译道
Way to Translation
2024 年 第 4 卷 第 2 期

机器翻译评估指标研究 *

牟正帅 [1] 奚瑞祺 [1,2] 成嘉垠 [1,2]

1．中南财经政法大学外国语学院AI翻译产教融合创新实践基地，武汉；
2．中南财经政法大学统计与数学学院，武汉

摘　要 | 机器翻译评估指标作为衡量翻译系统性能的重点研究对象之一，其研究不仅关系到翻译质量的判断，也对翻译技术的进步具有深远影响。随着深度学习技术的发展，大语言模型在机器翻译领域取得了显著进展，展现出强大的多模态翻译能力。然而，这些新模型也对传统翻译评估指标提出了新的挑战。针对这一问题，本研究系统性地分析了基于预训练语言模型的机器翻译评估指标。研究涵盖了基于词重叠、词向量、距离的评价指标，以及CIDEr、SPICE等其他指标，并从语言学的角度对各个指标进行了探讨分析。同时，对机器翻译质量评估方案进行了分类，包括单一方案评估、多方案对照评估和多方案结合评估，并对不同方案的优缺点进行了分析。最后，针对大语言模型翻译质量评估，提出了多方案结合评估的建议，以更全面地评估翻译的语义保持、文化适应性和语法正确性等诸多方面。本研究旨在为评估大语言模型翻译性能提供重要的理论参考。

关键词 | 机器翻译；预训练语言模型；翻译质量；翻译评估

一、引言

在当前全球化的时代背景下，语言的多样性和交流的需求日益增长，机器翻译（MT）技术作为打破语言障碍的关键工具，其应用范围和影响力持续扩大。然而，由于语言的复杂性以及文化差异的存在，机器翻译系统在翻译过程中可能产生误解、歧义或其他错误，影响翻译结果的准确性和自然性。因此，评估和保证机器翻译输出的质量变得尤为重要。

以ChatGPT为代表的一众大语言模型的底层架构为transformer模型，其不同于传统翻译模型（如RNN），它抛弃了传统的模型主题，只保留注意力机制，也抛弃了时序处理等处理方式，使用自注意力实现全局，并利用query，key，value三个矩阵对输入矩阵进行处理，计算权重并得出注意力分布（Vaswani J，et al.，2017）。不过ChatGPT对transformer模型进行了一定的调整。解码器的第一

* 指导教师：胡惮。

https://doi.org/10.35534/wtt.0402006

层和第二层进行对输入数据的处理，第三层对下一词进行预测并输出。因此，ChatGPT可以进行交互翻译和在一个文本框内进行连续翻译。利用prompt引导ChatGPT，形成交互式人机协同翻译模式，可以帮助提升人机翻译的翻译质量。用户可在翻译过程中提供译文干预，实时提供翻译辅助信息；同时机器也能及时获取用户的修改反馈，帮助优化翻译质量。

机器翻译质量评估（Quality Estimation，简称QE）旨在预测机器翻译结果的质量，并为每个翻译输出提供可靠的质量分数或反馈。既包含传统的机器翻译评估方法（韦佑武等，2022）如BLEU、METEOR，也包括新兴的WordVector、图论等方法，QE在实际应用中具有广泛的适用性，能够提供实时、客观的质量反馈，为用户和翻译服务提供者带来显著的价值。

然而，对机器翻译评估方法本身的系统性研究相对较少，尤其是对自动评估方法的深入分析。随着人工智能和深度学习技术的快速发展，机器翻译质量评估的研究和应用也在不断进步中。现代QE系统能够基于复杂的模型，如神经网络，词向量，预训练语言模型准确地预测翻译的准确性、流畅性和其他质量维度，甚至能够识别和指出翻译中的具体错误。这些进步不仅提高了MT系统的实用性和用户满意度，而且对于机器翻译技术的持续改进和优化提供了重要的支持（Bojar O，et al.，2017）。

本文旨在探讨基于预训练语言模型的机器翻译评估指标，分析现代QE系统如何利用深度学习技术提升翻译质量评估的准确性和效率，进而推动机器翻译技术的持续发展和优化。

二、机器翻译评估指标发展现状

笔者在知网上以“机器翻译质量评估”为主题进行发散性搜索。目前，研究主要集中在对机器翻译质量评估的现状、应用场景及发展建议。李行健（2022）在MQM模型的基础上，选择了三个具体且重要的量化指标作为机器翻译的质量测试手段，并制定了MQM模型的具体使用方法和医学文本的选择方法，促进了机器翻译在医学领域的应用。黄辉（2023）针对机器翻译质量评估的问题，对基于预训练模型的翻译质量评估算法，在低资源场景下特征融合和数据稀缺两方面的问题进行了研究，挖掘出预训练模型在机器翻译质量评估任务上的潜力。冯勤（2022）以指代消解为驱动，在利用有限和全局篇章信息的篇章神经机器翻译模型中，从观察到的源文和译文存在的上下文语义和指代差异出发，研究了不依赖参考翻译完成译文质量评估的方法。陆晓蕾、管新潮（2023）基于N-gram相似度、翻译编辑率、术语质量和向量相似度等评估指标，创新性地引入基于Python统计与词向量聚类进行量化实验，为ESP的跨学科研究提供了有益借鉴。叶恒、贡正仙（2023）提出了一种基于平行语料的翻译知识迁移方案，采用跨语言预训练模型XLM-R构建神经质量评估基线系统，并在此基础上提出了三种预训练策略以增强XLM-R的双语语义关联能力。陈磊、李泽世（2023）等结合学术论文标题翻译的特点，基于译文质量评估的三个方面，分析了以BLEU为代表的常用机器翻译质量评估指标的局限性，并尝试构建质性分析与量化统计相结合的机器翻译质量多维量化评估标准。韦晓保、陈巽（2023）基于错误分析理论，构建了译文评价多维指标体系，并通过运用熵权法完成了对译文质量评估矩阵的权重计算，为改善机器翻译质量提供了理论支撑。刘世界（2024）对六大主流翻译引擎的译文进行了定量定性评估，从四个方面出发总结研究成果，并提出改善模型、建立垂直语料库等相关建议，为未来机器翻译在涉海领域的应用提供了方向。

此外，国外学者的研究也为我们提供了宝贵的视角。例如，Callison-Burch等人的研究强调了机器翻译评估中统计显著性测试的重要性，这对于确保实验结果的可靠性具有关键作用。Hovy等人的工作则集中在内容要求上，为提高机器翻译的质量提供了指导。Shen等人通过词汇方法对机器翻译的质量进行评估，为自动评估方法的发展做出了贡献。Koehn的研究则关注于统计测试在机器翻译评估中的应用。同时，国外的研究者也在探索机器翻译后编辑工作量与翻译质量之间的关系，这对于理解机器翻译输出的实用性和改进翻译模型至关重要。

通过整合国内外的研究成果，我们可以更全面地理解机器翻译质量评估的复杂性和挑战性，以及如何通过多维度的方法来提高评估的准确性和效率。这些研究不仅为我们提供了丰富的理论支持，也为机器翻译技术的未来发展指明了方向。

三、基于词重叠的评价指标

基于词重叠的评价指标关注词汇分布的相似性，主要包括BLEU、ROUGE、METEOR、NIST和distinct。其中BLEU和METEOR常用于机器翻译任务，ROUGE常用于自动文本摘要。

（一）BLEU

BLEU（Bilingual Evaluation Understudy）是由Papineni等人于2002年提出的一种评估机器翻译质量的指标。该算法通过计算机器翻译与参考翻译之间的n元组（1-gram到4-gram）匹配度，来评估翻译的忠实度和流畅度。具体计算公式如下：

$$BLEU = BP * exp(\frac{1}{N}\sum_{n=1}^{N} ln(P_n))$$

$$P_n = \frac{\sum_i^E \sum_k^K min\left(h_k(c_i), min_{j\in M} h_k(s_{i,j})\right)}{\sum_i^E \sum_k^K min\left(h_k(c_i)\right)}$$

此处为利用均匀加权的几何平均法BLEU值计算公式，共有E个生成译文 c_i 和M篇相对应的参考译文 $s_{i,j}$。$h_k(c_i)$ 表示第k个N-gram在生成译文 c_i 中出现的次数，$h_k(s_{i,j})$ 表示表示第k个N-gram在参考译文 $s_{i,j}$ 中出现的次数，$min_{j\in M} h_k(s_{i,j})$ 相当于匹配与生成译文最相似的参考译文。

通过这个公式计算来自人工译文的N-gram与参考译文的匹配程度。当N取1时，计算词语忠诚度；当N大于1时，可计算词组流畅度；当N过大时，失去评估意义，一般最大取4。而当生成译文较多时，匹配率较高，为防止短句现象，设定简短惩罚因子BP，l_c 为生成译文长度，l_s 为参考译文长度。

$$BP = \begin{cases} 1 & if \quad l_c > l_s \\ e^{1-\frac{l_s}{l_c}} & if \quad l_c <= l_s \end{cases}$$

示例

Reference	It is a guide to action that ensures that the military will forever heed Party commands	BLEU
Candidate 1	It is a guide to action which ensures that the military always obeys the commands of the party	0.412
Candidate 2	It is a guide to action which that the military always the commands of the party	0.394
Candidate 3	It is to insure the troops forever hearing the activity guidebook that party direct	$5.510*10^{-155}$

在这一实例中，BLEU算法将比较机器译文与参考译文之间的N-gram匹配情况，以此得出机器翻译接近专业的人工翻译的程度。候选译文1与参考译文之间在词汇选择和结构上较为接近，表现出较高的一致性；相比之下，候选译文2和候选译文3虽然传达了相似的意思，但在词汇选择和句子结构上与参考译文的差异更大。

然而，从语义准确性和流畅性角度而言，BLEU的语义敏感度有限，因此语义不通顺但结构类似的句子也可能获得较高的分数。同时，BLEU对参考译文的依赖性较大，不同参考译文可能导致评分有较大差异。

（二）ROUGE

ROUGE（Recall-Oriented Understudy for Gisting Evaluation）是由Lin于2004年提出，用于评估自动译文质量的指标。它计算机器译文与参考译文之间的n元组（ROUGE-N）或最长公共子序列（ROUGE-L）的召回率，强调捕获参考译文中信息的多少，即在参考译文中出现的信息有多少被生成译文所覆盖。同时还存在两种方法的扩展版本，计算加权最长公共子序列召回率的ROUGE-W和允许跳跃二元组匹配的ROUGE-S。

1. ROUGE-N

ROUGE-N计算的是系统生成译文与参考译文之间的n-gram召回率，其公式如下：

$$ROUGE-N = \frac{\sum_{S\in\{ReferemceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S\in\{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)}$$

其中，$Count_{match}(gram_n)$ 表示在参考译文中与系统生成译文共有的n-gram数目，而 $Count(gram_n)$ 则表示在所有参考译文中n-gram的总数目。

	ROUGE-2	BLEU
Candidate 1	0.53	0.41
Candidate 2	0.47	0.39
Candidate 3	0.07	$5.51*10^{-155}$

在这一实例中，我们以ROUGE-2为例，BLEU和ROUGE指标在对不同候选翻译的评价上呈现出既有相似性又有差异性的特点。ROUGE指标着重于召回率，即参考文本中的信息有多少被候选文本成功捕获。对于Candidate 1与Candidate 2，BLEU

https://doi.org/10.35534/wtt.0402006

和ROUGE指标保持了一定的一致性，差异主要体现在Candidate 3，其在语言学的准确性上得分极低（由BLEU评分反映），但仍能够捕获少量的原文意义（由ROUGE评分反映）。

ROUGE-N评估体现了理解和生成语言的一个重要方面——即能够捕捉和重现语言的局部模式。N-gram的使用揭示了语言生成中的模式和规律性，这对于模拟人类语言处理过程至关重要。 不同的N值让ROUGE-N能够捕获不同长度的语言模式，从单个词（N=1，即ROUGE-1）到较长的短语或句子片段（例如，N=2或更大）。这反映了语言的多样性和复杂性，也使评估更加灵活和全面，间接反映了语言理解的深度。良好的N-gram重叠通常表明较好的内容和意义的保留，尽管这也暴露了ROUGE-N和其他基于形式匹配的评估方法的局限性——它们可能无法完全捕获语言的语义和语用层面。

2. ROUGE-L

ROUGE-L通过最长公共子序列（LCS）来评估参考译文与系统生成译文的相似度，旨在更好地捕捉长序列的相似性和句子结构的一致性。LCS（Largest Commen Sequence）允许非连续匹配，因此更能灵活地应对词序变化，在评估译文时能够考虑更多的语义和结构信息。

$$ROUGE-L=\frac{\sum_{s_i\in S_1}\max_{s_j\in S_2}LCS(s_i,s_j)+\sum_{s_j\in S_2}\max_{s_i\in S_1}LCS(s_i,s_j)}{\sum_{s_i\in S_1}length(s_i)+\sum_{s_j\in S_2}length(s_j)}$$

Reference 1	A B C D E F G H	ROUGE-2	ROUGE-L	Largest Commen Sequence	Length
Candidate 1	A B E F C D I	0.43	0.53	A B E F	4
Candidate 2	A B F C D E J	0.43	0.76	A B C D E	5

在这一对实例中，这两个candidate有相似的ROUGE-2得分，因为它们都包含了类似的词汇对，例如“A B”“C D”等。ROUGE-2会评估这些连续的bi-gram匹配。然而，在ROUGE-L评分中，ROUGE-L通过评估最长公共子序列（LCS）来测量句子结构的一致性和整体相似性。Candidate 2与参考译文保持了较高的词序和结构相似性，因为大部分词汇和短语的顺序与参考译文相似，即使不是完全连续的。而Candidate 1虽然包含了数量相同的关键信息，但由于相关单词的重新排序，其LCS与参考译文相比会稍短一些，因此在ROUGE-L评分中得到较低的评分。

ROUGE-L通过关注最长公共子序列，凸显了文本中单词顺序的重要性，单词顺序在很多语言任务的意义传达上意义重大。因此，ROUGE-L在评估生成译文时，能在一定程度上反映其在词汇使用、语法结构和语义保持等方面的能力。然而，ROUGE-L依然存在局限，尤其是在处理富有创造性的文本或需要深层语义理解的任务时，它可能无法完全区分生成译文的细微差别，评估效果不够完善。

3. ROUGE-W

ROUGE-W，即加权最长公共子序列（Weighted Longest Common Subsequence），是对ROUGE-L的扩展。与基本的最长公共子序列（LCS）方法相比，ROUGE-W进一步考虑了序列中连续匹配的长度，给予不同加权，从而区分不同空间关系下的LCS，更加精细地评估译文的质量，进一步改进对文本相似性的衡量。

在定义ROUGE-W时，首先需要计算两个句子（或译文）X和Y之间的加权最长公共子序列（WLCS）得分。通过动态规划方法，维护一个二维表格，记录到当前位置为止的最长公共子序列得分，以及连续匹配的长度。如果在某一位置上X和Y的单词相匹配，则根据之前连续匹配的长度（k）更新当前位置的得分，得分的更新取决于连续匹配长度的权重函数f。权重函数f需要满足某些性质，比如对于任意正整数x和y，有$f(x+y)>f(x)+f(y)$，这确保了连续匹配比非连续匹配获得更多的得分。之后使用一个特定的函数g对得分进行归一化处理，得到两种形式的得分：基于召回率的得分（R_{wlcs}）和基于精确度的得分（P_{wlcs}）。最后，结合这两种得分，计算出最终测度（F-measure）。

$$R_{wlcs}=g^{-1}\left(\frac{WLCS(X,Y)}{f(m)}\right)$$

$$P_{wlcs}=g^{-1}\left(\frac{WLCS(X,Y)}{f(n)}\right)$$

$$F_{wlcs}=\frac{\left(1+\beta^2\right)\cdot R_{wlcs}\cdot P_{wlcs}}{R_{wlcs}+\beta^2\cdot P_{wlcs}}$$

		ROUGE-L	ROUGE-W	Largest Commen Sequence	Length
Reference 1	A B C D E F G				
Candidate 1	A B C D X X X	0.67	0.53	A B C D	4
Candidate 2	A X B X C X D	0.67	0.76	A B C D	4

在这一对实例中，这两个candidate有相同的ROUGE-L得分，因为他们具有相同的最长公共子序列，但candidate 1保持了较长的连续匹配序列，而candidate 2 的匹配序列不连续，因此在计算权重函数f时不会获得较高的分数。

ROUGE-W不仅考虑了最长公共子序列的长度，还关注了其中单词的分布，以提供更精细的生成文本质量评估。通过加权，该算法更加注重连续匹配的长度，从而提高了对文本流畅性和连贯性的敏感度。通过调整权重函数f和权重值β，ROUGE-W能够更好地平衡精确度和召回率，根据不同任务的评估需求，提供了一种可调节的译文评估方法。ROUGE-W的引入体现了对语义完整性和叙述连贯性的重视，同时说明了自动文本生成评估方法向着更细致和全面的方向发展。然而，与其他自动评估指标一样，ROUGE-W也存在局限性，尤其是在评估语言创造性和复杂语义表达方面。

4. ROUGE-S

ROUGE-S是对ROUGE-N的扩展，旨在解决连续n-gram和LCS无法有效捕获的跳跃或间隔词组合的问题。其计算方法与ROUGE-N类似，但允许跳跃二元组（即允许词之间有间隔的二元组）的匹配，使得ROUGE-S在处理语言中的置换和修饰结构时更加有效。

$$R_{skip2}=\frac{SKIP(X,\ Y)}{C(m,2)}$$

$$P_{skip2}=\frac{SKIP(X,\ Y)}{C(n,2)}$$

$$F_{skip2}=\frac{\left(1+\beta^2\right)R_{skip2}P_{skip2}}{R_{skip2}+\beta^2P_{skip2}}$$

ROUGE-S通过允许跳跃二元组的匹配，能够捕捉更宽松的语序关系和结构灵活性，不再局限于严格的连续词序列，这使得它更适合评估包含复杂语言结构和丰富信息的文本。在处理复杂句子或信息密度高的段落时，与严格的连续词序列方法相比，ROUGE-S的跳跃二元组方法提供了更多的灵活性，有助于捕捉语义的完整性。ROUGE-S实现在连续词序列的ROUGE-N和基于最长公共子序列的ROUGE-L方法基础上，从不同角度，特别是在语言的灵活性和语义保持方面评估自动文摘或机器翻译的质量。同时，它还允许基于跳跃数量和权重等参数进行扩展。

（三）METEOR

METEOR（Metric for Evaluation of Translation with Explicit Ordering）是一种基于BLEU和ROUGE改进的评价指标，它同时考虑了同义词匹配、词干匹配和词序等因素。它使用调和平均数综合准确率和召回率，并引入词序惩罚因子，以反映词序对翻译质量的影响。具体计算公式如下。

$$P=\frac{m}{l_c}$$

$$R=\frac{m}{l_s}$$

m为一元组在生成译文中的匹配数，METEOR算法通过引入WordNet词典将同义词与词干匹配也增加匹配数，简化准确率P和召回率R的计算公式，再利用调和平均数计算准确率与召回率的综合指标，并设定词序惩罚因子Penalty，惩罚词序错误的情况。

https://doi.org/10.35534/wtt.0402006

$$F_{mean}=\frac{P*R}{\alpha P+(1-\alpha)R}$$

$$Penalty=\gamma\left(\frac{chunks}{m}\right)^{\beta}$$

$$METEOR=(1-Penalty)F_{mean}$$

chunk指匹配的单元组构成的连续词组也匹配的数量，通过调节α，β和γ可以控制召回率与准确率的权重以及惩罚因子的大小。METEOR算法综合考虑准确率与召回率，并将词序作为语句流畅度的衡量标准，但同时单元组计算会误判语法结构，也受到参考译文的选择影响，同样适合较短译文。

相比其他评价指标如BLEU和ROUGE，METEOR提供了一种更为复杂和细致的评估方法。METEOR考虑了词汇的精确匹配、同义词匹配，以及词序（即单词的相对位置）等因素，以此来评估机器翻译输出与人类翻译之间的相似度。METEOR试图更全面地捕捉翻译质量的多个维度，体现了对语言深层次特性的理解和重视，尤其是在保持语义完整性、尊重词序，以及平衡评估的准确性和全面性方面。

四、基于词向量的评价指标

词向量是利用高维空间中的向量表达词语信息，以便能够捕捉并表达词语之间的语义和语法关系，这些向量是通过机器学习模型从大量文本数据中学习得到的。现在较为常用的词向量表示方法有：Word2Vec，GloVe（Global Vectors for Word Representation），FastText等。基于词向量的评价指标关注语义分布的相似性，主要包括Embedding Average Score、Greedy Matching Score、Vector Extrema Score等。

（一）Embedding Average Score

该指标通过计算参考译文和生成译文中所有词向量的平均值，来获得两个文本的代表向量。然后，通过计算这两个向量之间的余弦相似度，评估生成译文与参考译文之间的语义相似度。这种方法简单直观，但由于没有权重因子，可能会忽略词汇的语义贡献差异。该算法利用的主要公式如下：

$$CosineSimilarity=\frac{\mathbf{V}_{ref}\cdot\mathbf{V}_{gen}}{\|\mathbf{V}_{ref}\|\cdot\|\mathbf{V}_{gen}\|}$$

其中，$\mathbf{V}_{ref}$和$\mathbf{V}_{gen}$分别是参考文本和生成文本的词向量平均值

Reference	The quick brown fox jumps over the lazy dog.	EAS	ROUGE-L
Candidate 1	The quick brown fox leaps over the lazy dog	0.954	0.75
Candidate 2	An agile fox jumps across a sleepy canine	0.717	0.125

在这一实例中，我们采用GoogleNews-vectors-negative300作为评估的词向量模型。观察到Candidate 2在语义上与参考句子保持了一定的一致性，但通过完全不同的词语表达了相似的意思（例如，“an agile fox”代替了“the quick brown fox”，“across a sleepy canine”代替了“over the lazy dog”）。在这种情况下，EAS仍能相对较好地识别出两个句子间的语义相似度（0.717），而ROUGE-L指标则因为缺乏直接的词汇重叠，给出了显著更低的评分（0.125）。这表明EAS在评估涉及较大词汇变动但语义保持一致的句子时的有效性和灵活性。

与依赖于字面重叠的ROUGE-L相比，EAS通过深入分析词向量的语义信息，为评估提供了一种更为细腻和全面的方法。这种方法尤其适用于那些在保持核心意义不变的同时，通过不同的表达方式传达信息的文本。

（二）Greedy Matching Score

此方法通过在参考文本和生成文本中寻找最相似的单词对，并计算它们词向量的余弦相似度，以评估两个文本的相似度。通过贪婪匹配算法，每个词都寻找与之最相似的对应词，从而更好地捕捉细粒度的语义匹配。然而，该方法可能过度强调单个单词间的匹配，而忽视整体语义结构。具体步骤如下：

（1）分词：将参考文本和生成文本分割成单词列表。

（2）遍历参考文本中的每个单词，计算其与生成文本中单词的余弦相似度，并记录最大相似度。

（3）遍历生成文本中的每个单词，寻找与之最相似的参考文本中的单词，并记录相似度。

（4）计算最终的贪婪匹配得分，即上述步骤中相似度的平均值，代表两段文本间的整体相似度。

https://doi.org/10.35534/wtt.0402006

Reference	The economic growth in the first quarter exceeded expectations.	GAS	EAS
Candidate 1	The first quarter's economic expansion was beyond expectations.	0.71	0.86
Candidate 2	Growth in economy has surpassed the first quarter's forecast.	0.73	0.82

在使用Embedding Average Score进行评估时，Candidate 1整体上与Reference在语义上非常接近，词汇的平均嵌入向量相似度高。而Candidate 2的词汇和句子结构与参考译文相比变化较大，因此整体的平均词向量相似度降低，得分较低。但在使用Greedy Matching Score进行评估时，由于其专注于词对词的最佳匹配，即使是在结构和词汇使用上有较大差异的情况下，Candidate 2中的某些词汇（如“economic growth”和“growth in economy”）会找到与参考译文中相对应词汇的高相似度匹配，从而获得较高的分数。这展示了Greedy Matching Score在评估单词之间细粒度语义匹配方面的独特优势。

（三）Vector Extrema Score

该方法通过在每个维度上取句子中所有词向量的最大值，构建代表句子的向量。然后，计算两个句子向量之间的余弦相似度。这种方法通过考虑词向量的极值，能更有效地捕捉文本的关键信息和语义特征。

Vector Extrema方法的核心思想是提取每个句子中词向量的最极端特征值来代表整个句子的语义。主要包含以下步骤：

（1）分词：将两个句子分割成单词列表。

（2）获取词向量：遍历两个单词列表，获取每个单词的词向量，形成两个包含有效词向量的列表。

（3）处理无有效词向量情况：如果某个句子中没有任何单词的词向量在模型中找到，返回默认值0.0。

（4）构建Vector Extrema向量：对每个维度使用np.max函数找出最大值，形成代表各自句子的Vector Extrema向量。

（5）计算余弦相似度：使用余弦相似度公式计算两个Vector Extrema向量之间的相似度，并通过1 – cosine转换确保相似度为正值。

余弦相似度的值范围从–1（完全不相似）到1（完全相似），这里通过1 – cosine转换，确保相似度为正值，越接近1表示相似度越高。

Reference	The president announced a groundbreaking climate policy initiative today	VES
Candidate 1	Today a revolutionary policy on climate was unveiled by the leaders	0.914
Candidate 2	The leader talked about the weather policy on TV today	0.884

这一实例展示了VES在强调句子中关键和强烈词汇的能力，尤其是在这些词汇在表达文本的核心语义和情感色彩方面起到决定性作用时。因此，VES提供了一个补充视角，特别适用于那些关键信息和语义强度对于整体意义理解至关重要的评估场景。

五、基于距离的评价指标

基于距离的评价指标是一种常见的“错误率”度量方法，类似于WER、PER等，这些指标主要用于评估机器翻译任务。

TER（Translation Edit Rate）是一种评估机器翻译输出质量的直观方法，它旨在量化人类编辑需要对机器翻译结果进行多少编辑，以使其与参考翻译完全匹配。TER通过计算将机器翻译输出转变为与至少一个参考翻译完全匹配所需的最少编辑次数（包括插入、删除、替换和移动词组）来实现，然后将这个数字标准化，以参考翻译的平均长度为基准。

HTER（Human–targeted Translation Edit Rate）是TER的一种变体，涉及通过人工编辑来创建“目标参考翻译”。这种方法通过让流利的目标语言使用者编辑机器翻译的输出，直到它既流畅又与原始参考翻译具有相同的意义，从而生成一个新的目标参考翻译。然后，利用这个单一的目标参考翻译来计算HTER，这反映了将机器翻译输出转变为具有相同意义且流畅的目标语言句子所需的最少编辑次数。与单一参考的TER相比，HTER与人类对机器翻译质量的评价具有更高的相关性。然而，HTER的计算需要人类的编辑，人工成本更高，花费时长更大，不适合用于机器

https://doi.org/10.35534/wtt.0402006

翻译系统的开发周期中，但它可以作为评价机器翻译质量的人类主观判断的可能替代。

六、其他评价指标

（一）TF-IDF

TF-IDF（Term Frequency-Inverse Document Frequency）是一种在自然语言处理和信息检索领域广泛应用的数学模型，旨在表示文档或文本数据的重要性。该模型通过词频（TF）和逆文档频率（IDF）两个统计量来量化词语在文档集合中的重要性。计算文档中所有词的TF-IDF值，并按一定顺序排列成向量，每个维度代表一个特定的词，其值是该词在文档中的TF-IDF值。TF-IDF向量能够很好地捕捉词语在文档中的重要性，同时降低文档集合中普遍出现的词的影响。通过这种方式，算法可以处理文本数据，进行文本相似度计算、文本分类、信息检索等任务。TF-IDF有效地表达出词语在特定文档中的相对重要性，为这些任务提供了一种强有力的特征表示方法。

（二）CIDEr

CIDEr（Consensus-based Image Description Evaluation）方法将每个句子视为一个“文档”，计算其TF-IDF向量，然后计算参考描述和生成描述的余弦相似度，以反映描述的一致性，模拟人类描述中的共识形成过程。具体步骤如下：

（1）将参考文本和翻译文本组合成文档集合。

（2）计算TF-IDF向量，得到TF-IDF矩阵。

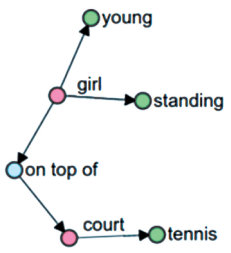
（3）使用线性核函数计算文档A和文档B的余弦相似度值。

Reference	It is a guide to action that ensures that the military will forever heed Party commands	CIDEr
Candidate 1	It is a guide to action which ensures that the military always obeys the commands of the party	0.580
Candidate 2	It is a guide to action which that the military always the commands of the party	0.562

CIDEr通过对参考描述集合中词汇的统计分析，尝试模拟人类在描述时的共识。这种方法体现了对语言使用中共识形成过程的认识，即人们在描述同一图像时往往会选择相同或相似的词汇和表达方式。

（三）SPICE（Semantic Propositional Image Caption Evaluation）

SPICE（Semantic Propositional Image Caption Evaluation）是一种评估图像字幕生成的方法。它利用图论将描述转化为语义结构，计算描述中的对象objects（图中红色）、属性attributes（图中绿色）和关系的匹配度relationship（图中蓝色），将单一语句转化为树状结构，强调对描述中深层语义的理解。以图1为例。



A young girl standing on top of a tennis court.

图1　SPICE 评估法实例

在完成映射构建后，进一步将scene graphs转化为元祖集合，例句转化元祖集合为：{（girl），（court），（girl，young），（girl，standing），（court，tennis），（girl，on-top-of，court）}。接着与METOR类似，计算元祖间匹配率，得到生成句评分。该方法更广泛运用于图片生成文字领域，需要更多相似例句集作为比较对象，对于机器翻译评估领域目前工作量较大。

$$G(c)=O(c),E(c),K(c)$$

$$T(G(c))\triangleq O(c)\cup E(c)\cup K(c)$$

$$P(c,S)=\frac{\left|T(G(c))\otimes T(G(S))\right|}{\left|T(G(c))\right|}$$

$$R(c,S)=\frac{\left|T(G(c))\otimes T(G(S))\right|}{\left|T(G(S))\right|}$$

$$SPICE(c,S)=F_1(c,S)=\frac{2\cdot P(c,S)\cdot R(c,S)}{P(c,S)+R(c,S)}$$

SPICE不仅仅依赖于词语的匹配，而更看重对于图像描述中深层语义信息的理解和评估。这也使得它能够在一定程度上处理并评估语言中的一些复杂现象，如同义词的使用、复杂句子结构等，在对语言多样性和复杂性有了更深层次的应用，有助

https://doi.org/10.35534/wtt.0402006

于推动自然语言生成技术向更高的准确性和自然性发展。

七、大语言模型翻译质量的评估方案探讨

（一）评估方案分类

单一方案评估：采用如BLEU或METEOR等单一指标进行翻译质量评估，此类方法操作过程简单，能快速地对大量翻译文本进行自动评估，从而节省评估时间和评估资源。但这种方法通常只从某一角度关注词汇和句法的对应关系，难以全面捕捉翻译的语义准确性、文化适应性和语境恰当性等多维度质量。此外，此方法也无法评估翻译的自然流畅度和目标语言用户的接受度。

多方案对照评估：运用多种评估工具进行平行比较，如结合BLEU、METEOR与人工评估。这种方法可以从不同角度分析翻译质量，得到更全面的结果，同时提供额外洞见。不同工具间的评分差异能够揭示模型在某些翻译维度上的具体弱点，为模型优化提供指向性建议。

多方案结合评估：采用综合评估模式，结合多种量化指标和定性评估来全面评价翻译质量。例如，可以结合自动评估的一致性和人工评审的深度反馈，针对具体语言资源特点提出一种或几种评议模型。通过加权得分分析模型在语义保持、文化适应性、语法正确性等方面的表现。但需要额外对加权评议模型的有效性进行验证，且需要投入较多的时间和资源。

（二）评估方案分析

翻译模型的发展经历了基于规则的模型、基于统计的模型、基于神经网络的模型以及基于大语言模型的预训练模型。大语言模型在自然语言处理领域取得了重大突破，显著提升了自然语言处理的能力。评估模型也随着翻译模型的发展而发展。我们应当注意到，大语言模型翻译相较于以往的翻译模型，最显著的特点便是其强大的自然语言生成能力。然而，翻译内容易出现“幻觉”（hallucination）现象，传统的单一评估模型难以对此类错误进行有效惩罚。多方案对照评估与多方案结合评估通过结合自动化和人工评估的优点，提供了最全面和深入的翻译质量评估，能够同时评估翻译的语义保持、文化适应性和语法正确性等多个方面。因此，对于旨在深入理解和优化翻译模型性能的项目，采用多方案对照评估与多方案结合评估是最佳选择。

参考文献

[1] Bojar O，Chatterjee R，Federmann C，et al. Findings of the 2017 Conference on Machine Translation（WMT17）[C]//Proceedings of the Second Conference on Machine Translation，2017.

[2] Callison-Burch C，Fordyce C，Koehn P，et al. Further Meta-Evaluation of Machine Translation：Strong Agreement，Strong Disagreement，and System Design [C]//Proceedings of the Second Workshop on Statistical Machine Translation，2007：122-129.

[3] Eduard Hovy，Margaret King，Andrei Popescu-Belis. Computer-Aided Specification of Quality Models for Machine Translation Evaluation [J]，2002.

[4] Joseph P Turian，Luke Shen，I Dan Melamed. 机器翻译的评估及其评估 [C]//Proceedings of Machine Translation Summit IX：Papers，New Orleans，USA，2003.

[5] Kishore Papineni，Salim Roukos，Todd Ward，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，Pennsylvania，USA. Association for Computational Linguistics，2002：311-318.

[6] Koehn P. Statistical Significance Tests for Machine Translation Evaluation [C]//Proceedings of the 2010 Workshop on Statistical Machine Translation，2010：10-18.

[7] Li J，Galley M，Brockett C，et al. A Diversity-Promoting Objective Function for Neural Conversation Models [J]. Computer Science，2015.

[8] Lin C Y. ROUGE：A Package for Automatic Evaluation of Summaries [C]//In Proceedings of

https://doi.org/10.35534/wtt.0402006

the Workshop on Text Summarization Branches Out（WAS 2004），2004.

［9］Vaswani J，Shazeer N，Parmar N，et al. Attention is all you need［J］. Advances in Neural Information Processing Systems，2017：6000–6010.

［10］Wei Zhao，Maxime Peyrard，Fei Liu，et al. MoverScore Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance［C］//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing（EMNLP–IJCNLP），Hong Kong，China. Association for Computational Linguistics，2019：563–578.

［11］陈磊，李泽世，王鹏晓，等. 学术论文标题翻译：常用机器翻译质量评估指标的局限性与多维量化评估标准的构建［J］. 科技传播，2023，15（16）：43–46，50.

［12］冯勤. 指代消解驱动的篇章神经机器翻译及质量评估的研究［D］. 苏州：苏州大学，2022.

［13］黄辉. 低资源场景下的翻译质量评估研究［D］. 北京：北京交通大学，2023.

［14］李行健. 基于多维质量指标模型的机器翻译质量评估实验［D］. 北京：首都经济贸易大学，2022.

［15］陆晓蕾，管新潮. 翻译质量评估的现状与对策：基于人文社科与自然科学文献的计量研究（1981—2021）［J］. 中国ESP研究，2023（1）：114–125，132.

［16］刘世界. 涉海翻译中的机器翻译应用效能：基于BLEU、chrF++和BERTScore指标的综合评估［J］. 中国海洋大学学报（社会科学版），2024（2）：21–31.

［17］韦佑武，李娜，赵良威. 机器翻译的译文质量、高频错误类型及解决对策研究： 基于机器翻译的发展史［J］. 现代语言学，2022，10（9）：1944–1949.

［18］韦晓保，陈巽. 基于熵权TOPSIS法的机器翻译译文测度［J］. 外国语（上海外国语大学学报），2023，46（6）：106–119.

［19］叶恒，贡正仙. 利用语义关联增强的跨语言预训练模型的译文质量评估［J］. 中文信息学报，2023，37（3）：79–88.

Research on Machine Translation Evaluation Metrics

Mou Zhengshuai[1] Xi Ruiqi[1,2] Cheng Jiayin[1,2]

1. Industry-Education Integration Innovation Practice Base for AI-powered Translation, School of Foreign Studies, Zhongnan University of Economics and Law, Wuhan;

2. School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan

Abstract: Evaluation metrics for machine translation, as a crucial research focus for measuring the performance of translation systems, significantly impact the judgment of translation quality and the advancement of translation technology. With the progress of deep learning, large language models have achieved remarkable advancements in the field of machine translation, showcasing strong multimodal translation capabilities. However, these new models also pose challenges to traditional translation evaluation metrics. This study systematically analyzes machine translation evaluation metrics based on pre-trained language models, including metrics based on word overlap, word vectors, and distance, as well as other metrics like CIDEr and SPICE. From a linguistic perspective, the study discusses and analyzes each metric. It also categorizes machine translation quality evaluation schemes into single-scheme evaluation, multiple-scheme comparative evaluation, and combined-scheme evaluation, analyzing their respective advantages and disadvantages. Finally, for evaluating the translation quality of large language models, the study suggests a combined-scheme evaluation to more comprehensively assess various aspects such as semantic retention, cultural adaptability, and grammatical correctness, aiming to provide important theoretical references for evaluating the performance of large language models in translation.

Key words: Machine translation; Pre-trained language models; Translation quality; Translation evaluation

https://doi.org/10.35534/wtt.0402006