

语义信息论的回顾与展望



Semantic Information Theory: Past and Future

辛港涛/XIN Gangtao, 樊平毅/FAN Pingyi

(清华大学, 中国 北京 100084)

(Tsinghua University, Beijing 100084, China)

DOI: 10.12142/ZTETJ.202302003

网络出版地址: <https://kns.cnki.net/kcms/detail/34.1228.TN.20230411.1249.004.html>

网络出版日期: 2023-04-11

收稿日期: 2023-01-27

摘要: 近年来, 语义通信受到业界的广泛关注, 但一套完整且有效的语义信息论框架仍未完全建立。将语义信息论分为语义熵、语义率失真、语义信道容量3个部分, 分别探讨语义信息的度量问题、语义编码与失真问题、最大语义通信量问题。同时, 从信息理论角度出发, 系统回顾和梳理了语义信息论的相关发展, 对未来语义信息论发展中可能存在的开放性问题进行了讨论。认为目前语义通信的发展处于初级阶段, 仍然有很多问题尚未解决。

关键词: 语义信息论; 语义通信; 信息论; 6G

Abstract: In recent years, semantic communications has received wide attention from academia and industry, but a complete and effective framework of semantic information theory has not been fully established. The semantic information theory is divided into three parts: semantic entropy, semantic rate-distortion, and semantic channel capacity, and the measurement of semantic information, semantic coding and distortion, and the maximum semantic traffic are then discussed. At the same time, the related work in semantic information theory is sorted out, and some open problems with high probabilities about the future of semantic information theory are discussed. It is believed that the development of semantic communications is in its initial stage, and there are still many unsolved problems.

Keywords: semantic information theory; semantic communication; information theory; 6G

近年来, 随着规模化无线通信的快速发展与智能处理需求的快速增多, 各种基于无线通信技术的新兴智能业务得到迅速发展, 这给通信技术带来了新的挑战。一方面, 这些新兴业务(如工业互联网、虚拟/增强/混合现实、元宇宙、全息通信)的成功高度依赖于在海量数据集上对大型学习模型的训练。此类应用承载的大量流量可能会使现存的网络容量趋于饱和。因此, 通信基础设施需要融入智能, 确保在必要的时间, 以合理的速率, 传输所需的流量。另一方面, 这些新兴业务需要极低的端到端时延, 因此通信技术需要考虑流量的相关性和紧迫性, 以最快、最可靠的方式提取和交付与任务相关的信息。这将导致通信网络架构从单纯追求高速率传输向追求智能化过渡^[1]。

语义通信是一种全新的通信架构, 它将用户对信息的需求及任务语义融入通信过程, 有望大幅度提高通信实施效率, 提升用户的体验, 并从根本上解决基于数据的传统通信协议中存在的跨系统、跨协议、跨网络等技术难题^[2]。C. E. SHANNON 和 W. WEAVER 在 1949 年发表的论文中提出了广义通信的 3 个层次的问题^[3], 分别为:

- 1) 技术层: 传输的符号是否准确?
- 2) 语义层: 传输的符号是否准确表达了任务预期含义?
- 3) 有效层: 接收的信息含义能否按照预期方式影响用户的行为?

从技术层到语义层, 通信的目标从符号的准确传输转变为语义的有效交换。经典的香农信息论建立在概率与统计的基础上, 忽略了具体任务的语义表示方式, 从技术层解决了两个基本问题: 临界数据压缩的值和临界通信传输速率的值。语义通信的数学理论和语义的数学表征可以归结为语义信息论的问题。尽管目前语义信息论没有公认且统一的理论框架, 但近些年来, 学术界和工业界对语义信息论的研究越来越多。2021 年, 华为公司提出的后香农时代信息通信技术 (ICT) 领域的十大挑战问题^[4], 将语义信息论列为基础理论的首要问题。本文中, 我们关注语义信息论的核心概念, 介绍语义熵、语义率失真和语义信道容量方面的进展。

1 语义熵

语义信息的度量问题是语义信息理论的基础。熵是对随机变量不确定性的度量, 语义熵是对语义不确定度或信息量大小的度量。到目前为止, 对语义熵直观且通用的数学描述

基金项目: 国家重点研发计划 [2021YFA1000500 (4)]

仍然是一个具有挑战性的难题：一方面语义基本内涵不易定义从而难以度量；另一方面人们对语义的产生机理和过程并不明确^[5-7]。

1) 语义熵起源于前香农时代，与自然语言的分析相关^[8]。针对语言任务，1952年，R. CARNAP和Y. BAR-HILLEL^[9]提出了语义熵的概念，用逻辑概率来衡量一个句子所包含的语义信息量，即：

$$H_s(e) = -\log m(e), \quad (1)$$

其中， $m(e)$ 是指事件 e 在所有可能情况下为真的概率，也就是事件 e 的逻辑概率。可以看出，一个句子为真的逻辑概率越大，包含的信息量就越小。但是，这产生了一个悖论，即任何事实与其自身矛盾时会具有无穷大的信息量，如 $H_s(e \wedge \neg e)$ （表示“事件 e 与事件非 e ”发生的概率）为无穷大。2004年，L. FLORIDI^[10]提出强语义信息理论，用与真实事件之间的距离来表示语义信息量的大小，解决了这一悖论。2011年，S. D'AFLOSIO^[11]基于真理相似性对语义信息进行了定量描述。L. FLORIDI和S. D'AFLOSIO的方法衡量的是某一事件相对于参考事件的信息，信息值总是在 $0 \sim 1$ 。然而，这些度量都依赖于参考事件的存在。从本质上讲，他们的工作提供了两个句子之间语义相似度的度量，而不是语义不确定性或信息量大小的测量。2011年，遵循R. CARNAP的定义，J. BAO、P. BASU等^[12]采用命题逻辑拓展了 $m(e)$ 的表示，进一步拓展了语言任务中语义熵的含义。

2) 针对智能任务，I. D. MELAMED^[13]提出了一种测量文本中单词的语义熵的方法。具体而言，对于一个词 w ，语义熵可以被表示为：

$$H_s(w) = H(Tlw) + N(w) = -\sum_{t \in T} p(tlw) \log p(tlw) + p(NULL|w) \log F(w), \quad (2)$$

其中， $H(Tlw)$ 代表了翻译不一致性，表示一个词被翻译成不同的方式所带来的不确定性， T 表示目标词的集合， $N(w)$ 表示 w 的空链接的贡献，表示从一种语言翻译到另一种语言时遇到困难的可能性， $F(w)$ 是 w 的频率。对于分类任务，LIU X. D.、JIA W. J.等^[14]引入匹配度和隶属度的概念来定义语义熵。隶属度是模糊集理论中的一个难以解析的概念，一般是根据经验来给定。定义 s 为语义概念， $\mu_s(x)$ 为每个 $x \in X$ 的隶属度。对于某个类别 C_j ，匹配度 $D_j(s)$ 定义为：

$$D_j(s) = \frac{\sum_{x \in X_{C_j}} \mu_s(x)}{\sum_{x \in X} \mu_s(x)}. \quad (3)$$

对于类别 C_j ，其语义熵定义为 $H_s(s) = -D_j(s) \log D_j(s)$ 。将 X 中所有类的语义熵相加，可以得到集合 X 上的整体语义熵。

A. CHATTOPADHYAY等^[15]提出一种量化任务相关的语义熵，他们将语义熵定义为关于数据 X 的语义查询的最小数量，其答案足以预测任务 V ，可以表示为：

$$H_s(X; V) \triangleq \min_E E_s \left[\left| \text{Code}_Q^E(X) \right| \right],$$

$$\text{s.t. } P(v | \text{Code}_Q^E(x)) = P(v|x), \forall x, v, \quad (4)$$

其中， $\text{Code}_Q^E(X)$ 表示使用语义编码器 E 从 x 中提取的查询向量。

3) 语义是对信号的理解，知识在语义的产生过程中充当了语义编码和表示的角色，发挥了重要作用。J. CHOI等^[16]基于逻辑概率，从知识库的角度研究一个句子的语义熵。具体而言，记知识库为 K ，用 $p[K \vdash q]$ 表示句子 q 相对于知识库 K 正确的概率，简化为 $p_q = p[K \vdash q]$ ，则 q 相对于 K 的语义熵为：

$$H_s(q) = -p_q \log p_q + (1 - p_q) \log (1 - p_q). \quad (5)$$

众所周知，扩展模式可以把简单的事物组合成复杂的系统，甚至产生智能。在人类的语言系统中，主谓宾、定状补的成分组成了句子，可以表达出单个单词所不可能具有的含义。受此启发，XIN G. T.和FAN P. Y.^[17]认为智能语义应当是可以扩展的，语义熵的表示形式要体现知识扩展的理念。在语义的扩展中，知识层面发生了碰撞。语义正是在知识碰撞作用下，与信号作用的产物。例如，Apple Inc.属于商业公司类别，thirteen属于数字类别，但是它们的碰撞可能会产生一个单词——iPhone，属于移动通信产品类别。用 X_1 和 X_2 表示信号， K_A^1 和 K_A^2 表示两个知识实例，则语义熵可以表示为 $H_s(X_1 \oplus X_2, K_A^1 \odot K_A^2)$ ，其中 $\oplus$ 表示扩展， $\odot$ 表示碰撞。

4) 不同于逻辑概率或统计概率模型，针对不同任务和背景，语义熵的推导形式也各不相同。A. KOLCHINSKY与D. H. WOLPERT^[18]将语义信息定义为描述系统与其所处环境之间关系的语法信息。M. KOUNTOURIS和N. PAPPAS^[19]使用 $R \acute{e}nyi$ 熵对语义信息进行度量。N. J. VENHUIZEN等^[20]基于背景知识的语言理解模型推导出语义熵。C. LU^[21]引入广义信息论，基于贝叶斯公式、逻辑概率和模糊集合等理论对语义信息进行数学度量描述。

2 语义率失真

在香农信息论中，率失真理论解决了失真情况下的编码

问题。若信源为 $X \sim p(x)$, 编码后的表示为 $\hat{X} \sim p(\hat{x})$, 率失真度量为 $d(x, \hat{x})$, 则率失真函数为:

$$R(D) = \min_{p(\hat{x}|x)} I(X; \hat{X}). \quad (6)$$

如果 $R > R(D)$, 则存在编码方法使得平均失真趋向于 D 。如果 $R < R(D)$, 则这样的码序列不存在。

在语义通信中, 在语义层面进行信息提取与编码表示, 有助于进一步压缩语义信息的冗余, 提取最重要的语义特征, 从而提高语义传输的有效性。J. LIU 等^[22]将语义率失真定义为:

$$\begin{aligned} R(D_s, D_a) &= \min I(X; \hat{S}, \hat{X}), \\ \text{s.t. } E\hat{d}_s(X, \hat{S}) &\leq D_s, E\hat{d}_a(X, \hat{X}) \leq D_a, \end{aligned} \quad (7)$$

其中, S 和 $\hat{S}$ 为发送方和接收方对语义的理解, X 和 $\hat{X}$ 则分别是语义表示。 D_s 表示语义失真, D_a 为表示信号失真, 分别代表语义编码和信道噪声所造成的失真。通过上述简化, 这个问题变为带有两个失真度量的香农率失真问题。基于此, GUO T. 等^[23]从边信息的角度, 对两个用户的语义率失真做了进一步的推广, 可表示为:

$$R(D_1, D_2, D_s) = \min I(X_1, X_2; \hat{X}_1, \hat{X}_2, \hat{S}|Y), \quad (8)$$

其中, X_1 、 X_2 分别是两个用户的语义表示, Y 是边信息。2022年, P. A. STAVROU 和 M. KOUNTOURIS^[24]进一步研究了在汉明失真度量下的特征。

语义率失真理论直接对应编码技术, 有效的语义编码工作具有高复杂性和挑战性。目前, 用双失真度量的方式来表示语义编码是较普遍的做法。Y. SHAO 等^[25]用语义失真 L_U 和语义成本 D_U 来定义语义编码的可达区域。语义失真反映了接收方和发送方在语义理解上的差异, 语义成本则代表了信息的简洁性或可理解性, 它可以是相应比特序列的长度。语义成本定义可表述为: 若存在一个语义编码方案 U 使得 $D_U = D$, $L_U = L$, 则称失真-成本区域 (L, D) 是可达的。P. AGHELI^[26]等多用户的语义编码进行了研究, 发送端向具有两个不同目标的接收端传输语义, 用户根据最大化两个接收端的语义感知效用函数的加权和, 分配所选数据包到达的最佳实际码字长度。Y. XIAO 等^[27]提出了策略语义通信的率失真理论。该将博弈论模型与率失真理论相结合, 以表征语义编码器和解码器之间的信息交互对通信失真性能的影响。

3 语义信道容量

信道容量是信息论的中心问题, 也是信息论中最著名的成果。对于一个输入为 X 、输出为 Y 的离散无记忆信道, 信

道容量被定义为:

$$C = \max_{p(x)} I(X; Y). \quad (9)$$

信道容量指出了在传输信息时可达速率的上界。所有小于信道容量 C 的码率都可在无差错条件下可达, 而所有大于信道容量的码率是不可无差错达到的。也就是说, 对任意的码率 $R < C$, 存在误差概率趋于 0 的一个 $(2^{nR}, n)$ 码序列; 反之, 如果码率 $R > C$, 那么误差概率将远离 0。

语义信道容量的相关工作较少。T. OKAMOTO^[28]认为语义信道容量可以表示为通过语义信道传输的语义信息的最大速率或者最大语义通信量与通信数据大小的比率。一方面, 在该速率下, 可以通过语义信道发送语义信息, 并在语义信道的输出端恢复信息; 另一方面, 在信道上的语义数据信息传输速率不可能超过信道的语义容量。

2011, J. BAO 等^[12]对语义信道容量做出如下定义:

$$C_s = \sup_{P(Z|X)} \{I(X; V) - H(Z|X) + \overline{H}_s(V)\}. \quad (10)$$

对任意 $\epsilon > 0$ 和 $R < C_s$, 存在一种编码策略, 使得语义错误的最大概率小于 ϵ 。其中, X 是信源, V 是传输任务, Z 是语义表示, $I(X; V)$ 是信源 X 和传输任务 V 之间的互信息, $H(Z|X)$ 表示编码的语义不确定度, $\overline{H}_s(V)$ 是任务 V 的接收消息的平均逻辑信息。

语义信道容量可以高于或低于香农信道容量 $I(X; V)$, 这取决于语义编码策略和接收机解释所收到消息的能力。若 $\overline{H}_s(V) - H(Z|X) > 0$, 则意味着接收者可以处理语义歧义, 接收者可以用收到的语句回答问题, 语义压缩可以实现更高的传输速率。在这种情况下, C_s 高于香农容量。若 $\overline{H}_s(V) - H(Z|X) < 0$, 接收者不能解决原本存在的语义歧义, 接收者无法根据收到的语句回答问题, C_s 低于香农容量。

4 语义信息论展望

如果说目前语义通信处于幼儿期, 那么语义信息论则处于婴儿期, 仍然有大量开放性问题尚未解决, 例如:

1) 语义、语义熵、语义编码器是否存在通用的数学表示形式? 还是可以分类表示?

2) 语义通信的基本性能极限是什么? 由于语义通信旨在让传输的符号精确地传达内容含义, 因此语义通信的基本性能极限难以用一个具体的容量值描述, 应由技术层和语义层所共同决定, 其形式应体现多维特点。涉及的极限问题包括但不限于网络容量、最优编码、数据传输速率对语义传输速率的影响等, 可能还需要兼顾物理层通信速率与语义解码

器表达能力之间的均衡特征。

3) 基于深度学习的语义通信系统的理论框架应如何发展? 目前深度学习被广泛用于信源信道联合编码和面向任务的通信结构设计, 在智能处理服务中逐渐显现其重要性。探索基于深度学习的语义通信的理论框架并给出理论指导, 包括其合适的性能指标与语义相似度度量等是非常必要的。

5 结束语

语义通信作为一种新的通信结构, 改变了传统的数据传输范式, 有望为大型智能处理服务带来新气象。然而, 目前的语义通信研究刚刚起步, 有大量的工作值得进一步研究, 如知识共享框架、资源调度与分配、语义噪声建模与度量等。本文从语义通信的理论角度出发, 依次回顾了语义熵、语义率失真和语义信道容量等相关概念, 梳理了语义信息论的相关发展, 同时对未来语义信息论存在的开放性问题进行了简单讨论。希望这些问题能引起专家学者的关注, 助力他们为语义信息论的发展做出更多的贡献。

参考文献

- [1] GÜNDÜZ D, QIN Z J, AGUERRI I E, et al. Beyond transmitting bits: context, semantics, and task-oriented communications [J]. IEEE journal on selected areas in communications, 2023, 41(1): 5–41. DOI: 10.1109/JSAC.2022.3223408
- [2] 石光明, 肖泳, 李莹玉, 等. 面向万物智联的语义通信网络 [J]. 物联网学报, 2021, 5(2): 26–36. DOI: 10.11959/j.issn.2096–3750.2021.00209
- [3] SHANNON C E, WEAVER W, BURKS A. The mathematical theory of communication [J]. The bell system technical journal, 1948: 379–428
- [4] 徐文伟, 张弓, 白铂, 等. 后香农时代 ICT 领域的十大挑战问题 [J]. 中国科学: 数学, 2021, 51(7): 1095–1138. DOI: 10.1360/SSM–2021–0013
- [5] 石光明, 高大化, 杨曼曦, 等. 信号的语义刻画与度量 [J]. 电子学报, 2022, 50(9): 2068–2078. DOI: 10.12263/DZXB.20210238
- [6] 刘传宏, 郭彩丽, 杨洋, 等. 面向智能任务的语义通信: 理论、技术和挑战 [J]. 通信学报, 2022, 43(6): 41–57. DOI: 10.11959/j.issn.1000–436x.2022117
- [7] QIN Z J, TAO X M, LU J H, et al. Semantic communications: principles and challenges [EB/OL]. [2023–01–22]. <https://arxiv.org/abs/2201.01389>
- [8] ZHANG P, XU W, GAO H, et al. Toward wisdom-evolutionary and primitive-concise 6G: a new paradigm of semantic communication networks [J]. Engineering, 2022(8): 60–73. DOI: 10.1016/j.eng.2021.11.003
- [9] CARNAP R, BAR-HILLEL Y. An outline of a theory of semantic information [EB/OL]. (1952–10–27) [2023–01–12]. <https://dspace.mit.edu/handle/1721.1/4821>
- [10] FLORIDI L. Outline of a theory of strongly semantic information [J]. Minds and machines, 2004, 14(2): 197–221. DOI: 10.1023/B: MIND.0000021684.50925.c9
- [11] D'ALFONSO S. On quantifying semantic information [J]. Information, 2011, 2(1): 61–101. DOI: 10.3390/info2010061
- [12] BAO J, BASU P, DEAN M K, et al. Towards a theory of semantic communication [C]//Proceedings of 2011 IEEE Network Science Workshop. IEEE, 2011: 110–117. DOI: 10.1109/NSW.2011.6004632
- [13] MELAMED I D. Measuring semantic entropy [EB/OL]. [2022–12–28]. <https://www.semanticscholar.org/paper/Measuring-Semantic-Entropy-Melamed/4e107ca440ad32ccf279545a3b44228af587da42>
- [14] LIU X D, JIA W J, LIU W Q, et al. AFSSE: an interpretable classifier with axiomatic fuzzy set and semantic entropy [J]. IEEE transactions on fuzzy systems, 2020, 28(11): 2825–2840. DOI: 10.1109/TFUZZ.2019.2945239
- [15] CHATTOPADHYAY A, HAEFFELE B D, GEMAN D, et al. Quantifying task complexity through generalized information measures [EB/OL]. (2021–01–01) [2022–01–20]. <https://paperswithcode.com/paper/quantifying-task-complexity-through-generalized-information-measures>

- complexity-through
- [16] CHOI J, LOKE S W, PARK J. A unified view on semantic information and communication: a probabilistic logic approach [C]//Proceedings of 2022 IEEE International Conference on Communications Workshops (ICC Workshops). IEEE, 2022: 705–710. DOI: 10.1109/ICCWorkshops53468.2022.9814642
 - [17] XIN G T, FAN P Y. EXK-SC: a semantic communication model based on information framework expansion and knowledge collision [J]. Entropy (Basel, Switzerland), 2022, 24(12): 1842. DOI: 10.3390/e24121842
 - [18] KOLCHINSKY A, WOLPERT D H. Semantic information, autonomous agency and non-equilibrium statistical physics [J]. Interface focus, 2018, 8(6): 41. DOI: 10.1098/rsfs.2018.0041
 - [19] KOUNTOURIS M, PAPPAS N. Semantics-empowered communication for networked intelligent systems [J]. IEEE communications magazine, 2021, 59(6): 96–102. DOI: 10.1109/MCOM.001.2000604
 - [20] VENHUIZEN N J, CROCKER M W, BROUWER H. Semantic entropy in language comprehension [J]. Entropy, 2019, 21(12): 1159. DOI: 10.3390/e21121159
 - [21] LU C. From Bayesian inference to logical Bayesian inference: a new mathematical frame for semantic communication and machine learning [EB/OL]. [2023–01–26]. https://link.springer.com/chapter/10.1007/978-3-030-01313-4_2
 - [22] LIU J K, ZHANG W Y, POOR H V. A rate-distortion framework for characterizing semantic information [C]//Proceedings of 2021 IEEE International Symposium on Information Theory (ISIT). IEEE, 2021: 2894–2899. DOI: 10.1109/ISIT45174.2021.9518240
 - [23] GUO T, WANG Y, HAN J, et al. Semantic compression with side information: a rate-distortion perspective [EB/OL]. [2023–01–25]. <https://arxiv.org/abs/2208.06094>
 - [24] STAVROU P A, KOUNTOURIS M. A rate distortion approach to goal-oriented communication [C]//Proceedings of 2022 IEEE International Symposium on Information Theory (ISIT). IEEE, 2022: 590–595. DOI: 10.1109/ISIT50566.2022.9834593
 - [25] SHAO Y, CAO Q, GUNDUZ D. A theory of semantic communication [EB/OL]. [2023–01–26]. <https://arxiv.org/abs/2212.01485>
 - [26] AGHELI P, PAPPAS N, KOUNTOURIS M. Semantic source coding for two users with heterogeneous goals [C]//Proceedings of GLOBECOM 2022 – 2022 IEEE Global Communications Conference. IEEE, 2023: 4983–4988. DOI: 10.1109/GLOBECOM48099.2022.10001230
 - [27] XIAO Y, ZHANG X, LI Y Y, et al. Rate-distortion theory for strategic semantic communication [C]//Proceedings of 2022 IEEE Information Theory Workshop (ITW). IEEE, 2022: 279–284. DOI: 10.1109/ITW54588.2022.9965825
 - [28] OKAMOTO T. A unified paradigm of organized complexity and semantic information theory [EB/OL]. [2023–01–26]. <https://arxiv.org/abs/1608.00941>

作者简介



辛港涛, 清华大学在读博士研究生; 主要研究领域为信息论、语义通信、图像编码。



樊平毅, 清华大学教授; 主要研究领域为 6G 无线通信网络与机器学习、语义信息论与广义信息论、大数据处理理论以及智能网络与系统检测; 承担多个国家“973”“863”项目、移动专项、重点研发计划、国家自然科学基金和国际合作项目; 已发表论文 200 余篇, 出版学术著作 4 部, 申请国家发明专利 30 余项、国际专利 5 项。