基于熵函数的语义信息边界研究

Research on Semantic Information Boundary Based on Entropy

唐雪/TANG Xue[1,2]，许进/XU Jin[1,2]，冯雨龙/FENG Yulong[1,2]
（1. 移动网络与移动多媒体技术全国重点实验室，中国 深圳 518055；
2. 中兴通讯股份有限公司，中国 深圳 518057）
(1. China State Key Laboratory of Mobile Network and Mobile Multimedia Technology, Shenzhen 518055, China;
2. ZTE Corporation, Shenzhen 518057, China)

DOI：10.12142/ZTETJ.202506009
网络出版地址：https://link.cnki.net/urlid/34.1228.TN.20251219.1012.008
网络出版日期：2025-12-19
收稿日期：2025-05-15

摘要：人工智能的不断发展助力语义通信不断成熟，语义通信越来越被证明是未来“智能体”之间通信的新范式，然而语义度量理论发展缓慢且极不成熟。提出了语义度量的4条基本假设，试图将经典信息理论与强语义信息理论等归入统一的语义度量理论中，实现概率视角下语义度量方式与真性距离视角下语义度量方式的有机结合。在梳理语义度量理论发展历程的同时，试图将其中的要素融汇贯通，进一步给出相关概念，并基于熵函数对语义信息的边界进行探索。

关键词：Shannon极限；语义通信；人工智能；语义度量

Abstract: The ongoing advancement of artificial intelligence is facilitating the maturation of semantic communication, steadily establishing itself as a new communication paradigm for future "intelligent agents". However, the development of theories for semantic measurement has been sluggish and remains underdeveloped. Four basic assumptions of semantic measurement are proposed, aiming to incorporate classical information theory and strong semantic information theory into a unified semantic measurement theory, and to realize the organic combination of semantic measurement methods from the probability perspective and those from the truth distance perspective. This paper attempts to explore the boundary of semantic information based on the entropy function, while combing through the development of semantic measurement theory and trying to integrate its elements.

Keywords: Shannon limit; semantic communication; artificial intelligence; semantic measurement

引用格式：唐雪, 许进, 冯雨龙. 基于熵函数的语义信息边界研究 [J]. 中兴通讯技术, 2025, 31(6): 61-69. DOI: 10.12142/ZTETJ.202506009
Citation: TANG X, XU J, FENG Y L. Research on semantic information boundary based on entropy [J]. ZTE technology journal, 2025, 31(6): 61-69. DOI: 10.12142/ZTETJ.202506009

1948年，香农发表“A Mathematical Theory of Communication”一文[1]，高度概括了人类对于通信的认识，指出数字化与编码是设计高效且可靠通信系统的必由之路。基于此，通信行业经历了几十年的繁荣发展。直到今天，经典通信的实现基本逼近香农理论极限，通信技术的发展速度开始迟缓，甚至不再聚焦于通信“本身”。通信领域的研究者提出的一部分新技术开始在复杂度、有效性和可靠性之间相互折中，而另外一些新技术则依赖于材料、器件等方向的突破。就像19世纪的物理学一样，可以说经典通信的大厦已经建成，但其发展进程中仍存在两个关键的待解难题。

第一个难题源于信息论本身的假设条件。这一假设条件使其在实际通信系统搭建中存在很大的局限性，不能面面俱到地刻画人类的实际通信问题。原因有两点：一方面人是智慧生物，其交流的信息往往是非平稳的，非各态历经的，而概率论中没有有效的工具来刻画这类问题；另一方面，人类信息还具有模糊性，比如文字的“韵味”或乐谱的“旋律”。其实，在香农的论文出版之后，当时的通信学家们就注意到了这个问题，并被WEAVER总结为通信的三层问题：语法问题、语义问题和语用问题[2]。

第二个难题源于人工智能（AI）带来的不确定性。自人工神经网络（ANN）展现出强大的非线性拟合能力以来，尤其是近期大模型对自然语言建模的能力，这一特性更为凸显。ANN似乎为原本概率论中无法解决的问题带来了新的思路，即采用ANN近似逼近[3]。有人用流形理论解释ANN的这种非线性拟合能力。对于一个圆，直接在一维坐标上表示是一群点的集合，但是在二维坐标上可以用半径来表示，将其投影回一维坐标，则为一个点，此时这个圆就是二维空

基金项目：国家重点研发计划项目（2020YFB1807202）

间中的一维流形。通信数据无论是文本、图片、视频，大都是从人的角度，即三维空间建模的，但是ANN具有从更高维和更多维建模的能力，这就使得现实数据可以借助ANN在高维进一步表示、压缩和传输。而流形理论中的高维建模能力，其表现之一就是强大的非线性拟合能力。

随着5G技术的逐渐落地应用，国际电信联盟（ITU）发布了6G愿景示意图[4]。未来6G强调数据处理的内容在不断增多，比如沉浸式通信、通感一体化等。相比于前几代通信，6G愿景强调通信方面的内容在减少，唯一的新变化是AI与通信的融合。而语义通信起源于经典通信中的联合编码。联合编码则是最早试图采用ANN代替或增强经典通信系统的研究之一。因此，语义通信不仅与最早的通信研究有关，而且符合未来6G所重点关注的AI融合技术方向。但是，语义通信的理论基础还不够成熟，虽然它与经典通信同时提出，但是香农在对经典通信做了大量假设后，基于热力学中的熵函数得到了一套完备的描述体系。而语义度量理论面对的是非线性问题，即便想要转化为单纯的数学模型也极为困难。本文在梳理语义度量理论发展历程的基础上，尝试从熵函数的角度切入，得到关于语义度量的一些有用结果。

1 语义度量理论的早期发展

语义通信的研究主要沿3条技术路径并行推进。第一条路为基于香农经典信息理论（CIT）的发展。基于CIT的分离定理[1]，当前的通信系统一般采用信源信道编码分离的方式搭建。但实际应用场景往往不满足分离定理成立的假设条件，因此人们试图测算联合编码方式能实现的增益，联合编码由此成为语义通信的开端。第二条路为语义度量理论的发展，包括WEAVER提出关于通信的三层问题[2]、经典语义信息理论（CSIT）[5]、强语义信息理论（TSSI）[6-7]等，它们共同构成了语义通信的理论基础。第三条路为AI的发展历程，包括“图灵机”的提出[8]、梯度回传算法的实现[9]、深度学习[10]以及现在的大模型[11]等。语义通信主要是在联合编码系统的基础上，基于语义度量理论，利用ANN实现并发展的。如图1所示，语义通信架构在经典通信架构的基础上进行了扩展。本小节主要介绍语义度量理论的早期发展阶段和其中的主要思想。

1.1 针对通信的三层问题

CIT的提出引发了学界的热烈讨论，并迅速形成两种对立观点。一派坚定地支持香农理论，认为其完备地描述了通信所面临的工程问题；另外一派则认为CIT存在极大局限，无法刻画人类实际通信背后的深刻意义。在这样的背景下，当时的语言学大师WEAVER在CIT文章再版时，与香农讨论后为其补充了一段序言，将通信问题拓展为三层问题[2]，一般将其翻译为：

- 语法问题：通信系统能多准确地传输符号？
- 语义问题：通信系统能多准确地传输符号含义？
- 语用问题：传输的符号含义能多准确地完成通信任务？

CIT几乎完美地解决了语法问题。但受限于工程实现需求，CIT将语义问题和语用问题排除在外，香农后来也曾在率失真理论文章中再次提到了人类对于符号的理解，即语义问题，称其无法通过简单的误码率来描述，而需涉及“可懂度”的概念。

语义问题的本质是建立语义空间到数据空间的映射，这一问题对应非线性编码过程，因此在统计理论中难以直接描述和求解。更进一步地，一般认为解决语义问题是解决语用问题的基础，只有理解了传输符号背后的含义，才能进一步将其映射为通信任务的完成程度。语义通信就是针对这两个问题背后复杂的统计理论而诞生的。可以认为，语义通信是语义问题和语用问题的综合解决方案。

1.2 经典语义信息理论

与WEAVER通信理论几乎同时诞生的是BAR-HILLEL与CARNAP提出的第一个对命题语义信息的描述[5]，该理论一般被称为经典语义信息理论，简称为CSIT。CSIT沿袭了CIT的思路，但是将统计概率替换为语言的逻辑概率，并只针对一些逻辑命题。CSIT认为一个命题包含的语义信息与其逻辑概率成反比，即一个命题出现的逻辑概率越大，其语义就越小。例如，A=“小白是条狗或猫”，其逻辑概率大于B=“小白是条狗”，因而A的语义信息量就大于B。CSIT定义的语义信息公式为：

$$\mathrm{cont}(A) = {}_{df}1 - \Pr(A) \tag{1},$$

$$\inf(A) = {}_{df} - \log_2[\Pr(A)] \tag{2},$$

其中，$\Pr(A)$表示命题A出现的逻辑概率，$\mathrm{cont}(A)$与$\inf(A)$分别表示描述命题A语义信息的两种形式。

但是，CSIT的定义产生了一个悖论，人们称其为BAR-HILLEL-CARNAP PARADOX（BCP），即如果是矛盾命题，比如$A \wedge \overline{A}$，由于其产生的逻辑概率为0，那么按照CSIT，其语义信息量为无穷大，这显然是违背常识的。BAR-HILLEL与CARNAP也给出了相关解释，他们认为“矛盾本身就包含了太大的信息量”是其成为矛盾的原因，但这样的解释显然不能让人满意。之后也有很多学者沿着BAR-

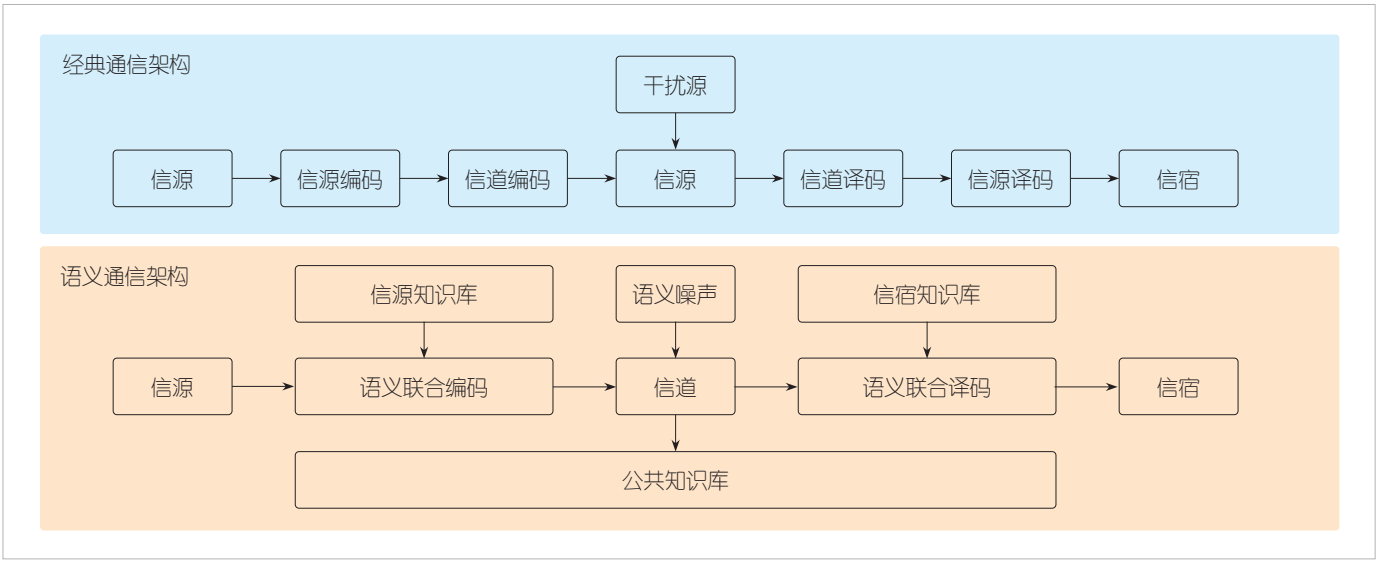


图1　经典通信系统架构与语义通信系统架构对比图示

HILLEL与CARNAP提出的研究路径描述信息量，但是都无法完美地解决这个悖论问题[12]。

1.3 强语义信息理论

直到2001年，FLORIDI认识到BCP出现的原因在于没有规避语义真性的问题[6]。在CIT中，由于人为规定了0log0 = 0，因此自然规避了不存在的信息。CSIT同样沿用了概率方法来描述语义信息，但是当命题语义不为真，即为矛盾命题时，就会出现语义信息趋于无穷大的情况。为了解决BCP问题，FLORIDI将语义真性封装在语义信息度量标准内，提出了另外一套语义度量标准[7]，一般被称为强语义信息理论，简称为TSSI。TSSI中存在一个真命题，其他命题与真命题存在一定距离，该距离属于[–1,1]区间。其中，–1代表全假，即该命题中没有一种情况符合真命题；1代表该命题符合所有情况的真命题，但是不止符合当前情况下这一种真命题；0则为当前条件下的真命题。语义信息公式定义为：

$$g(s) = 1 - f^2(s) \tag{3},$$

其中，$f(s)$表示命题s到真命题之间的距离，$g(s)$表示命题s包含的语义信息。

TSSI通过规定距离的方式定义语义信息，避免了BCP。但是由于FLORIDI本人是一位语言学家，其定义的语义信息计算公式过于唯象，公式之间缺乏推导和逻辑。这导致TSSI只适用于一些简单的逻辑命题。当命题之间需要进行逻辑运算，或者面对复合命题时，TSSI便难以有效处理。

1.4 语义真性与概率统计的结合

2011年，D’ALFONSO提出了一种将概率和距离融合起来的语义度量理论[13]。该理论针对TSSI直接将矛盾排除在外，并对数学描述不完备的问题进行改进，解决了以下问题：数学描述不完备问题（定义了6种真性距离与语义信息之间的关系）、语用偏好问题（距离公式中添加权重）、矛盾不可计算问题（引入介于语义真假的第三态）、与CSI度量不一致问题（距离定义中加入概率描述），并且引入了效用与决策理论来分析语义信息（距离函数中引入效用参数）。这6种距离与语义信息之间的运算关系分别为：

• 距离与语义关系1：$\Delta_{\min}(A,T) = \Delta_{\min}(w_a,w_T)$，其中$w_a \in W_A$。

• 距离与语义关系2：$\Delta_{\max}(A,T) = \Delta_{\max}(w_a,w_T)$，其中$w_a \in W_A$。

• 距离与语义关系3：$\Delta_{\text{sum}}(A,T) = \sum_{w_a \in W_A} \Delta(w_a,w_T) / \sum_{w_b \in B} \Delta(w_b,w_T)$。

• 距离与语义关系4：$\Delta_{\text{av}}(A,T) = \sum_{w_a \in W_A} \Delta_{\max}(w_a,w_T) / |W_A|$。

• 距离与语义关系5：$\Delta^{\gamma}{}_{\min}(A,T) = \gamma\Delta_{\min}(A,T) + (1-\gamma)\Delta_{\max}(A,T)$，其中$0 \leqslant \gamma \leqslant 1$表示权重。

• 距离与语义关系6：$\Delta^{\gamma\lambda}{}_{\text{ms}}(A,T) = \gamma\Delta_{\min}(A,T) + \lambda\Delta_{\text{sum}}(A,T)$，其中$0 \leqslant \gamma \leqslant 1$与$0 \leqslant \lambda \leqslant 1$表示权重。

在上述运算中，$\Delta(w_a,w_T)$表示命题w_a与真性命题w_T之间的真性距离，$\Delta(A,T)$表示命题w_a在真性命题w_T定义的场景T中具有的语义信息A。

D’ALFONSO结合CSIT与TSSI的思想，给出了一种新

的$\Delta(w_a,w_T)$，其改进的语义信息公式为：

$$\mathrm{val}(w)=\frac{t}{n\times 2^n} \tag{4}$$

其中，n表示逻辑空间中逻辑命题的数量，t表示命题w在实际情况中包含的所有真命题的数量。需要说明的是，公式（4）已经将真命题及其存在的场景整合在内了，因此该公式更倾向于描述命题w的性质，而非与真性命题的距离。

2 后香农时代的语义度量理论

事实上，上一节提及的语义度量理论均针对简单的逻辑命题，目前还没有一套通用的理论能够有效衡量人类语言中语义信息的衡量体系。但是，也有很多研究者从各个角度，对语义信息进行了尝试性描述，比如代数信息理论[14-15]、通用语义信息理论[16-17]、语义编码理论[18]等。本节举出几个对语义通信影响较大的度量方式。

2.1 经典语义信息理论的扩展

对语义通信影响最大的是BAO等在2011年发表的研究[19]。该研究首次提出语义通信系统中的本地知识库与共享知识库概念，仍以熵函数为核心理论思路。BAO等认为，扩展CSIT可以使该理论体系完备地描述语义信息。该理论同样对语义描述的适用范围做了严格假设。在该研究中，语义信源发出的消息被假定为真，这样就避免了BCP。定义信源语义信息为一个元组$\{W_s,K_s,I_s,M_s\}$，信宿语义信息也为一个元组$\{W_r,K_r,I_r,M_r\}$：

- W_s,W_r：表示信源/信宿可能观察到的世界模型。
- K_s,K_r：表示信源/信宿的背景知识库。
- I_s,I_r：表示信源/信宿的推理程序。
- M_s,M_r：表示信源/信宿用来编码/解码消息的消息生成/解释器。

与CSIT不同之处在于，该研究重新定义了消息x的逻辑概率与语义信息：

$$m(x)=\frac{\mu(W_x)}{\mu(W)}=\frac{\sum\limits_{w\in W,w\mapsto x}\mu(w)}{\sum\limits_{w\in W}\mu(w)} \tag{5}$$

$$H_s(x)=-\log_2(m(x)) \tag{6}$$

其中，$\mu(W_x)$表示消息x在世界模型中对应模型集合的逻辑概率，$\mapsto$表示一般命题可满足关系，即消息x的模型，$\mu(W)$表示世界模型的逻辑概率。需要注意的是，在CSIT的定义中，分子为命题x包含的所有原子命题数，分母为当前情况下的所有原子命题，这是两者的主要差异。

如果存在知识库K，那么上述语义信息量被定义为逻辑概率的条件信息：

$$m(x|K)=\frac{\sum\limits_{w\in W,w\mapsto K,x}\mu(w)}{\sum\limits_{w\in W,w\mapsto K}\mu(w)} \tag{7}$$

$$H_s(x|K)=-\log_2(m(x|K)) \tag{8}$$

需要注意的是，上述条件熵是由逻辑概率函数定义的，其与经典统计中的概率不同，没有可加性，这也是我们认为该描述存在的最大问题。3.1节将会解释没有可加性的量并不能被熵函数唯一表示。更进一步地，这里的逻辑概率没有给出具体的计算方式。而我们认为，逻辑概率与统计概率之间的映射关系才是求解语义信息量的核心问题。

2.2 其他语义度量理论

KOLCHINSKY等[20]将语义信息定义为“物理系统对其环境的信息，这种信息对于系统随时间维持自身存在是因果必需的”。该定义基于系统及其环境的内在动力学，同时借鉴了信息理论和非平衡统计物理学的思想，产生了信息度量的非负分解，并将其分为“有意义的比特”和“无意义的比特”。更进一步地，该定义为表达与“语义信息”相关的一系列概念提供了一个连贯的量化框架，如“信息价值”“语义内容”和“代理性”。

KOUNTOURIS等[21]则利用AI来定义语义信息，基于过程动力学、信号稀疏性、数据相关性和语义信息属性，将AI技术中的信息生成、信息重构与经典通信中的信息传输进行有机结合。他们的实验基于语义进行采样和通信，展示了重构误差和驱动误差成本，以及生成的无信息样本数量。

华为香港理论研究部同样希望借助AI来研究语义信息[22-23]。但是由于AI本身是一个“黑盒子”，难以进行量化，语义信息和AI理论研究的过程是相辅相成的。他们认为，语义通信一定是语义问题和语用问题的结合，不能分开考虑。语义信息度量问题可能是ANN可存储信息量的问题，它采用Korner熵来描述AI状态数与语义信息量[24]。在假定ANN模型没有推理能力的前提下，Korner熵可以近似估计ANN模型参数与可存储状态数之间的关系。然而截至目前，该研究尚未取得实质进展。

3 语义度量的假设条件

基于上述语义度量理论的发展历史，本文结合当前已有的研究，对语义度量中的假设进行总结和分析，试图找到能够用来描述语义信息的相关量。

3.1 熵函数与语义信息

香农利用热力学中的熵公式度量信息量并非偶然[1]，因为由符号表示的信息满足4条公理，即对称性、扩展性、极值性与可加性。符合这4个公理描述的信息量可以被度量为熵，且这种表示是唯一的——这被称为熵的唯一性定理。在熵唯一性定理证明中，可加性是该定理存在的必要条件，而其他3项则可以适当弱化。如下为4条公理与熵唯一性定理[25]：

• 熵的对称性：对于任意K，$H_K(\boldsymbol{P})$对$\boldsymbol{P}$的所有分量的连续和对称。

• 熵的扩展性：对于任意K，有$H_{K+1}(P_1,P_2,\cdots,P_K,0)=H_K(P_1,P_2,\cdots,P_K)$。

• 熵的极值性：对于任意K，有$H_K(P_1,P_2,\cdots,P_K)\leqslant H_K(\frac{1}{K},\cdots,\frac{1}{K})$。

• 熵的可加性：对于$M=\sum_{i=1}^{K}m_i$，存在：

$$H_M(P_1Q_{11},P_1Q_{21},\cdots,P_1Q_{m_11},P_2Q_{12},P_2Q_{22},\cdots,P_2Q_{m_22},\cdots,P_KQ_{m_KK})=H_K(P_1,P_2,\cdots,P_K)+\sum_{k=1}^{K}P_kH_{m_k}(Q_{1k},Q_{2k},\cdots,Q_{m_ik}) \quad (9)。$$

• 熵唯一性定理：令$H_K(P_1,P_2,\cdots,P_K)$是概率矢量$\boldsymbol{P}=(P_1,P_2,\cdots,P_K)$的非负实函数，其中$P_K\geqslant 0$，$\sum_{k=1}^{K}P_k=1$。若$H_K(P_1,P_2,\cdots,P_K)$满足上述4条公理，则有$H_K(\boldsymbol{P})=-\lambda\sum_{k=1}^{K}P_k\log(P_k)$，其中$\lambda$为一个正常数。

熵能否作为语义信息的唯一度量指标，取决于语义是否满足上述4条公理。

事实上，人类语言具有非平稳性、模糊性，语言之美也正体现在其模糊性中。正是得益于这种模糊性，人类才能欣赏各种文章、乐章等，甚至从中体会出一些“非语义”“非语法”的信息，也正是这种模糊性才使得语言并非简单的词汇叠加，而是在叠加过程中形成有结构、有组织的序列。这种构成使得该叠加产生了1+1>2的效果，否则就会出现“断章取义”的问题。因此，语义需要依托语法编排才能存在。语义信息既不满足对称性，更不满足可加性。熵公式并不能作为语义信息的唯一度量形式。语义信息是一种更高维的信息存在形式。此外，可加性这一公理的本源来自于范畴论，若要进一步研究其内禀假设，还需要结合范畴论中的相关理论展开分析，这里不再赘述。

3.2 关于语义信息的假设

类似于熵的4条公理，在语义信息漫长的研究历程中人们也总结出了4条假设[6–7,13]。如果将消息m中具有的语义信息量标记为σ，则有：

• 数据假设：σ是基于一个或者一组数据d的。

• 格式假设：数据d是格式良好的。

• 意义假设：每一个数据d是具有意义的，且记为δ_d。

• 真性假设：σ是具有真性的。

结合上述4条假设，基于CIT的经典通信要解决的问题是如何编码数据d，语义通信要解决的问题是如何编码σ。假设组成消息m的数据为$\{d_i;i=1,\cdots,n\}$，σ并非是每一个数据d_i的意义δ_{d_i}的简单叠加，而是在d_i变得格式良好的过程中，进一步消除不确定性，使得σ小于等于δ_{d_i}的直接叠加，因此类似于消息m为数据d的函数，σ也是关于δ_{d_i}的一个函数：

$$m=wf_d(d) \quad (10),$$

$$\sigma=wf_\delta(\delta_d) \quad (11),$$

其中，$wf_d()$表示数据格式函数，$wf_\delta()$表示意义格式函数。

3.3 语义空间定义

为了更好地描述上述语义信息假设，本文进行如下定义：

• 数据d：人能够感受到的且可以记录世界状态变化的物理载体。

• 数据空间$D=\{d\}$：某一种物理载体的集合。

• 语义空间$\sigma=wf_D(\delta_D)$：某一种物理载体D所包含的意义δ_D，以及其上所有格式函数$wf_D()$所产生的语义信息集合。

针对上述对语义信息的分析可以发现，语义信息是基于数据存在的。一句话的含义不会因是否有通信而发生变化，语义信息是数据的内禀属性。因此，在语义信息的研究中，不考虑信息产生方和接收方的信息，即认为通信双方是绝对理性客观的。类似地，由熵函数描述的数据信息，也是客观存在的，不会因是否存在通信而改变。

此外，也有多种物理载体d可以承载语义，比如汉语、英语、图片、视频等。然而，不同载体记录语义信息的能力是等效的。基于数字化的信息时代都建立在该假设之上。因为从数字的起源来看，其本质也是一种自然语言，同样被视为一种语义信息载体。试想，对于一个不懂英文的人，面对英文时看到的就是26个字母得到的编码，无法感知其语义含义；而懂英文的人由于知道英文的语法与词意，因此可以

解读其语义含义。对于最早的程序员而言，一些机器码是有具体含义的，因为他们了解机器码的语法。而对普通人而言，机器码可能就是一堆无意义的0和1组成的序列。我们可以将上述假设归纳为[6,26]：

• 属性假设：意义是数据的内禀属性。

• 等价假设：不同物理载体承载语义信息的能力等价。

4 语义信息熵的定义与性质

语义信息不能使用熵公式进行唯一刻画，但或许可以借助这种形式确定其度量的上下界。基于数据假设，数据度量应该是语义信息度量的一种特殊形式。本文从该特殊形式出发，对理想中的极限情况进行定义。

4.1 理想语义空间与数据空间

考虑一个理想中的数据集$D=\{d_i\}$，其内单个数据能够独立无关联地承载语义信息，并且承载的语义空间与其数据空间一一对应。此时该数据集描述的语义具有渐进均分性，基于熵公式计算的数据信息量，等于该数据集承载的语义信息量。此时$wf_D()$与$wf_\delta()$不产生作用，且格式假设不存在，那么理想数据集可以定义为：

• 元数据空间D：一种物理载体$D=\{d_i\}$，其内单个数据能够无关联地承载语义信息，并且承载的语义空间$\{\delta_D\}$与其数据空间D一一对应。其中，每一个d_i被称为一个元数据。

• 元语义空间Δ：元数据空间D所承载的语义空间$\{\delta_D\}$。其中，每一个元数据d_i所代表的意义δ_{d_i}被称为一个元语义。

事实上，香农在定义通信模型的过程中，明确指出不考虑符号背后的含义，认为符号是独立同分布的，具有渐进均分性，是一种典型集，并且样本概率近似为总体概率，从而近似得到数据熵。该假设是一个很强的假设（此时格式假设不存在），在这种假设下符号独立，因此可以认为其承载的语义信息也是独立的。可以看出，香农通信模型中的信息量为语义信息的一种特殊情况，元数据空间的语义信息就属于这种情况。

但是，一旦考虑符号背后的含义，以及含义之间的关联和语法赋予的关联，符号就不再是独立同分布的，因此不再属于典型集，甚至不是线性的、平稳的、各态历经的。样本概率无法近似总体概率，数据熵无法描述语义信息，经典信源编码、信道编码定理不再适用。例如，选择不同的符号种类或者符号组合作为单个数据时，所选数据单元的概率分布会存在差异，最终计算得到的数据熵也不相同。这是因为，数据本身具有含义，描述含义的物理载体具有一定格式。数据分布的改变反映了数据格式的改变。数据分布实际上包含了格式假设，使用统计概率代替了数据中一部分语义关联性。因此，语义信息与不同物理载体的数据分布相关。

此外，经典信息理论基于数字化和编码，这使得熵函数可以实现精确计算。但是如果考虑信息的原本描述，其往往是连续的，而非离散的。连续变量的熵由微分熵和绝对熵构成，其信息量趋于无穷，无法计算。近似计算需要对绝对熵进行离散处理，本质是对描述物理世界的数据进行重新采样。而语义编码一般基于ANN实现，具有概率意义，属于连续变量。因此，语义信息也与信源的采样方式或者数据描述物理世界的方式有关，这一般被称为语义采样。

图2为从数据空间转为语义空间的示意图。在图2（a）所示的数据空间中，单个数据在经典信息理论中是独立同分布的。图2（b）则进一步纳入数据的语义属性。意义相近的数据1与数据2属于同一个语义空间，针对其数据的编码距离也应更近。此时数据之间存在关联，不再是独立分布，无法用典型集描述。图2（c）展示了数据到语义之间的非线性映射关系。由于数据存在多义、歧义，因此在意义上的关联是非线性的。真实物理载体的单个数据并非只具有意义，而是同样包含一定的关联信息。虽然本文认为考虑了语义后的数据集属于非典型集，无法直接用熵函数描述，但是符合典型集的数据为非典型集的一种极限情况。或许图2（c）展示的语义信息可以基于极限情况，由图2（a）中数据的熵函数进行逼近。

4.2 语义信息定义与特征

在实际情况中，一般物理载体的数据格式$wf_D()$与语义格式$wf_\delta()$并不等价，因此，其熵函数计算得到的信息量并不等于其语义信息量。按照3.2节中的四大假设，以及4.1节中的分析，本文将实际物理载体承载的语义信息分为两部分考虑。

对于实际数据集$E=\{e_i;i=1,2,\cdots,n\}$，其中n表示数据集E中具有单个数据的个数，单个数据e_i之间都具有语义关联性，此时熵函数无法描述语义。针对某消息$M=\{e_j;j=m_1,m_2,\cdots,m_k\}\subseteq E$由$E$中的$k$个单数据组成，那么定义消息$M$的语义信息量：

组成消息M的数据$\{e_j\}$本身包含的语义信息量与消息M基于的实际数据E所包含的语义关联信息量，在语义空间中的总和可以表示为：

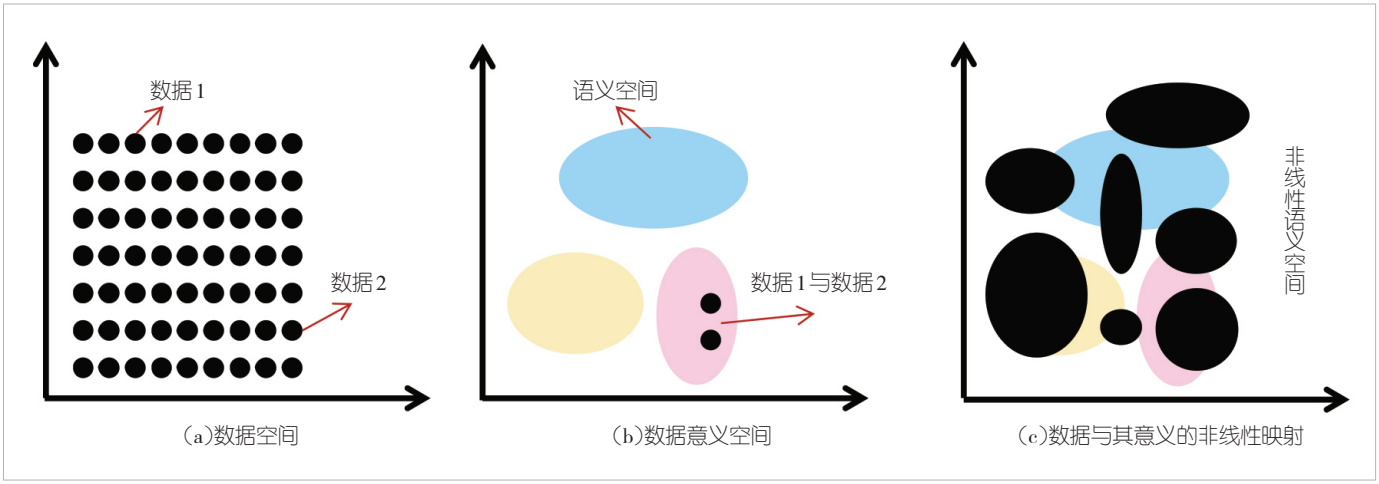


图2　数据空间与语义空间之间映射的非严格表示

$$S(M) = wf_{\delta}(\delta_M) = \lambda(\delta_M) \otimes cor(\delta_E) \quad (12),$$

其中，⊗表示高维运算，是在不能确定语义格式函数$wf_{\delta}()$具体形式时的权宜表示。$cor(\delta_E)$表示实际数据E所包含的所有单个数据之间的语义关联，称为语义关联信息。$\lambda(\delta_M)$表示组成消息M的单个数据意义的集合$\{\delta_{e_j}; j = m_1, m_2, \cdots, m_k\}$，称为语义特征信息。

图3为语义信息的可视化示意图。其中，蓝色表示物理世界的整体描述，即本文定义的数据。横坐标表示数据的语义关联信息。由于语义关联信息包含了单个数据意义之间的关联，以及语法赋予的确定性信息，因此无法严格地示意出来。纵坐标表示数据的语义特征信息。元数据空间中的数据意义集合既等价于语义特征信息，也等价于语义信息。然而，实际数据空间中的数据，其意义之间存在关联。

需要说明的是，数据假设体现在语义特征信息里，但同时实际数据的语义特征信息也包含了一部分格式假设，因为

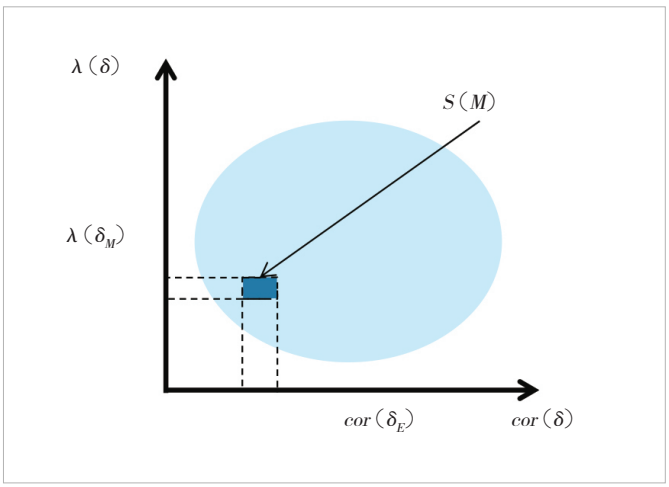


图3　语义信息的非严格示意图

其存在一定的分布特征。而语义关联信息则是格式假设的另一大体现，它不仅涵盖了分布特征，还融入了真性假设。此外，语义关联信息属于确定性信息，其作用相当于消除语义特征信息中的一部分不确定性。语义信息的核心性质如下：

• 语义特征信息一定大于等于语义信息。

• 语义关联信息包括词义关联信息，即词义与词义之间的关联。这一类信息类似于自然语言处理（NLP）中的词嵌入向量，其与基于真性距离的TSSI相关。

• 语义关联信息包括语法关联信息，即上下文之间的关联。这一类信息类似于NLP中的语言模型，其与基于逻辑概率的CSIT相关。

根据四大假设，语义信息包含数据本身的意义与意义编排。这里将数据本身的意义视为语义特征信息，将意义编排产生的效果视为语义关联信息。根据4.1节中定义的语义空间，这些关联或许无法通过离散变量完整描述，可能需要采用连续变量描述。这就使得原本基于经典概率空间的熵函数，扩展到基于概率分布函数的熵函数。其中，微分熵或许可以用来表示上述语义特征信息，绝对熵可以用来表示语义关联信息。

4.3 熵函数逼近下的语义信息

为了得到语义信息的界限，需要进一步研究4个假设。

基于数据假设和格式假设，可以得到：

• 定理一：任意自然语言的数据熵下确界小于等于其携带的语义信息：

$$S(\delta_E) \geqslant \inf_{E \subset \Omega} H(E) \quad (12),$$

其中，Ω表示所有自然语言集合。

基于数据假设，语义信息一定需要数据才能存在，即数

据是语义存在的必要条件。因此定理一的成立可以直接证明。针对消息M，在元数据空间D中，其数据信息等于语义信息：

$$H_D(M) = S(M) \tag{13}$$

其中，$H_D(M)$表示采用物理载体D表达消息M熵的大小。但是对于真实数据E，其相比于元数据，由于存在格式假设，必然存在数据间关联。该关联体现在真实数据的不同分布形式中，即元数据分布必然比真实数据分布更均匀。这是由于关联关系消除了一部分不确定性。因此，真实数据熵下确界小于等于元数据熵：

$$H_E(M) \leqslant H_D(M) \tag{14}$$

推广到一般情况，可以得到数据熵下确界小于等于其携带的语义信息，即$S(\delta_E) \geqslant \inf_{E \subset \Omega} H(E)$。

此外，基于意义假设，可以得到：

• 定理二：任意自然语言的数据熵上确界大于其携带的语义信息：

$$\sup_{E \subset \Omega} H(E) > S(\delta_E) \tag{15}$$

其中，Ω表示所有自然语言集合。

基于意义假设，所有数据都具有一定的意义。针对消息M，在元语义空间Δ中，组成消息M的所有单个数据意义的集合，即该消息的语义特征信息等于语义信息，也等于：

$$\{\delta_d\}_{d \subset M \wedge d \subset D} = \lambda(\delta_M) = H_D(M) \tag{16}$$

其中，$\{\delta_d\}_{d \subset M \wedge d \subset D}$表示采用元数据空间$D$中消息$M$包含的所有单数据意义的集合。但是真实数据的语义信息小于等于其语义特征信息：

$$S_E(M) \leqslant \lambda(\delta_M) \tag{17}$$

推广到一般情况，可以得到数据熵的上确界大于等于其携带的语义信息。但是由于在语义空间中，格式假设一定存在，因此等于并不成立，因此$\sup_{E \subset \Omega} H(E) > S(\delta_E)$。

根据上述两个定理，可以得到语义信息界为：

$$\sup_{E \subset \Omega} H(E) > S(\delta_E) \geqslant \inf_{F \subset \Omega} H(F) \tag{18}$$

其上界说明，数据熵描述的信息由于未考虑数据背后意义的关联性，均可以依靠语义信息进行进一步压缩。而下界说明，语义编码本质是找到一个最有效率的物理载体，该载体能够充分依靠语义关联性减少数据信息量，从而进行高效的语义描述。

更进一步地，熵的上确界与下确界之所以存在，本质是考虑了数据背后的含义以及含义之间的关联。对于经典信息理论，数据属于典型集，样本概率可以近似为总体概率，因此数据熵能够收敛到一个确定的数据，并且这个数值可以用样本估计出来。但是，一旦考虑数据的语义，这种数据集不再是独立同分布，样本概率无法逼近总体概率，数据不再属于典型集，熵无法被准确估计，或者说此时估计的熵与所使用的数据集和数据采样方式均有关。因而，熵存在上确界与下确界。

针对消息M，语义编码需要找到合适的数据集与采样方式，尽可能有效地描述M的语义。此时的数据集与采样方式决定了语义关联信息，该信息为确定性信息，并被存储在知识库中。依靠知识库，语义通信可以进一步消除语义的不确定性，实现通信数据压缩，节省通信资源。根据定理一，其编码的最小长度为：$\inf_{F \subset \Delta} H(F)$，所能节省的最大数据量为$\sup_{E \subset \Delta} H(E) - \inf_{F \subset \Delta} H(F)$。

5 结束语

本文从经典通信出发，详细介绍了其面临的主要困境，并指出语义通信可能是突破该困境的方式之一。而语义通信的研究由来已久，主要分为3条路，其中最重要也是发展最缓慢的是语义度量理论。在语义度量的漫长发展历史中，使用逻辑概率与真性距离描述语义信息是最重要的两类思想。此外，也不乏基于物理系统、基于AI等的语义信息度量理论。本文综合上述研究，提取出语义度量的4条基本假设，试图将CIT与TSSI等归纳入统一的语义度量理论中，实现概率视角下语义度量方式与真性距离视角下语义度量方式的有机结合，给出语义空间、语义信息等的具体定义。同时，基于熵函数，采用理想实验得到了语义信息的上下界，从而对语义编码进行实际指导。

致谢

感谢中兴通讯股份有限公司算法部胡留军、郁光辉、梁楚龙等专家对本研究的帮助！

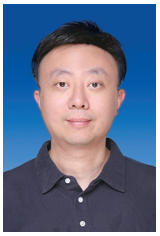
参考文献

[1] SHANNON C E. A mathematical theory of communication [J]. Bell system technical journal, 1948, 27(3): 379–423. DOI: 10.1002/j.1538–7305.1948.tb01338.x
[2] WEAVER W. Recent contributions to the mathematical theory of communication [EB/OL]. [2025–11–05]. https://courses. ischool. berkeley.edu/i218/s15/Weaver_Recent–Contributions.pdf.
[3] HORNIK K, STINCHCOMBE M, WHITE H. Multilayer feedforward networks are universal approximators [J]. Neural networks, 1989, 2(5): 359–366. DOI: 10.1016/0893–6080(89)90020–8
[4] ITU. Framework and overall objectives of the future development of IMT for 2030 and beyond: ITU M.2160 [S]. 2023
[5] BAR–HILLEL Y, CARNAP R. Semantic information [J]. The British

journal for the philosophy of science, 1953, 4(14): 147–157. DOI: 10.1093/bjps/iv.14.147
[6] 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
[7] FLORIDI L. Is semantic information meaningful data? [J]. Philosophy and phenomenological research, 2005, 70(2): 351–370. DOI: 10.1111/j.1933–1592.2005.tb00531.x
[8] TURING A. On computable numbers, with an application to the entscheidungsproblem (1936) [M]//The Essential Turing. Oxford University PressOxford, 2004: 58–90. DOI: 10.1093/oso/9780198250791.003.0005
[9] RUMELHART D E, HINTON G E, WILLIAMS R J. Learning representations by back–propagating errors [J]. Nature, 1986, 323(6088): 533–536. DOI: 10.1038/323533a0
[10] LECUN Y, BENGIO Y, HINTON G. Deep learning [J]. Nature, 2015, 521(7553): 436–444. DOI: 10.1038/nature14539
[11] ZHAO W X, ZHOU K, LI J, et al. A survey of large language models [EB/OL]. (2023–03–31) [2025–11–05]. https://arxiv. org/abs/2303.18223
[12] BACCHUS F I. Representing and reasoning with probabilistic knowledge [M]. Cambridge: MIT Press, 1993
[13] D’ALFONSO S. On quantifying semantic information [J]. Information, 2011, 2(1): 61–101. DOI: 10.3390/info2010061
[14] KOHLAS J, SCHNEUWLY C. Information algebra [M]. Berlin: Springer–Verlag Berlin Heidelberg, 2009
[15] LANGEL J. Logic and information: a unifying approach to semantic information theory [EB/OL]. [2025–11–05]. https://docslib. org/doc/4260093/logic–and–information–a–unifying–approach–to–semantic–information–theory
[16] JUBA B, SUDAN M. Universal semantic communication I [EB/OL]. [2025–11–05]. https://people. seas. harvard. edu/~madhusudan/papers/2007/juba–full.pdf
[17] JUBA B, SUDAN M. Universal semantic communication II: a theory of goal–oriented communication [EB/OL]. [2025–11–05]. https://people.csail.mit.edu/madhu/papers/2008/usc2.pdf
[18] WILLEMS F M J, KALKER T. Semantic compaction, transmission, and compression codes [C]//Proceedings of International Symposium on Information Theory. IEEE, 2005: 214–218. DOI: 10.1109/ISIT.2005.1523325
[19] BAO J, BASU P, DEAN M K, et al. Towards a theory of semantic communication [C]//Proceedings of IEEE Network Science Workshop. IEEE, 2011: 110–117. DOI: 10.1109/NSW.2011.6004632
[20] KOLCHINSKY A, WOLPERT D H. Semantic information, autonomous agency and non–equilibrium statistical physics [J]. Interface focus, 2018, 8(6): 20180041. DOI: 10.1098/rsfs.2018.0041
[21] 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
[22] ORLITSKY A, ROCHE J R. Coding for computing [C]// Proceedings of IEEE 36th Annual Foundations of Computer Science. IEEE, 1995: 502–511. DOI: 10.1109/SFCS.1995.492580
[23] YUAN D H, GUO T, BAI B, et al. Lossy computing with side information via multi–hypergraphs [C]//Proceedings of IEEE Information Theory Workshop (ITW). IEEE, 2022: 344–349. DOI: 10.1109/ITW54588.2022.9965914
[24] HARANGI V, NIU X Y, BAI B. Generalizing Körner’s graph entropy to graphons [J]. European journal of combinatorics, 2023, 114: 103779. DOI: 10.1016/j.ejc.2023.103779
[25] FEINSTEIN A. A new basic theorem of information theory [J]. Transactions of the IRE professional group on information theory, 1954, 4(4): 2–22. DOI: 10.1109/TIT.1954.1057459
[26] WU J. The beauty of mathematics in computer science [M]. Boca Raton, FL: Chapman and Hall/CRC, Taylor & Francis Group, 2019. DOI: 10.1201/9781315169491

作 者 简 介

唐雪，中兴通讯股份有限公司副总裁、无线及算力产品经营部战略架构总经理；主要负责无线及算力核心产品全球战略规划、核心技术规划、高端市场建设等。



许进，中兴通讯股份有限公司无线算法部部长；主要从事新型调制编码、网络编码、语义通信等新技术的预研；发表论文10余篇。

冯雨龙，中兴通讯股份有限公司算法工程师；主要研究领域为语义通信、人工智能、机器学习；发表论文7篇。