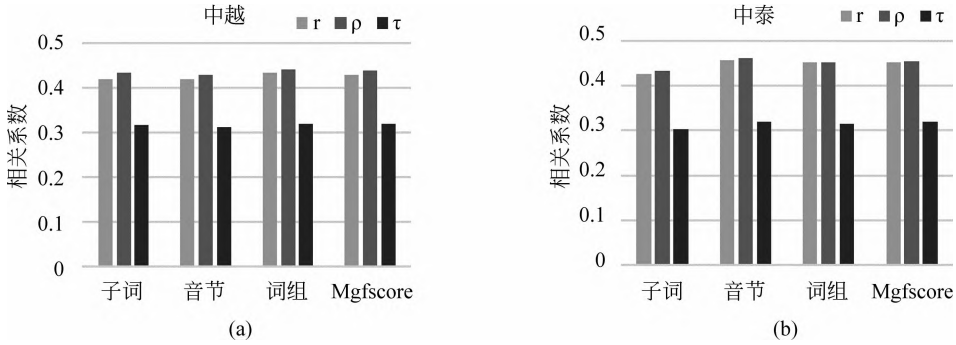


图 3　不同粒度下的实验结果

在图 3(a)与图 3(b)中可以发现，针对 r、ρ 相关系数来说，在中越测评任务上，基于词组粒度的评价与人工评价相关系数最高；在中泰测评任务上，基于音节粒度的评价与人工评价相关系数最高；但无论是在中泰还是中越测评任务中，本文提出的评价方法 Mgfscore 与人工评估的相关系数 r、ρ 均维持在高于中位数水平，由此可见，结合多粒度特征的评价方法具有较高的稳定性。针对相关系数 τ 来说，Mgfscore 与单粒度评价中的最优结果相近。由此，进一步证明了结合多粒度的评价方法具有更加稳定的性能。

4.2　各预训练语言模型及层数对性能的影响

为了验证本文提出方法在不同预训练语言模型上的有效性，以及探究预训练语言模型层数对本文提出的评价方法的性能的影响，本节在 MBERT 与 XLM(xlm-mlm-100-1280)[22] 两个预训练语言模型上进行实验，计算不同层表示下 Bertscore/Mgfscore 的 F 值与人工评价的皮尔逊相关系数 ρ，结果如图 4 所示。

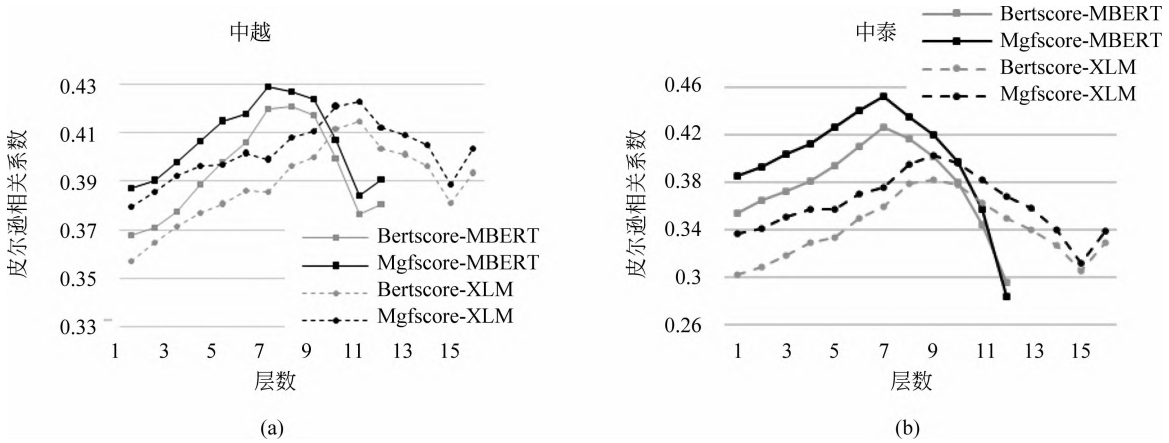
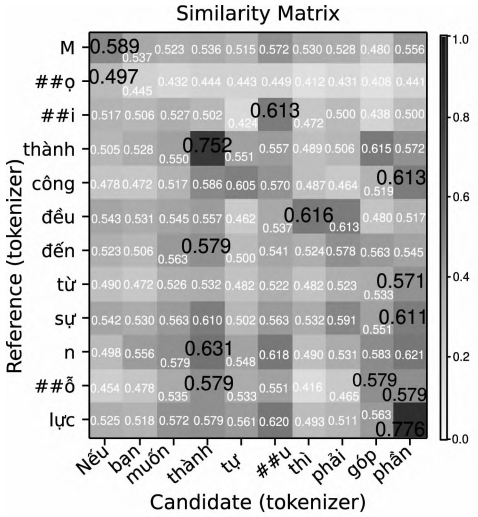


图 4　各预训练语言模型在不同层数下的实验结果

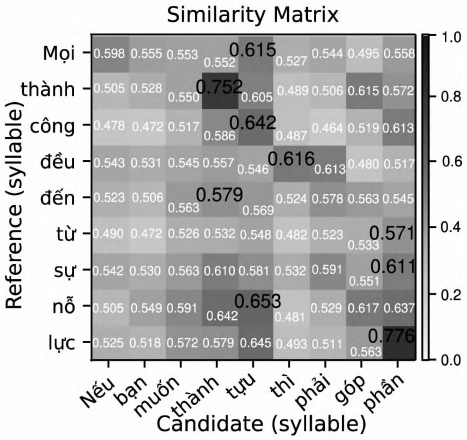
从图 4 可以看出，一方面，在不同的预训练语言模型中，本文提出的方法 Mgfscore 与 Bertscore 相比表现更好。另一方面，在中越与中泰的测评任务中，Mgfscore 或 Bertscore 与人工评估的相关性并没有随着模型层数的增加而不断上升，而是在中间层数取得了更好的结果，说明中间层数的表示包含了更多的语义信息。同时，Mgfscore 与 Bertscore 比较，在低层数表示时性能提升较大，这是由于低层数的表征中包含了更多子词本身的信息，此时更加明显地体现出了基于多粒度的方法能提取更全面的信息。而随着层数的增加，子词表征向量提取到了更多上下文的信息以及一些与语义无关的信息，此时带来的错误传播也更为明显，以致于 Mgfscore 较 Bertscore 的性能提升出现了减缓的趋势。

5 实例分析

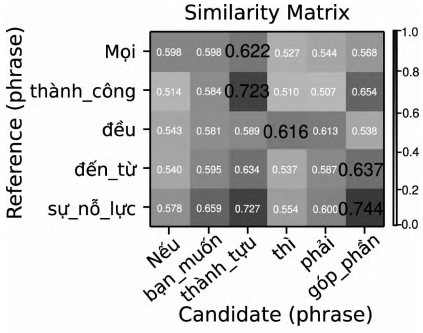
为了进一步证明多粒度特征的文本生成评价方法的有效性，本节对图 1(b)中示例，使用 Mgfscore 评价并进行分析，各粒度匹配结果如图 5 所示。



(a) 子词粒度



(b) 音节粒度



(c) 词组粒度

图 5 Mgfscore 各粒度匹配结果

对比图 5(a)，图 5(b)，图 5(c)，参考句中“Mọi(所有)”是一个音节，在子词匹配中切分为“M”“##ọ”“##i”三个子词，依次与“Nếu(如果)”“Nếu”“##u(tựu 子词)”匹配；在音节匹配中，“Mọi”的子词“M”“##ọ”“##i”合并，与生成句中“tựu(thành tựu 成就)”匹配；在词组粒度上，“Mọi”与生成句中“thành tựu(成就)”匹配；子词匹配中，仅使用 MBERT 表征时学习到上下文信息的子词语义向量进行匹配，不能判定它绝对的对或错。但存在大量黏着语素的语言中，仅使用不能独立表达语义的子词进行匹配并不合理。可以看出参考句中“Mọi(所有)”从子词粒度到词组粒度的匹配，是一个语义匹配更加合理的过程，相似度值也有所提升，说明了使用音节粒度与词组粒度匹配的有效性。

在子词匹配上，参考句中词组“thành công(成功)”拆开后无法完整表达成功之意，“thành”与生

成句“thành(thành tựu 成就)”匹配，相似度值为0.752，其中“công”与“phần(góp phần 付出)”匹配；在音节匹配上，“công”与“tựu(thành tựu 成就)”匹配，相比于子词匹配更加合理；在词组匹配中，参考句中“thành công(成功)”与“thành tựu(成就)”匹配，相似度值为0.723。从理论上来说，相比于子词或音节粒度匹配，词组粒度匹配更加合理，但是在词组粒度上相似度值出现了下降的情况。由此可见，当子词匹配出现不正确的情况时，使用更粗粒度的匹配很可能会出现错误传播。因此，仅基于单一粒度观察并不全面，而在多粒度上观察既保留了子词匹配中更精确的相似度值，也保留了粗粒度匹配的语义合理性，提取了更加全面的特征信息。

6　结束语

本文针对越南语、泰语等存在大量黏着语素的语言，提出了一种基于多粒度语义相似度的评价指标，通过 MBERT 提取不同粒度的语义特征信息，基于此综合考虑不同粒度语义单元的相似度，改善了传统基于统计的评价方法中存在的词汇多样性评价能力弱的问题，也更好反映了生成文本与参考文本间的语义关系。在中泰及中越语言对测评任务中，相比于 BLEU、ROUGE 以及 Bertscore 等评价指标，本文提出的多粒度特征评价方法均与人工评价结果更接近。

参考文献

[1] Ma Q, Wei J, Bojar O, et al. Results of the WMT19 metrics shared task: segment-level and strong MT systems pose big challenges[C]//Proceedings of the Fourth Conference on Machine Translation, 2019: 62-90.

[2] Rei R, Stewart C, Farinha A C, et al. COMET: a neural framework for MT evaluation[J]. arXiv preprint arXiv: 2009.09025, 2020.

[3] Takahashi K, Sudoh K, Nakamura S. Automatic machine translation evaluation using source language inputs and cross-lingual language model[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 3553-3558.

[4] Snover M, Dorr B, Schwartz R, et al. A study of translation edit rate with targeted human annotation[C]//Proceedings of the AMTA, 2006, 25(7): 223-231.

[5] 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. 2002: 311-318.

[6] Lin C Y. Rouge: a package for automatic evaluation of summaries[C]//Proceedings of the Text Summarization Branches Out, 2004: 74-81.

[7] Banerjee S, Lavie A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments[C]//Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 2005: 65-72.

[8] Denkowski M, Lavie A. Meteor universal: language specific translation evaluation for any target language [C]//Proceedings of the 9th Workshop on Statistical Machine Translation, 2014: 376-380.

[9] Guo Y, Hu J. Meteor++ 2.0: adopt syntactic level paraphrase knowledge into machine translation evaluation[C]//Proceedings of the 4th Conference on Machine Translation, 2019: 501-506.

[10] Ng J P, Abrecht V. Better summarization evaluation with word embeddings for ROUGE[J]. arXiv preprint arXiv: 1508.06034, 2015.

[11] Ganesan K. Rouge 2.0: updated and improved measures for evaluation of summarization tasks[J]. arXiv preprint arXiv: 1803.01937, 2018.

[12] Thompson B, Post M. Automatic machine translation evaluation in many languages via zero-shot paraphrasing[J]. arXiv preprint arXiv: 2004.14564, 2020.

[13] Gao Y, Zhao W, Eger S. SUPERT: towards new frontiers in unsupervised evaluation metrics for multi-document summarization[J]. arXiv preprint arXiv: 2005.03724, 2020.

[14] Zhang T, Kishore V, Wu F, et al. BERTScore: evaluating text generation with BERT[J]. arXiv preprint arXiv: 1904.09675, 2019.

[15] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv: 1810.04805, 2018.

[16] Phatthiyaphaibun W, Korakot Chaovavanich C. PyThaiNLP: Thai natural language processing in Python[OL]. http://doi. org/10.5281/zenodo, 2016, 3519354.

[17] Vu T, Nguyen D Q, Nguyen D Q, et al. VnCoreNLP: a Vietnamese natural language processing toolkit [J]. arXiv preprint arXiv: 1801.01331, 2018.

[18] Kryściński W, Keskar N S, McCann B, et al. Neural text summarization: a critical evaluation[J]. arXiv preprint arXiv: 1908.08960, 2019.

（下转第 63 页）

贾猛(1995—)，硕士研究生，主要研究领域为自然语言处理。
E-mail：1499115378@qq.com

张桂平(1960—)，博士，教授，主要研究领域为自然语言处理与机器翻译、知识工程与知识管理。
E-mail：zgp@ge-soft.com

王裴岩(1983—)，通信作者，博士，工程师，主要研究领域为自然语言处理与机器学习、知识工程与知识管理。
E-mail：wangpy@sau.edu.cn

(上接第 53 页)

[19] Hu B, Chen Q, Zhu F. LCSTS: a large scale Chinese short text summarization dataset[J]. arXiv preprint arXiv: 1506.05865, 2015.

[20] Erkan G, Radev D R. LexRank: graph-based lexical centrality as salience in text summarization[J]. Journal of Artificial Intelligence Research, 2004, 22: 457-479.

[21] Chaganty A T, Mussman S, Liang P. The price of debiasing automatic metrics in natural language evaluation[J]. arXiv preprint arXiv: 1807.02202, 2018.

[22] Lample G, Conneau A. Cross-lingual language model pretraining[J]. arXiv preprint arXiv: 1901.07291, 2019.

赖华(1966—)，硕士，副教授，硕士生导师，主要研究领域为智能信息处理。
E-mail：405904235@qq.com

黄于欣(1983—)，通信作者，博士，主要研究领域为自然语言处理、文本摘要等。
E-mail：huangyuxin2004@163.com

高玉梦(1995—)，硕士研究生，主要研究领域为自然语言处理、文本摘要。
E-mail：625150695@qq.com