计算机科学
COMPUTER SCIENCE

语义通信系统的性能度量指标分析

姜胜腾，张亦弛，罗鹏，刘月玲，曹阔，赵海涛，魏急波

引用本文

姜胜腾，张亦弛，罗鹏，刘月玲，曹阔，赵海涛，魏急波. 语义通信系统的性能度量指标分析[J]. 计算机科学, 2022, 49(7): 236-241.

JIANG Sheng-teng, ZHANG Yi-chi, LUO Peng, LIU Yue-ling, CAO Kuo, ZHAO Hai-tao, WEI Ji-bo. Analysis of Performance Metrics of Semantic Communication Systems [J]. Computer Science, 2022, 49(7): 236-241.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.211200071

语义通信系统的性能度量指标分析

姜胜腾　张亦弛　罗　鹏　刘月玲　曹　阔　赵海涛　魏急波
国防科技大学电子科学学院　长沙 410073
(jiangshengteng@nudt.edu.cn)

摘　要　语义通信系统是目前通信领域的研究热点,但是该领域尚未建立起成熟的评价体系,导致不同性能度量指标下设计的语义通信系统的性能也各不相同。文中主要针对语义通信系统,介绍了基于精确率的性能度量指标、基于召回率的性能度量指标、基于精确率与召回率相结合的性能度量指标以及基于词向量空间模型的性能度量指标;并详细阐述了语义通信中各种性能度量指标提出的背景、意义、主要算法思想和适用范围,对比分析了每一种性能度量指标间的差异和优缺点。最后,总结了现阶段语义通信性能度量指标面临的问题,并展望了语义通信系统中性能度量指标研究的未来发展方向。

关键词:语义通信;性能度量;文本评价;词向量
中图法分类号　TN919

Analysis of Performance Metrics of Semantic Communication Systems

JIANG Sheng-teng,ZHANG Yi-chi,LUO Peng,LIU Yue-ling,CAO Kuo,ZHAO Hai-tao and WEI Ji-bo
College of Electronic Science and Technology,National University of Defense Technology,Changsha 410073,China

Abstract　Semantic communication system is currently a hot research topic in the communication field,but a mature evaluation system has not been established in this field,which leads to the different performance of semantic communication systems designed under different performance metrics. This paper mainly focuses on semantic communication systems,introducing performance metrics based on precision,performance metrics based on recall,performance metrics based on the combination of precision and recall,and performance metrics based on word vector space models. It also elaborates on the background,significance,main algorithm ideas,and scope of application of various performance metrics in semantic communication,and analyzes and compares the differences,advantages and disadvantages of each performance metric. Finally,it summarizes the problems faced by semantic communication performance metrics at this stage,and points out the future development direction of performance metrics research in semantic communication system.

Keywords　Semantic communication,Performance measurement,Text evaluation,Word vector

1　引言

5G 技术的普及打开了信息世界新的大门,人类社会也走进了万物互联的新时代。未来,人们期望下一代无线通信系统能够以更高的通信质量、更快的传输速率来支持更加多元的应用。然而,随着现代通信技术的飞速发展和成熟,其传输速率已经逐步逼近香农容量,这导致未来通信技术的发展将面临一个瓶颈期[1],即如何突破香农极限传输速率[2]。作为一种全新的通信架构,语义通信是解决这一瓶颈的有效途径之一[3-4]。

语义通信模型的结构如图 1 所示,其中语义指待传输数据背后的含义。相比图 2 中的传统通信模型的结构,语义通信并不关注收发消息间的一致性,而是关注发送方与接收方对消息的理解是否相同[5]。这意味着传统通信系统的性能度量指标(如误码率和误比特率等指标)已不再适用于语义通信[6]。这些性能指标在评价语义通信系统时不能直接反映语义通信系统的性能,误码率较高时并不一定会影响语义通信系统正常的语义交互,如同义词替换。因此,语义通信系统的性能度量指标需要反映收发消息间的语义相似程度。

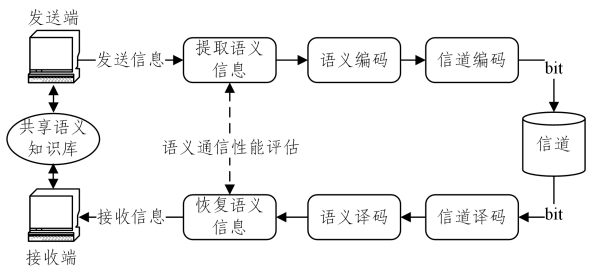

图 1　语义通信模型图

Fig. 1　Diagram of semantic communication model

到稿日期:2021-12-06　返修日期:2022-04-04
基金项目:国家自然科学基金(61931020,U19B2024,62001483)
This work was supported by the National Natural Science Foundation of China(61931020,U19B2024,62001483).
通信作者:曹阔(caokuo90@sina.cn)

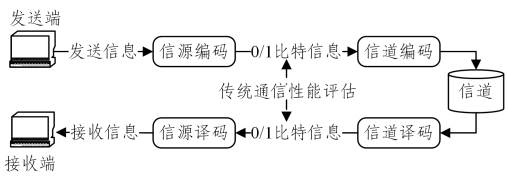

图 2 传统通信模型图

Fig. 2 Diagram of traditional communication model

由于目前对语义通信领域的研究还不够深入，关于语义通信系统的性能度量指标尚未建立完整的体系，缺乏对各种性能度量指标的优缺点的对比和总结，导致部分相关研究工作存在度量指标选择不合理的现象，并且相关研究工作之间无法直接进行对比和衡量。为解决这些问题，本文主要对目前语义通信领域内常用的性能度量指标进行归纳和总结。Papineni 等[7]提出了一种基于 n-gram 精确率的评价指标 BLEU(Bilingual Evaluation Understudy)，该指标的核心思想是使用机器进行文本评价，解决了人工评估成本过高的问题，但是该评价指标存在无法评价高级语义关系的问题。BLEU 自提出后被不断地改进与创新，Doddington 在其基础上通过引入 n-gram 的信息量[8]来得出每个 n-gram 的对应权重，进而计算加权得分，使评价结果更接近人工评价。Lin[9]对 BLEU 进行了优化，将 n-gram 精确率修改为召回率，实现了更好的文本真实性评价。Banerjee 等在文献[7]与文献[9]的研究的基础上将精确率与召回率相结合，提出了一种可以识别高级语义关系的评价方法[10]。另一方面，Agirre 等在神经网络的基础上，借助词向量训练模型，提出了基于词向量空间的语义评价[11]，这种评价方法能够表示一定程度的高层语义，如时态、近义词等。Vedantam 等[12]将词向量模型与精确率指标相结合，提出了一种可以评价文本信息量的性能度量指标。

本文介绍了语义通信系统中各种性能度量指标的提出背景与意义，并通过实验测试分析了不同性能度量指标的优缺点以及适用范围，为语义通信系统性能度量指标的选取提供了参考。

2 基于精确率的 BLEU 度量指标

语义通信系统的度量指标最早被用于文本评价，文本评价是自然语言处理中的一个重要领域，包括文本生成、机器翻译、语义相似等多种任务[13]。文本评价任务主要的评价方式是人工评估，这种评价方式服从人的思维逻辑性，评价结果合理且灵活性强，但是需要消耗大量的人力成本与时间成本[14]，这也限制了该领域的发展。

为了解决这一问题，Papineni 等于 2002 年提出了基于精确率的 BLEU 性能度量指标[7]。该指标计算待评价文本和参考文本间相同的单词数与待评价文本单词总数的比值，并将其作为精准率用于反映两者之间的相似程度。BLEU 在精确率的基础上被进一步优化，加入了 n-gram 精确率(文本中连续的 n 个单词被视为一个 n-gram)与短句惩罚因子，以进一步提高文本评价任务的准确性。BLEU 的核心思想是用机器评估代替人工评估，该方法被提出后，文本评价需要消耗大量人力成本的问题迎刃而解。BLEU 凭借其高效性以及与人类评价的高度相关性[15]直接取代了人工评估，成为了该领域通用的评估方式。

BLEU 是基于 n-gram 精确率的性能度量指标，其表达式为：

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{1}$$

其中，BP 是短句惩罚因子，目的是为了惩罚因句子过短而导致部分语义丢失的影响，可以表示为：

$$BP = \begin{cases} 1, & c > s \\ e^{(1-s/c)}, & c \leqslant s \end{cases} \tag{2}$$

其中，c 和 s 分别为待评价文本与参考文本的长度。w_n 为权重因子，其数值为 $1/N$，可使各阶 n-gram 服从均匀分布；p_n 代表参考文本 S 与待评价文本 C 间的 n-gram 精确率，可以表示为：

$$p_n = \frac{\sum_{n\text{-}gram \in S} Count_{\text{clip}}(n\text{-}gram)}{\sum_{n\text{-}gram' \in C} Count(n\text{-}gram')} \tag{3}$$

$$Count_{\text{clip}} = \min(Count, Max\_Ref\_Count) \tag{4}$$

其中，$Count$ 是 n-gram 在待评价文本中出现的次数，Max_Ref_Count 表示该 n-gram 在参考文本中的最大出现次数，最大匹配数通过取这两者的最小值来避免参考文本中的 n-gram 被重复匹配。对式(1)－式(4)进行分析可以得出，BLEU 在 n-gram 精确率的基础上引入了短句惩罚因子与最大匹配数来解决语义丢失和 n-gram 重复匹配的问题，并将其计算出的 n-gram 精确率作为评价分数，分数越高代表待评价文本与参考文本的语义越接近。

BLEU 首次提出了使用机器代替人类进行评估的理念，并且效果突出。由于机器自动评估是机器翻译系统与语义通信系统开发和评估的基础[16]，BLEU 自提出后受到了文本评价任务的青睐，被视为一种重要的评价指标[17]，且被广泛应用于自然语言处理、语义通信等多类任务中[18]。但是 BLEU 也存在一些问题[19-20]，例如在设计过程中没有考虑到高级语义，如单词时态、同义词、近义词等，因此 BLEU 在评价高级语义任务时效果并不理想。

3 基于召回率的 ROUGE 度量指标

文本评价在初期一直都是采用统计学方法来评价准确度与流畅度[21]，随着机器学习和神经网络技术的发展和进步，文本评价任务逐步开始利用神经网络方法来进行处理。由于神经网络强大的补充能力，得到的文本基本不会出现流畅性问题。然而，其缺点也很明显，即虽然恢复的文本流畅度高，但是补充信息过多会干扰用户判断，这将导致用户无法判别信息是否真实[22]。为了解决这一问题，Lin 于 2003 年提出了 ROUGE(Recall-Oriented Understudy for Gisting Evaluation)评价指标[9]。ROUGE 是一种基于召回率的文本评价指标，它是在 BLEU 的基础上改进而来的，核心思想是利用 n-gram 召回率来代替精确率进行度量。其中，召回率指待评价文本和参考文本间的相同的单词数与参考文本单词总数的比值，它反映了待评价文本中正确的 n-gram 数目。

由于文本评价任务具有多样性，针对不同任务时 ROUGE 度量指标的侧重点也各不相同。现在常用的指标主要为 ROUGE-N 和 ROUGE-L，其中 ROUGE-N 主要用来计算

待评价文本与参考文本在 n-gram 上的召回率，ROUGE-L 主要是将 ROUGE-N 中使用的 n-gram 替换为待评价文本与参考文本间的最长公共子序列，具体表达式为：

$$ROUGE\text{-}L=\frac{(1+\beta^2)R_{LCS}P_{LCS}}{R_{LCS}+\beta^2 P_{LCS}} \tag{5}$$

其中，R_{LCS} 与 P_{LCS} 分别是召回率和精确率，具体表达式为：

$$R_{LCS}=\frac{LCS(C,S)}{len(S)} \tag{6}$$

$$P_{LCS}=\frac{LCS(C,S)}{len(C)} \tag{7}$$

其中，C 是待评价文本，S 是参考文本，$LCS(C,S)$ 表示待评价文本与参考文本间最长公共子序列的长度，$len(S)$ 是参考文本的长度，$len(C)$ 是待评价文本的长度。式(5)中，β 表示召回率与精确率的权重比，由于在实际使用中 β 数值较大，因此 ROUGE-L 关注的重点是召回率而不是精确率。对 ROUGE-L 的计算式进行分析可以得出，ROUGE-L 本质上是优化后的 n-gram 召回率，其得分反映了待评价文本与参考文本间的语义相似程度，得分越高，文本间的相似程度就越高。

相比 BLEU，ROUGE 在评价文本语义任务中性能有一定的提升，但由于未考虑高级语义关系，因此提升效果并不明显，故应用范围不及 BLEU。与此同时，由于 ROUGE 更加关注召回率，因此在有多个候选文本时其性能表现会更好，但是这一特点并不适用于语义通信，因为语义通信针对的是收发消息间的单文件任务，候选文本相对匮乏。ROUGE 的缺点也很明显：1)与 BLEU 一样，ROUGE 没有考虑到高级语义关系；2)ROUGE 不能评价待评价文本的流畅度。这些缺点都会对复杂文本任务的评估产生不利影响。

4 基于精确率与召回率相结合的 METEOR 度量指标

随着文本评价的快速发展，为了保证性能度量指标的有效性，性能度量指标必须要与人类的判断具有高度相关性。首先，无论是在文本评价还是语义通信方面，人类评价才是最具权威的；其次，性能度量指标之间应当具有一致性，即语义相近的文本在同样的指标下应该得到相似的分数；最后，还需要保证度量指标的通用性与可靠性。然而，BLEU 与 ROUGE 在处理各项任务时无法兼顾上述所有要求。为解决这个问题，Lavie 等研究了精确率和召回率的权重与人类判断相关性之间的关系[23]，并提出了 METEOR(Metric for Evaluation of Translation with Explicit Ordering)这一新的性能指标[10]。

METEOR 是基于精确率与召回率的调和平均的性能指标。相比 BLEU 与 ROUGE，METEOR 在文本整体层面进行评价，而非单词层面。为了解决 BLEU 无法评价高级语义关系这一问题，METEOR 在设计过程中引入了 WordNet 单词集，该单词集是一个包含同义词组的大型词典，从而扩充了高级语义关系。除此之外，为了更好地评价句子整体的流畅性，METEOR 引入了词块的概念。

在设计性能度量指标时，单纯考虑精确率会忽略信息的真实性，只考虑召回率会忽略流畅度性能。鉴于这种情况，METEOR 采用了精确率与召回率调和平均的方式来进行评估，具体表达式为：

$$F_{\text{mean}}=\frac{P_m R_m}{\alpha P_m+(1-\alpha)R_m} \tag{8}$$

其中，P_m 为精确率，R_m 为召回率，α 为精确率与召回率间的调和平均比例。同时，为了更好地评价句子的流畅度，在词块的基础上设计了惩罚因子 Pen，具体表达式为：

$$Pen=\gamma\left(\frac{ch}{m}\right)^{\theta} \tag{9}$$

其中，ch 代表词块，m 是参考文本与待评价文本间相同的单词数目，γ 是惩罚因子的最大值，θ 是惩罚因子的衰减指数。综上，METEOR 可以表示为：

$$METEOR=(1-Pen)F_{\text{mean}} \tag{10}$$

在使用 METEOR 进行评价时，式(8)与式(9)中出现的 α,γ,θ 等参数需要根据实际任务进行调试。METEOR 通过对精确率与召回率进行调和平均，在保留了 BLEU 与 ROUGE 优势的同时，解决了 BLEU 与 ROUGE 无法评价高级语义关系的问题。METEOR 在识别高级语义关系方面的突出能力使其更加适合评估语义通信系统的性能。文献[24]中的语义图像命题系统就使用了 METEOR 作为其性能度量指标，并取得了优异的评价结果。但是 METEOR 也存在着一些问题，即 METEOR 依赖于特定数据集调试的超参数(α,γ,θ)，如果待评估语言不在 WordNet 数据集中，用户则需要寻找对应数据集并进行超参数调试，这将大幅度增加 METEOR 的使用难度，而且重新调试的参数效果也无法得到保障。因此 METEOR 在使用上存在一定的局限性，适用于一些特定任务，普适性不强。

5 基于空间词向量模型的语义相似度

神经网络不仅可以用来改善恢复文本的流畅度，也可以直接用于设计文本语义评价方法。在使用神经网络处理文本评价任务时，需要将文本转化为词向量再输入神经网络中，其中词向量技术是将自然语言中的单词转化为高维空间中的向量。在这个高维语义向量空间内，语义相近的单词对应的词向量在向量空间中的位置接近，而语义差别较大的词向量在空间中的位置较远[25]。词向量训练模型能够很好地反映单词间的语义关系，因此可以通过对比收发消息词向量间的相似度来判断两者的语义是否相同。

语义相似度的概念于 2012 年被首次提出，主要用于评估两段文本的语义相似程度[11]，语义相似度评价也是语义通信任务中语义恢复的关键问题[26]。在计算语义文本相似度前，首先需要使用词向量模型获取文本对应的词向量，目前性能较好的词向量模型主要有 Word2vec 模型[27]、词袋模型[28]、Glove 模型[29]和 BERT 模型[30]。语义相似度的定义也多种多样[31]，其中文献[11]提出的语义文本相似度的定义使用得较为广泛，计算式为：

$$sim_v(s_1,s_2)=\frac{\boldsymbol{v}(s_1)\boldsymbol{v}(s_2)}{\|\boldsymbol{v}(s_1)\|\ \|\boldsymbol{v}(s_2)\|} \tag{11}$$

其中，$\boldsymbol{v}(s_1)$ 与 $\boldsymbol{v}(s_2)$ 分别代表句子 s_1 与 s_2 的词向量，可以通过对句子中每个单词的词向量进行合成来得到句子的整体向量，这样才能更加精确地反映句子深层次的语义关系。从式(11)可以看出，语义相似度的物理意义就是两个句子词向量夹角的余弦值，向量夹角反映出了空间位置的相关性，同时

也代表了句子间的语义相似关系。

虽然语义相似度所采取的思想较为简单,但是向量夹角代表的语义关系让抽象的语义更加具体。然而,语义文本相似度也存在一些问题,其度量结果依赖于词向量训练模型的性能,不同的词向量训练模型得出的结果不同。因此,在使用语义文本相似度进行评估时,需要选取性能较好的词向量来训练模型。

6 语义通信系统评价指标对比

通过上述分析,语义通信系统的性能评价指标大致可以分为基于自然语言模型的评价方法与基于词向量模型的语义相似度评价方法。其中基于自然语言模型的评价方法符合人类认知的语言规律和逻辑性评价,但是灵活性较差,例如句型变换和同义词替换等高级语义问题会导致评价出现偏颇。基于词向量模型的语义相似度评价方法使用神经网络技术在大数据样本下生成词向量来表示语义信息,处理方式较为灵活,可以评价高级语义信息,但是其逻辑解释性较差,忽略了句子结构间的内在关联性,从而导致部分语言规律丢失。表 1 与表 2 分别列出了对这两类评价方法中的常见指标进行对比的结果。

表 1 基于自然语言模型的评价方法常见指标对比

Table 1 Comparison of common indicators of evaluation methods based on natural language models

评价指标	参考文献	优点	缺点
BLEU	文献[7]	以精确率为基础计算评价分数,泛用性强,效果接近人工评价	未考虑高级语义情况,在评价高级语义任务时效果不理想
ROUGE	文献[9]	以召回率为基础计算评价分数,考虑了传输信息间的相关性	未考虑句子结构,指标分支种类繁多,选取过程十分复杂
CIDER	文献[12]	通过引入词频-逆文件频率,可以评价文本蕴含的信息量	未考虑语法结构,过度依赖语义解析
METEOR	文献[10]	结合精确率与召回率进行设计,扩充了高级语义关系,同时可评价句子流畅性	依赖于特定数据集调试,使用上存在局限性,普适性不强

表 2 基于词向量模型的语义相似度评价方法常见指标对比

Table 2 Comparison of common indicators of semantic similarity evaluation methods based on word vector models

评价指标	参考文献	优点	缺点
Similarity score	文献[11]	首次提出通过计算词向量距离来评价信息语义,可以发现信息内部蕴含的深层语义	选取的模型直接使用嵌入词向量进行拼接获取句向量,误差较大
DSSM Similarity score	文献[32]	用字向量作为输入,提高模型的范化能力;精确度高	使用词袋模型导致语序信息和上下文信息丧失;预测结果不可控
CLSM Similarity score	文献[33]	通过卷积层提取滑动窗口上下文信息,通过池化层提取全局上下文信息,使上下文信息得以保留	无法捕捉间隔较远的上下文信息

语义可以视为一种客观知识,知识是通过对大量经验数据进行总结以及根据基本逻辑模型推理得到的。换言之,语义是在海量经验数据以及基本语言模型的基础上进行表征的。因此,语义通信系统必然是一个数据和模型双驱动的系统。语义通信系统的评价体系可以从经验数据和语言模型两个层面进行划分。其中基于词向量模型的语义相似度评价方法研究的是经验数据层面的指标,基于自然语言模型的评价方法研究的是语言模型层面的指标。综上,现有研究主要针对其中一个层面展开,无法准确刻画语义通信系统的性能。两个层面之间的融合评价是必然的发展趋势,并且两个层面之间的权重选择需要根据特定任务、特定对象和特定场景来决定。

7 实验分析

为了更加直观地反映各种性能度量指标的效果,我们选用 STS 数据集的训练集来进行实验分析。STS 数据集是专注于语义文本相似性基准测试的数据集[34],其训练集中共有 5749 对英文句子,在这些句子对中存在各种类型的语义关系,同时在数据集中也给出了对应的人工评价结果。在实验分析中,我们使用 NLTK 工具包[35-36]对数据集中的每一对句子依次使用上述的性能度量指标进行评估,然后将每种性能度量指标得到的所有测试结果取均值后与人工评价结果进行对比,以反映各种性能度量指标的评价效果。各种性能度量指标的平均评价结果如表 3 所列。

表 3 基于 STS 数据集的各种性能度量指标评分对比

Table 3 Comparison of various performance metrics based on STS data set

性能度量指标	得分
BLEU	0.4782
ROUGE	0.5349
METEOR	0.4814
语义相似度	0.4487
人工评价	0.4463

从表 3 可以看出,与人工评价分数最为接近的性能度量指标是基于词向量模型的语义相似度评价方式,而基于精确率与召回率的性能度量指标得到的分数整体偏高,这是由于 STS 数据集中存在很多结构相似但语义差距较大的句子对。对于这些“陷阱”,只考虑单词间的精确率与召回率而不考察深层语义关系的性能度量指标就会给出较高的评价结果,如 BLEU,ROUGE 与 METEOR。

为了更加直观地反映各度量指标的特点,我们从 STS 数据集中随机选取了 10 组数据,并给出这 10 组数据在不同度量指标下的得分,结果如表 4 所列。以测试数据第二组句子为例,在语义相似的情况下,BLEU,ROUGE 与 METEOR 给出的评价结果虽然不及语义相似度更接近人工评价结果,但是差距并不明显。而第七组句子中,两个句子语义相差很大,这种场景下 BLEU,ROUGE 与 METEOR 给出的分数却高于人工评价与语义相似度。这是由于这些指标陷入了上文提到的“陷阱”中,导致评分过高,不能正常评价语义关系。综上,在设计语义通信系统时,需要根据具体情况来选择性能度量指标,在语义关系较为简单时可以考虑使用 METEOR 等指标来降低评价复杂度。而在语义关系复杂的情况下,需要选择语义相似度指标来挖掘更深层次的高级语义。

表 4 STS 数据集中 10 组随机数据各种性能度量指标评分对比

Table 4 Comparison of the scores of various performance metrics for ten sets of random data in STS data set

序号	测试数据	BLEU	ROUGE	METEOR	语义相似度	人工评价
1	A plane is taking off. An air plane is taking off.	0.6667	0.8000	0.7782	0.9800	1.0000
2	The man hit the other man with a stick. The man spanked the other man with a stick.	0.8889	0.8750	0.8819	0.9172	0.8400
3	A man pours oil into a pot. A man pours wine in a pot.	0.7143	0.7143	0.6914	0.6233	0.6400
4	A woman is frying fish. A woman is cooking fish.	0.8000	0.8000	0.7500	0.8000	0.8000
5	A woman is peeling a potato. A woman is peeling an apple.	0.6667	0.6667	0.7500	0.3500	0.4000
6	A man is fishing. A man is exercising.	0.7500	0.6667	0.7361	0.0500	0.1000
7	A cat is playing a piano. A man is playing a guitar.	0.6667	0.6667	0.6250	0.0758	0.1200
8	Fish are swimming. A fish is swimming.	0.2500	0.3333	0.3226	0.9023	0.8266
9	A man is mowing a lawn. A woman is cutting a lemon.	0.5000	0.5000	0.2500	0.0100	0.0400
10	A man is lifting a car. A man is climbing a wall.	0.6667	0.6667	0.6250	0.3264	0.2780

8 未来发展方向

未来，语义通信将使用人工智能作为支撑内生技术，并结合现代通信技术来推动通信领域的快速发展。本文认为语义通信系统的性能度量指标在以下几个方面有着巨大的发展潜力。

(1)基于任务导向的可灵活调整的语义评价指标：语义通信技术的发展将推动其在不同任务场景下的应用，这就要求语义通信性能度量指标可以根据任务导向灵活调整语义评估策略，提高性能度量指标在特定任务场景下的针对性和准确性。

(2)基于深层语义词向量模型的评价指标：随着人工智能的发展，词向量训练技术也会不断更迭。深层语义特征可以压缩语义，减少传输资源的消耗，同时提高语义通信的准确性，词向量模型必然朝着提取深层语义特征的方向发展。

(3)主观、客观评价相结合的语义评价指标：BLEU，ROUGE 与 METEOR 这种类脑的主观评价方法符合人脑认知的语言规律和逻辑性评价，但是灵活性较差；另一方面，语义相似度这种基于大数据样本的客观评价方法，处理方式灵活，但逻辑性和解释性较差，导致部分语言规律丢失。综上，为了使评价结果更接近人工评估，需要综合主观评价与客观评价的优势，进一步研究主观评价、客观评价相结合的语义评价指标。

(4)语义通信系统性能度量指标的优化选择：在语义通信任务中，性能度量指标的选择尤为关键。不同通信任务的侧重点并不相同，信息之间的语义偏差程度也不同。为了保证语义通信系统性能评价更为合理，需要深入研究语义通信系统性能度量指标的优化选择问题。

结束语 本文针对语义通信系统，介绍了 4 类语义通信中常用的性能度量指标，并对这 4 种性能度量指标进行了详细的分析，讨论了这些性能度量指标的优缺点和适用性。分析结果显示，不同的性能度量指标有着不同的适用范围，在简单语义场景下可以采用 BLEU 指标或 ROUGE 指标，来实现评价复杂度和传输性能的良好折衷；而复杂语义场景下需要采用 METEOR 指标或语义相似度指标来挖掘句子的深层语义。

参 考 文 献

[1] SHI G M, LI Y Y, XIE X M. Semantic communications: outcome of the intelligence era[J]. Pattern Recognition and Artificial Intelligence, 2018, 30(3): 289-299.

[2] SHANNON C. A mathematical theory of communication[J]. Bell System Technical Journal, 1948, 27(3): 379-423.

[3] BAO J, BASU P, DEAN M, et al. Towards a theory of semantic communication[C]// 2011 IEEE Network Science Workshop. IEEE, 2011: 110-117.

[4] GÜLER B, YENER A, SWAMI A. The semantic communication game[J]. IEEE Transactions on Cognitive Communications and Networking, 2018, 4(4): 787-802.

[5] WENG Z Z, QIN Z J, LI G. Semantic communications for speech signals[C]// ICC 2021 - IEEE International Conference on Communications. IEEE, 2021: 1-6.

[6] JUBA B, SUDAN M. Universal semantic communication I[C]// Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing. 2008: 123-132.

[7] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: a method for automatic evaluation of machine translation[C]// 40th Annual Meeting of the Association for Computational Linguistics. 2002: 311-318.

[8] DODDINGTON G. Automatic evaluation of machine translation quality using n-gram cooccurrence statistics[C]// Proceedings of the 2nd International Conference on Human Language Technology Research. 2002: 138-145.

[9] LIN C Y. Rouge: a package for automatic evaluation of summaries[C]// Proceedings of the Workshop on Text Summarization Branches Out. 2004: 74-81.

[10] 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 Ex-

trinsic Evaluation Measures for Machine Translation and/or Summarization. 2005:65-72.

[11] AGIRRE E,BANEA C,CER D,et al. SemEval-2016 Task 1:semantic textual similarity,monolingual and cross-lingual evaluation[C]// Proceedings of SemEval-2016. 2016:497-511.

[12] VEDANTAM R,ZITNICK C,PARIKH D. Cider:consensus-based image description evaluation[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition. IEEE,2015:4566-4575.

[13] GOMAA W,FAHMY A. A survey of text similarity approaches [J]. International Journal of Computer Applications,2013,68(13):13-17.

[14] HOVY E. Toward finely differentiated evaluation metrics for machine translation[C]// Proceedings of the EAGLES Workshop on Standards and Evaluation. 1999:127-133.

[15] POPOVIC M. CHRF:character n-gram F-score for automatic MT evaluation[C]// Proceedings of the Tenth Workshop on Statistical Machine Translation. 2015:392-395.

[16] MATHUR N,BALDWIN T,COHN T. Tangled up in BLEU:reevaluating the evaluation of automatic machine translation evaluation metrics[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020:4984-4997.

[17] REITER E. A structured review of the validity of BLEU[J]. Computational Linguistics,2018,44(3):393-401.

[18] XIE H Q,QIN Z J. A lite distributed semantic communication system for Internet of Things[J]. IEEE Journal on Selected Areas in Communications,2021,39(1):142-153.

[19] SMITH A,HARDMEIER C,TIEDEMANN J. Climbing Mont BLEU:the strange world of reachable High-BLEU translations [C]// Proceedings of the 19th Annual Conference of the European Association for Machine Translation. 2016:269-281.

[20] STENT A,MARGE M,SINGHAI M. Evaluating evaluation methods for generation in the presence of variation[C]// Computational Linguistics and Intelligent Text Processing. 2005:341-351.

[21] WHITE J,CONNELL T O,MARA F O. The ARPA MT evaluation methodologies:evolution,lessons,and future approaches [C]// Proceedings of the First Conference of the Association for Machine Translation in the Americas. 1994:193-205.

[22] WISEMAN S,SHIEBER S,RUSH A. Challenges in data-to-document generation[C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017:2253-2263.

[23] LAVIE A,SAGAE K,JAYARAMAN S. The significance of recall in automatic metrics for MT evaluation[C]// 6th Conference of the Association for Machine Translation in the Americas. 2004:134-143.

[24] ANDERSON P,FERNANDO B,JOHNSON M,et al. SPICE:semantic propositional image caption evaluation[C]// 14th European Conference on Computer Vision. 2016:382-398.

[25] BENGIO Y,DUCHARME R,VINCENT P,et al. A neural probabilistic language model[J]. Journal of Machine Learning Research 3,2003,3(6):1137-1155.

[26] MAJUMDER G,PAKRAY P,GELBUKH A,et al. Semantic textual similarity methods,tools,and applications:a survey[J]. Computaciony Sistemas,2016,20(4):647-665.

[27] MIKOLOV T,CHEN K,CORRADO G,et al. Efficient estimation of word representations in vector space[J]. arXiv:1301.3781,2013.

[28] WANG Q,XU J G,CHEN H,et al. Two improved continuous bag-of-word models[C]// 2017 International Joint Conference on Neural Networks. IEEE,2017:2851-2856.

[29] PENNINGTON J,SOCHER R,MANNING C. Glove:global vectors for word representation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. 2014:1532-1543.

[30] DEVLIN J,CHANG M,LEE K,et al. Bert:pre-training of deep bidirectional transformers for language understanding[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. 2019:4171-4186.

[31] JIN X L,ZHANG S W,LIU J. Word semantic similarity calculation based on Word2vec[C]// 2018 International Conference on Control,Automation and Information Sciences. IEEE,2018:12-16.

[32] HUANG P S,HE X D,GAO J F,et al. Learning deep structured semantic models for web search using clickthrough data[C]// Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. 2013:2333-2338.

[33] SHEN Y L,HE X D,GAO J F,et al. A Latent Semantic Model with Convolutional-Pooling Structure for Information Retrieval [C]// Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. 2014:101-110.

[34] CER D,DIAB M,AGIRRE E,et al. SemEval-2017 Task 1:semantic textual similarity multilingual and crosslingual focused evaluation[C]// Proceedings of the 11th International Workshop on Semantic Evaluation. 2017:1-14.

[35] WANG M,HU F H. The application of NLTK library for python natural language processing in corpus research[J]. Theory and Practice in Language Studies,2021,11(9):1041-1049.

[36] BIRD S. NLTK:the natural language toolkit[C]// Proceedings of the COLING/ACL on Interactive Presentation Sessions. 2006:69-72.

JIANG Sheng-teng, born in 1998, postgraduate. His main research interests include semantic communication and so on.

CAO Kuo, born in 1990, Ph.D, lecturer. His main research interests include cooperative communications and physical layer security of wireless communications.

(责任编辑:何杨)