中国科学: 信息科学 2022 年 第 52 卷 第 5 期: 907–921
SCIENTIA SINICA Informationis
论文

面向智能体的语义通信: 架构与范例

张亦弛[1], 张平[2], 魏急波[1], 赵海涛[1*], 熊俊[1], 张姣[1]

1. 国防科技大学电子科学学院, 长沙 410073
2. 北京邮电大学网络与交换技术国家重点实验室, 北京 100876
* 通信作者. E-mail: haitaozhao@nudt.edu.cn

收稿日期: 2020–12–08; 修回日期: 2021–03–20; 接受日期: 2021–05–24; 网络出版日期: 2022–05–12

国家自然科学基金 (批准号: 61931020, U19B2024, 62001483) 和湖南省科技创新计划 (批准号: 2020RC2045) 资助项目

摘要 语义通信关注的是语义的准确传递, 而不是数据或通信符号的准确传输, 因此可以达到用更低的通信量实现更高效信息交互的目标, 是适用于智能体交互的高效通信机制. 本文分析总结了与语义通信紧密相关的一些研究工作, 探讨了一种面向智能体的语义通信架构, 并分析了该架构下各模块的功能; 进一步, 从有效性和可靠性两个方面分析了语义通信性能的指标和评估方法; 最后, 通过一个范例来说明语义通信的实现过程, 证明其相较于传统通信方式可以显著降低通信量.

关键词 语义通信, 智能体通信, 语义相似度, 语义编码, 语义译码

1 引言

1948 年香农 (Shannon) 在其划时代的著作《通信中的数学原理》中, 抽象出了信息传输的经典系统模型, 提出了信息熵的概念, 并给出了著名的信道容量公式, 为现代通信系统的设计奠定了理论基础. 香农信息论的基本假设之一是数据的独立同分布 [1], 在其建立信息熵的过程中忽略了信息的含义. “在当时通信工程的特殊需求的背景下, 忽略信息的语义并且把重心放在研究可靠信道问题上是合理的, 但是不意味着永远应该这样做 [2]”.

Weaver [3] 进一步研究通信的深层含义, 并提出了 3 个层次的通信: 第 1 层次为技术问题, 它主要解决 “通信符号如何准确传输” 的问题; 第 2 层次为语义问题, 它主要解决 “传输的符号如何精准传达含义” 的问题; 第 3 层次为效用问题, 它主要解决 “收到的含义如何以期望的方式有效影响行为” 的问题. 自香农建立信息论以来的七十多年, 学者们就如何逼近香农限做出了大量卓越贡献, 这些工作主要集中在第 1 个层次. 近几年, 随着人工智能、自然语言处理等相关支撑技术的快速发展, 通信设备的智能化水平和对外界的认知能力不断增强, 为深入开展第 2 层次的语义通信问题研究提供了可能, 语义通信也逐渐成为通信领域的一大研究趋势 [4].

引用格式: 张亦弛, 张平, 魏急波, 等. 面向智能体的语义通信: 架构与范例. 中国科学: 信息科学, 2022, 52: 907–921, doi: 10.1360/SSI-2020-0379
Zhang Y C, Zhang P, Wei J B, et al. Semantic communication for intelligent devices: architectures and a paradigm (in Chinese). Sci Sin Inform, 2022, 52: 907–921, doi: 10.1360/SSI-2020-0379

语义通信的核心不在于数据通信, 而是在于传输数据含义或内容通信 [2], 因此它不是以数据或通信符号的准确传递, 而是以将含义准确传递给对方为目标. 一个经典的例子是 “研表究明, 汉字序顺并不定一影阅响读”, 这句话中尽管有 “误码”, 但当看完这句话后并不影响我们对其表达意思的理解.

与传统通信相比, 语义通信可以获得更大的通信潜能. 首先, 发送端可以更好地理解通信的目的和环境, 消除冗余数据的传递. 随着通信和信息技术的不断进步, 系统产生的数据量也日渐庞大, 例如地理信息系统中大量交互产生的复杂数据流带来了大量数据冗余 [5]. 发送端根据对通信目的和历史通信过程的理解分析对需要传输的数据进行处理, 可从源头上避免大量冗余的传输, 同时可减少有用信息淹没在大量冗余带来的数据分析干扰, 有利于缓解无线通信系统中无线资源日趋紧张的问题. 其次, 语义通信中的背景知识有助于提高尤其是在带宽有限、信噪比较低或误码率较高的不理想通信环境中通信系统的有效性和准确性 [6]. 通信双方积累的上下文信息、个体通信目的等背景知识有助于在较差的通信环境中以及部分信息丢失的情况下对接收信息进行智能纠错和恢复, 保证了信息传输的有效性和可靠性的同时还避免或减少信息反复重传占用额外通信资源的情况. 第三, 有助于智能体间的通信. 有学者认为语义通信是智能体间必然的通信方式 [2]. 大量通信、计算功能融合的智能设备的出现, 以及数据驱动的人工智能技术的快速发展为研究语义通信技术打下了基础 [7]. 一方面, 因为智能体间更多的是交互任务驱动的业务, 语义通信能帮助其更精准地理解上下文. 另一方面, 人与智能体间通信过程中, 语义推理、语义理解是通信问题的直接组成部分 [8]. 第四, 语义通信还能达到更好地隐蔽通信效果. 因为语义通信主要是基于收发双方的上下文知识建立起来的, 同样的一句话, 对于了解背景的友方而言可能包含重大信息, 而对于不了解背景的窃听者而言可能毫无意义. 第五, 传输数据对于网络中不同接收方有相同价值的假设不再适用于物联网、云服务等新技术. 仅仅考虑接入时间、频谱等无线资源使频谱利用率最大化的方法不能满足新技术对通信网络自动化、智能化、服务多样化的需求, 而语义通信考虑信息含义和用户对信息的需求, 可在根本上实现跨系统、网络、人机的协议至最终实现 “万物透明智联” [8].

语义通信作为一种新型通信方式, 将对经典的通信体系带来颠覆性影响, 目前还有很多基础性问题没有研究清楚. 本文将分析面向智能体的语义通信的基本原理和实现架构, 探讨其评价指标, 并通过一个范例来说明语义通信的可行性及有效性. 本文剩余内容分成以下几个部分: 第 2 节介绍与语义通信相关的研究背景和现状; 第 3 节提出一种面向智能体的语义通信架构, 并阐述了架构中所涉及各个要素的功能; 第 4 节通过一个范例来说明语义通信的实现过程; 第 5 节分析了评价语义通信时应考虑的要素, 并给出了有效性和可靠性两个方面的评价指标及其计算方法; 第 6 节将范例与传统的方法进行了比较来说明语义通信的效能; 第 7 节对本文的研究进行了小结.

2 语义通信相关研究

尽管语义通信这一概念最近才重新被大家关注, 但就其本身的内涵和目标而言, 已有不少相关的研究工作, 包括信息中心网络 (information-centric networking)、基于上下文感知的通信 (context-aware communication)、知识为中心的通信和网络 (knowledge centric networking) 等, 这些研究本质上都是向着高于数据通信的第 2 层次的通信进行探索. 在本节中, 我们将对这些相关研究和语义通信的发展脉络进行介绍和分析.

2.1 信息中心网络

信息中心网络侧重于研究网络框架来应对未来网络的新需求, 主要包括高效的信息传递和支持移

动性等. Xylomenos 等[9] 对信息中心网络的网络体系结构进行了总结, 具体而言, 以下 7 个研究项目基本包括了信息中心网络的核心功能: (1) Berkeley 的面向数据的网络体系结构项目 DONA[10]; (2) 欧盟资助的发布 – 订阅网络技术项目 PURSUIT 及其前身发布 – 订阅互联网路由范例项目 PSIRP; (3) 可扩展和自适应的互联网项目 SAIL 及其前身未来互联网架构设计项目 4WARD; (4) 内容感知网络项目 COMET; (5) CONVERGENCE 项目; (6) 美国发起的基于命名数据的组网项目 NDN 及其前身以内容为中心的网络项目 CCN; (7) Mobility First 项目, 以及法国资助的采用 NDN 架构的 ANR Connect 项目.

2.2 基于上下文感知的通信和网络

基于上下文感知的通信, 其核心目的是对智能体所处环境进行特征化描述, 从而以更适配环境的方式完成通信. 上下文信息分成了 3 大类: 操作类上下文、通信类上下文和数据类上下文[1)]. 基于上下文的通信技术在认知无线电网络中有很好的应用前景. 最典型的例子是次级用户利用上下文感知信息检测主用户的存在[11]、用于信息安全的传输[12], 以及将上下文信息应用于路由选择等网络层决策[13~17]. 特别地, 文献 [17] 考虑了不同的中继优先级, 如最短可用距离、路由可靠性和信用信息, 利用上下文感知来提供可靠的路由, 使次级用户能够准确地估计满足其 QoS 要求所需的传输资源, 并且定义了新的路由度量. 随着机器学习技术的快速发展, 将其应用于上下文感知过程用来提升通信网络关键性能逐渐成为趋势, 代表性的工作如文献 [18~22] 等.

2.3 知识为中心的通信和网络

知识层的概念在 2003 年被提出[23], 它把知识作为网络体系结构的重要特征, 但是这种框架一直没有得到实现. 2018 年由 Wu 等[24] 提出知识为中心的通信网络 (knowledge centric networking, KCN), 其设计理念是试图通过面向服务的知识 (例如用户通信意图) 和面向网络的知识解决用户数量快速增长和冗余数据带来的网络负载能力不足等问题. 所构想的以知识为中心的通信网络可以通过信息聚合和过滤提取基本知识, 利用学习算法和推理预测从数据中获得高级知识. 收集、处理这些知识使网络可以理解周围的环境并对新状况及时反应, 相应地减少网络中传输的数据量并且最大化网络资源, 在此过程中可以自动发展出新的网络协议来调整动态网络资源适配传输的多样性. 知识为中心的通信网络将更符合现在复杂动态时变网络系统的传输形式, 但是目前还缺少具体的系统实现方案. 如何进行大数据收集处理、知识生成和管理、基于知识学习的网络优化等方面仍有待研究.

2.4 语义通信

语义通信的基础问题主要包括: 如何利用语义知识实现数据压缩和可靠通信? 语义编译码如何与现有经典编译码问题构建联系? 语义编码是否存在与第 1 层次通信系统中与香农限相类似的上限值? 应该考虑采用什么指标表征语义通信系统的有效性和可靠性? 等.

目前一部分工作从信息论角度探讨了语义通信压缩编码问题. Carnap 等[25] 首先提出了基于逻辑概率的语义信息理论 (semantic information theory). Bao 等[5] 定义了语义通信无损压缩编码过程中诸多概念 (例如语义噪声、语义冗余等) 和通用语义通信模型 (generic model of semantic communication), 并设计了度量语义信息的方法. Juba 等[26] 基于收发两方字符出现先验概率分布可能不一致的前提条件下设计了有损压缩编码方案, 并通过信息论证明了模糊性对于压缩的必要性. 该压缩编码方案通

1) Wireless innovation forum top 10 most wanted wireless innovations, version V4.0.02. 2015. https://www.wirelessinnovation.org/assets/workproducts/Reports/.

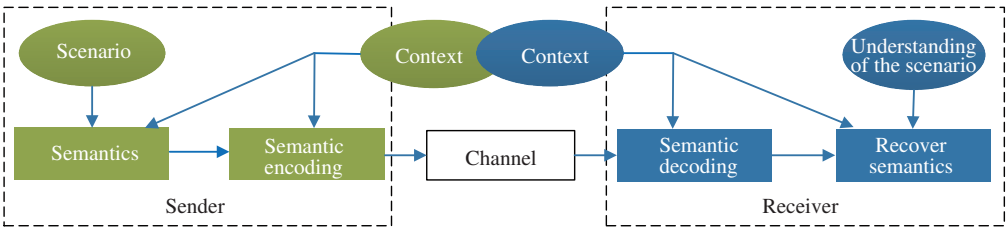

图 1　(网络版彩图) 语义通信架构

Figure 1　(Color online) The framework of semantic communication

过构建二分图建立了知识库中信息与编码的映射关系. Güler 等 [27] 引入第三方, 利用博弈论方法从语义相似度角度设计有损压缩编码方案, 其目标为最小化端对端平均语义错误.

另一些工作从具体语义编译码模型设计出发, 主要采用深度学习方法、自动编译码结构实现文本、图片等语义通信系统. Xie 等 [6] 提出基于深度学习的文本传输语义通信系统 DeepSC 和适用于物联网的 Lite 分布式语义通信系统 L-DeepSC 提高系统容量和减少语义错误. Bourtsoulatze 等 [28] 提出基于深度学习的图片传输语义信源信道联合编译码设计. Weng 等 [29] 在 DeepSC 的基础上提出语音传输语义通信系统并证明其在低信噪比下具有良好的传输性能. Shi 等 [30] 提出一个面向语义保真度的通信框架并设计了一个面向语义保真度的音频传输实例.

在语义通信网络和协议层面上, Shi 等 [8] 提出了语义通信网络架构并总结了 5 个开放问题包括: 不同语境下语义多义问题、多模态语义协同感知分析问题、动态环境下的语义感知/识别和预测问题、语义知识库自学习/更新问题和语义通信网络安全问题. Popovski 等 [31] 提出在现有的协议结构上增加一个语义平面用于在协议栈的所有层信号处理和信息过滤并总结了语义平面的技术路线和应用场景.

3　面向智能体语义通信的架构

语义通信作为 “最高级智能体” 的人类之间的沟通方式, 对面向智能体的语义通信具有很好的借鉴意义. 在熟悉的人与人之间, 有时只需简单的眼神或者手势就能传达意思, 这其实是一种非常高效的沟通方式. 要达到这个目标需要具备两个前提条件: 首先, 收发双方需要具有上下文知识, 即对当前通信的背景和进程有所了解; 其次, 收发双方要根据当前的情景和上下文知识选择最合适的信息交互模式 (如编码方式、手势的选择也可看作是一种编码方式), 从而实现准确表达. 鉴于此, 语义通信的架构如图 1 所示.

3.1　上下文知识

上下文知识是发送方 (简写为 S) 和接收方 (简写为 D) 不断积累的 “知识”. 它主要包括两部分, 第 1 部分是 S 和 D 公共的知识, 它可以是在生产制造时赋予智能体设备的, 也可以是设备在使用周期内由控制节点定期分发给它们的; 第 2 部分是智能体根据其本身所处的环境和历史通信过程所学习到的知识. 因此, 通信双方的上下文知识可能相同, 但也可能不尽相同, 后者与自然通信模式更接近, 因为不同生物个体并非上下文知识完全一致才能有效可靠的交流 [4]. 在经典通信中, 基于 S 和 D 的上下文知识相同这一假设已开展了大量的研究, 如果能进一步扩展到允许 S 与 D 的上下文知识不同时,

将扩展智能体通信的适用场景.

3.2 发送方与接收方

在以 “达意通信” [2] 为目标的语义通信架构下, S 要根据当前的通信情景、所积累的上下文等背景知识决定发送什么内容给 D, 以及以什么样的方式 (如语义编码方式) 发送给 D. D 根据收到的语义编码和上下文知识等背景知识来恢复语义, 然后理解情景. 此外, 作为交互信息内容中重要组成部分的反馈信息会引入额外的通信开销, 但必要的反馈信息可以更新、扩充 S 和 D 的背景知识, 并可提升通信系统的有效性和可靠性. 根据观测通信环境、与接收方 D 通信积累的上下文内容等背景知识, 发送方 S 与接收方 D 通过权衡发送反馈信息的通信开销和可获得的收益将必要的反馈信息有效地发送给对方. 然而, 以上过程建立在 S 和 D 对所发送的内容有一定的智能认知基础上.

如果赋予 S 与 D 更高的智能和更多的自主权, 为了达到降低传输量或隐蔽通信的目的, S 与 D 之间就会形成语义通信的博弈. 一方面, S 会通过观测环境、与 D 交互积累的背景知识会预测: D 有什么不知道的、怎么发送信息 D 才能更好地理解自己的意图, 然后来决定发什么和怎么发给 D (即 S 会根据 D 的理解程度来调整发送的策略). 另一方面, 基于通信环境和与 S 交互积累的背景知识, D 解码的过程其实也是在解析 S 的语义是什么、S 希望 D 理解的情景是什么.

3.3 语义的编解码

传统的编译码方法假设通信符号是独立同分布, 所以每个符号编码后的码字是独一无二的, 否则译码端无法区分发送端发送的符号. 在传统的编译码基础上, 语义编译码拓展了一个新维度, 即语义维度. 通信符号的特征空间也在概率分布特征的基础上增加了语义特征. 语义特征 (例如通信的场景、目的, 通信内容的上下文、语境等) 可改变传输中的符号概率分布, 且传输中上下文符号间可能存在一定相关性. 将这些语义特征作为接收端 D 的背景知识, 利用符号间语义相关性可帮助接收端在较差通信环境下恢复错误甚至丢失的符号, 还可以帮助接收端在不同的上下文下区分相同码字的不同符号. 这也为减少传输量提供了一个新思路, 利用语义特征作为发送方 S 的背景知识优化码字分配过程, 可以实现高效的信息表达能力. 当对相同数据有效编码后所需的期望码长长度降低后, 可以节省传输开销以及存储开销, 达到 “言简意赅”. 因此, 基于语义的编译码设计以准确可靠达意和传输相同信息量占用更少的数据量为目标.

具体语义编译码中, 每个通信符号的语义特征可由 $\mathcal{K}$ 个不同语义特征构成, 即 $\boldsymbol{x}=(\omega_1,\omega_2,\ldots\omega_{\mathcal{K}})$ $(\omega_i\in\mathbb{R})$ 表示 $\mathcal{K}$ 维语义特征空间中的语义向量. 当所有符号语义向量都为互异的 one-hot 向量 (one-hot 向量中只能有一位取 1 其余取 0) 时就退化到传统通信中符号间相互独立的假设 (one-hot 向量与其他 one-hot 向量间是正交的). 符号间的语义特征关系可以由语义向量间的相关性表示. 第 4 节范例中详细介绍了一种接收端和发送端利用语义特征的编译码方法.

4 范例: 一种语义通信的实现方法

4.1 基本思路

为了实现高效通信达义, 我们可以基于语义进行压缩编码, 下述两种情况可以采用相同的编码: 无需区分的同义词、语义相差很大的词语. 前者是因为我们不必进行区分; 而后者是因为它们有极小的概率出现在同一语境下, 或者出现在同一语境时, 通过上下文语义等先验信息可以很容易地判断其所

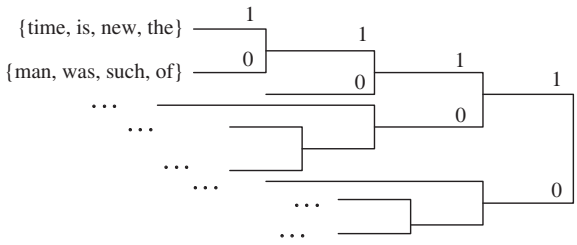

图 2　最优二叉树

Figure 2　The optimal binary code tree

对应的词语. 因此, 我们可以利用深度学习算法来学习上下文语义等相关特征, 从而进行最高效的语义编码和解码, 达到言简意赅的语义通信效果.

4.2　编码过程

本范例的语义编码中不要求每个单词对应一个编码码字, 只要不引起混淆, 一个码字可以分配给多个单词. 为了便于接收方有效译码以及增加通信的可靠性, 语义编码过程需要让相同码字所对应的不同单词之间的语义距离尽可能大. 考虑到不同词性的单词间的语义距离较大, 我们设计了基于词性的语义编码方法. 不失一般性, 第一, 将训练语料库中的所有单词 ($\mathcal{N}$ 个) 按其词性标记分成 4 个大类 (记为 $P=4$), 即名词类 η、动词类 v、形容词类 a 和其他类 o. 第二, 将每类字词按照日常中出现的频率由高到低降序排列. 第三, 将排好序的 4 类平行合并, 即每类中频率最高的字词组成叶子节点 $A_0=(\eta_0,v_0,a_0,o_0)$, 次高的组成下一个叶子 $A_1=(\eta_1,v_1,a_1,o_1)$. 以此类推, 将得到 M 个叶子节点 $A_i=(\eta_i,v_i,a_i,o_i)$, $i=0,\ldots,M-1$, 则 $M\leqslant\mathcal{N}$. 同时每个叶子节点的权重为该叶子节点所含的 4 个字词的频次之和. 举例而言, 在某个训练语料库中, 名词类的 'time' (出现 1597 次), 动词类的 'is' (10108 次), 形容词类的 'new' (1635 次) 和其他类的 'the' (69968 次) 分别在 4 类中出现频次最多, 那么叶子节点 A_0 为 {time, is, new, the}, 且其权重为 4 个词的频次和 83308. 第四, 自底向上构建最优二叉树, 即每次取权重最低的两个叶子节点合并生成新的节点, 并采用两个叶子节点的权重的求和值更新新叶子的权重. 此外, 给这两个要合并的节点按照权重的高低分别加标签 '1' 和 '0'. 第五, 将该合并过程逐次进行重复, 直到最后只剩下两个节点合并形成根节点, 最优二叉树生成. 树上 M 个节点都被分配了码字, 该码字是从根节点到该节点的标签相连, 所获的编码为无重复前缀码. 例如在图 2 的二叉树中从根节点到叶子节点 A_0 的标签相连为 1111, 即 time, is, new 和 the 这 4 个词被分配的码字均为 1111.

4.3　解码过程

一条消息中的每个字符都不是相互独立或无关的. 多个相邻字符组合在一起可以形成语义实体. 从这个意义上说, 当对分配给多个字符的某个码字译码时, 潜在的上下文语义关系可以帮助语义接收方从多个候选词语中推测出发送方发送的那个单词. 所以语义译码器不是一个一个地单独解码字符, 而是将连续相邻的字符作为上下文一起恢复. 假设一个长度为 n 的序列 $\boldsymbol{s}$ 表示为 $\boldsymbol{s}=(w_1,w_2,\ldots,w_n)$, $n\in\mathbb{Z}^+$. 由于序列 $\boldsymbol{s}$ 中第 i 位 ($i\in[1,n]$) 码字可以对应多个字符 w_i^k ($k\in\mathbb{Z}^+$), 序列 $\boldsymbol{s}$ 对应有多种排列组合的译码结果. 基于上下文的语义译码过程需要在所有可能的排列组合 $\boldsymbol{s}\in\mathcal{S}$ 中找到可能性最大

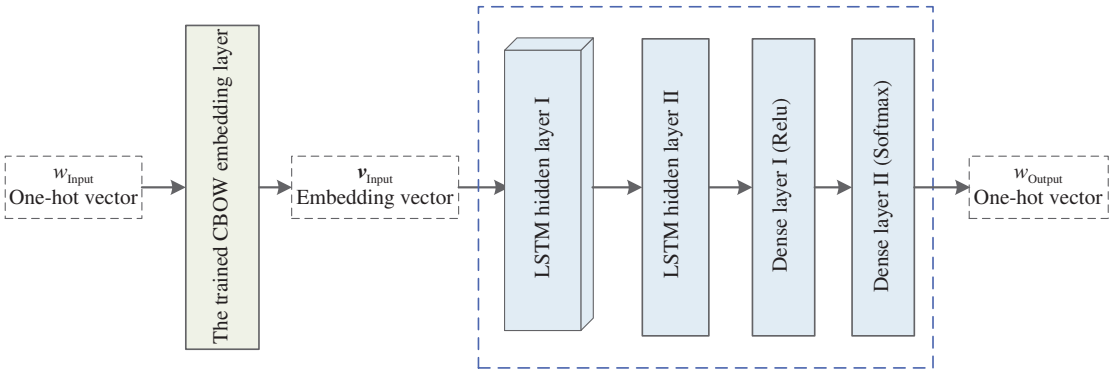

图 3 (网络版彩图) 基于 LSTM 的学习模型的架构

Figure 3 (Color online) The structure of the LSTM-based learning model

或联合概率 $\Pr(\boldsymbol{s}) = \Pr(w_1, w_2, \ldots, w_n)$ 最大的序列 $\boldsymbol{s}^*$, 可将其表示为

$$\boldsymbol{s}^* = \arg\max_{\boldsymbol{s}\in S} \Pr(\boldsymbol{s}). \tag{1}$$

为了求解该问题, 我们提出了一种基于上下文的语义译码算法. 首先, 设计深度学习模型对通信信息中的上下文关联特征进行学习; 然后, 将学习得到的特征提供给 N-gram 模型的公式 (3) 中得到联合概率 $\Pr(\boldsymbol{s})$ 值 ($\boldsymbol{s}$ 为一个长度为 n 的序列); 最后, 根据目标函数公式 (1), 利用动态规划算法最大化 $\Pr(\boldsymbol{s})$ 值, 实现从很多的候选序列 $\boldsymbol{s} \in \mathcal{S}$ 中寻找到最优解或最优序列 $\boldsymbol{s}^*$ 作为解译码结果输出. 下面具体介绍这 3 个过程.

4.3.1 联合概率 $\Pr(w_1, w_2, \ldots, w_n)$ 的计算

采用 N-gram 模型对联合概率建模, 该模型通过刻画字词的顺序来描述字词间的上下文关系. 因此, 联合概率 $\Pr(w_1, w_2, \ldots, w_n)$ 的求解过程是一个 Markov 链. 如式 (2) 所示, 式中每个字符的出现与前面历史字符都是相关的,

$$\Pr(\boldsymbol{s}) = \Pr(w_1, w_2, \ldots, w_n) = \Pr(w_1)\Pr(w_2|w_1)\cdots\Pr(w_n|w_1, w_2, \ldots, w_{n-1}). \tag{2}$$

然而, 随着字词间距离的增加, 距离越远的两个词的相关性逐渐降低. 因此, N-gram 的 Markov 假设是序列中的每个字符仅与前面 N 个历史字符是相关的, 可将联合概率公式 (2) 简化为

$$\Pr(\boldsymbol{s}) = \Pr(w_1, w_2, \ldots, w_n) = \prod_{i=1}^{n} \Pr(w_i|w_1, w_2, \ldots, w_{i-1}) \approx \prod_{i=1}^{n} \Pr(w_i|w_{i-N+1}, \ldots, w_{i-1}). \tag{3}$$

4.3.2 上下文语义关联特征的学习

基于上下文语义之间的历史相关性, 本小节设计多层 LSTM 网络来学习上下文语义关联特征 $\Pr(w_i|w_{i-N+1}, \ldots, w_{i-1})$, 该特征将作为求解联合概率 $\Pr(w_1, w_2, \ldots, w_n)$ 过程的已知输入.

图 3 显示了所设计的多层 LSTM 网络结构, 它由一个 CBOW (continuous bag-of-words) 模型 [32]2) 训练好的嵌入层作为输入层、3 个非线性隐藏层 (LSTM 隐藏层 I 和 II, Dense 层 I) 和一个输出

2) CBOW 模型为 3 层全连接神经网络, 主要用于生成词向量和计算字词之间相似度. 其训练好的输入层 – 隐藏层矩阵为嵌入层 (embedding layer). 该模型为自然语言程序设计任务中十分重要的词向量模型之一, 具体可参考文献 [32].

层 (Dense 层 II, 非线性激活函数为 Softmax) 组成. 多层 LSTM 神经网络输入为需要预测的中心词的周围几个词的 one-hot 向量[3)] $w_{\text{Input}} = [w_{i-\lceil L/2\rceil}, \ldots, w_{i-1}, w_{i+1}, \ldots, w_{i+\lceil L/2\rceil-1}]$, 多层 LSTM 神经网络的输出为预测目标函数的 one-hot 向量 $w_{\text{Onput}} = w_i$. 它的原理是利用中心词的前后 L 个词 $w_{\text{Input}} = [w_{i-\lceil L/2\rceil}, \ldots, w_{i-1}, w_{i+1}, \ldots, w_{i+\lceil L/2\rceil-1}]$ 来预测中心词 $w_{\text{Onput}} = w_i$. 具体而言, 网络的输入为中心词上下文的 one-hot 向量 w_{Input}, 它们经过嵌入层得到相对应的词向量或者特征向量 (embedding vector)[4)] $\boldsymbol{v}_{\text{Input}} = [\boldsymbol{v}_{w_{i-\lceil L/2\rceil}}, \ldots, \boldsymbol{v}_{w_{i-1}}, \boldsymbol{v}_{w_{i+1}}, \ldots, \boldsymbol{v}_{w_{i+\lceil L/2\rceil-1}}]$, 再输入到网络的多层隐藏层. 网络输出层的激活函数为 Softmax, 该函数将多个神经元的输出映射到 $(0, 1)$ 区间, 输出值即为概率 $\Pr(w_{\text{Output}}|w_{\text{Input}})$. 通过梯度下降方法训练使式 (4) 中多层 LSTM 网络的损失函数 E 最小化, 此时网络输出层即所求的基于上下文推导当前词的概率.

$$E = -\log \Pr\left(w_{\text{Output}}|w_{\text{Input}}\right). \tag{4}$$

多层 LSTM 网络允许递归网络学习多个步骤和创造长期记忆, 相比 CBOW 只有一层线性隐藏层, 多层 LSTM 网络有 3 个非线性隐藏层 (LSTM 隐藏层 I 和 II, 以及图 3 中的 Dense 层 I), 可以更好地学习复杂的上下文语义关联特征 $\Pr(w_i|w_{i-L+1}, \ldots, w_{i-1})$.

4.3.3 动态规划算法求解

在上文中, 我们用 N-gram 模型和多层 LSTM 网络模型分别刻画和求解序列中字词间上下文的相关性. 本小节利用动态规划算法求解全局最优的序列 $\boldsymbol{s}^*$ ($\boldsymbol{s}^* \in S$), 使得 $\Pr(\boldsymbol{s}^*)$ 在所有候选的解译序列中数值最大. 由于序列排列组合状态过多, 贪心算法和枚举遍历算法均无法快速得到该问题的全局最优解. 我们设计状态压缩的动态规划算法来求解目标函数得到全局最优. 因为该模型的目标函数可以分解成多个更小的子问题, 其中一些子问题是重叠子问题 (在递归算法中重叠子问题会被重复访问). 而动态规划算法将子问题的多个最优结果存储在一个表格中, 可以避免重复计算, 降低从大量潜在组合中寻找全局最优解的时间复杂度.

该算法从最小子问题开始, 然后将这些较小子问题的解组合起来, 不断得到更大子问题的解, 直到得到原最优问题的解. 对于一个长度为 n 的序列 $\boldsymbol{s} = (w_1, w_2, \ldots, w_n)$, $n \in \mathbb{Z}^+$, 假设算法中一个上下文窗口内含有上下文共 N 个单词, 即上下文窗口大小设为 N, $N \ll n$. 先考虑 $\boldsymbol{s}$ 序列中前 N 个字词最优组合, 再逐渐增加子问题的规模, 即考虑前 $N+1$ 个字词的组合、前 $N+2$ 个字词的组合, 逐渐递归一直到求解出长度为 n 的 $\boldsymbol{s}$ 序列的全局最优解.

具体来说, 首先考虑序列 $\boldsymbol{s}$ 一个上下文窗口大小的子问题, 即序列 $\boldsymbol{s}$ 前 N 个字词排列组合使得 $\Pr(w_1, w_2, \ldots, w_N)$ 概率值最大, 该最大概率值记为 $\mathcal{P}[(w_1^k, \ldots, w_N^k)]$, 具体求解过程如下所示:

$$\mathcal{P}[(w_1^k \ldots w_N^k)] \stackrel{\text{def}}{=} \max \sum_{i=1}^{N} \ln \Pr(w_i|w_{i-N+1}, \ldots, w_{i-1}). \tag{5}$$

下一步依次递归求解第 i 个子问题, $i = N+1, N+2, \ldots, n$. 最后直到递归求解 $\boldsymbol{s}$ 序列中全 n 个字词的最优组合 ($i = n$ 的情况), 即目标函数公式 (1) 的最优解.

每个子问题的求解过程需要用到上一个子问题多个最优的子序列概率值. 如图 4 所示, 第 i 个子问题 (即求解 $\boldsymbol{s}$ 序列中前 i 个字词能排列组成的最优子序列且假设最后 N 位的状态为图中实框所示, $i > N$) 的反向状态转移过程可以表征为选择 w_{i-N}^l 值使最后 N 位的状态为虚框所示的前 $i-1$ 长最

3) one-hot 向量为 "一位有效" 编码, 即对 $\mathcal{N}$ 个状态进行编码, 每个状态都是独一无二的. 在任意时候, 只有一位有效 (取 1), 其余位置取 0 的 $\mathcal{N}$ 维向量.

4) 词向量 (embedding vector) 就是用一个相对低维的向量对单词进行表征, 是单词的特征向量.

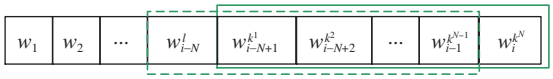

图 4 (网络版彩图) 第 i 个子问题求解的状态转移过程

Figure 4 (Color online) The state transition process of the ith subquestion

优子序列的概率值和由虚框上文推出下个词为 $w_i^{k^N}$ 的概率值之和的最大值, 状态转移过程如下所示:

$$\mathcal{P}[\mathcal{S}_i(k^1\cdots k^N)] = \max_{w_{i-N}^{l}}\{\mathcal{P}[\mathcal{S}_{i-1}(l\ k^1\cdots k^{N-1})] + \ln \Pr(w_i^{k^N}|w_{i-N}^{l}w_{i-N+1}^{k^1}\cdots w_{i-1}^{k^{N-1}})\}, \tag{6}$$

其中, 每个状态值为 i 位长序列 $\mathcal{S}_i$ 的概率值 $\Pr(S_i)$. 图 4 中的实框代表 i 位长的序列中最后一个上下文窗口内 (最后 N 位) 为 $w_{i-N+1}^{k^1}\cdots w_{i-1}^{k^{N-1}}w_i^{k^N}$ 组合, 即式 (6) 中的 $\mathcal{S}_i(k^1\cdots k^N)$. 虚框代表 $i-1$ 位长的序列中最后一个上下文窗口内 (最后 N 位) 为 $w_{i-N+1}^{l}\cdots w_{i-1}^{k^{N-1}}$ 组合, 即式 (6) 中的 $\mathcal{S}_{i-1}(lk^1\cdots k^{N-1})$.

5 语义编译码的评价指标

语义编译码的评价可以从有效性和可靠性两个方面考虑, 但其含义不同于经典的通信系统. 有效性主要是考察通信过程所传输的比特流是否减少, 可以通过编码长度来表示. 在可靠性方面, 传统的通信系统目的是准确地传输和接收通信符号或数据, 并且假设每个符号对于意思的表达同等重要和同一符号对于不同接收端的价值是相同的, 可见将误码率和误比特率作为衡量信息传输性能好坏的标准是合理的. 但是由于语义通信是以达意通信 (而不是符号的无误传输) 为目标, 在误码率和误比特率的基础上语义通信需要一个更合理地面向内容的语义度量来表征信息含义传输的能力. 举例来说, 假如将一个词解译为其同义词时, 从信息含义的角度出发传输是有效的, 但从误码率和误比特率角度出发发送端需要重传. 又比如将一个词解译成其近义词的语义损失和解译成其反义词的语义损失并不相同. 我们通过发送和解译出信息的语义相似度来考察其可靠性. 下面我们将分别给出评价指标的具体计算方法.

5.1 编码长度

动态码长 $\bar{\ell}^d$ 是指在实际应用过程中, 随着语义通信的进行, 动态测量所得到的整体码长长度, 它反映了在某一具体应用场景下, 利用语义通信所能达到的实时压缩效果.

5.2 语义相似度

随着 “感知智能” 到 “认知智能” 概念的提出, 以及图片/视频数据文本化描述 (image/video caption) 技术和多模态学习 (multi-modal machine learning) 技术的发展, 图像、视频等数据的语义相似度问题可转化为文本语义相似度问题, 文本语义相似度问题可以作为其他形式甚至全息数据语义相似度问题的基础. 此外, 文本数据相对由音频频谱序列向量/像素点向量所构成矩阵的音频/图片数据更为抽象, 难以仅通过简单字面距离度量模型比较数据间的差别. 以下度量文本符号间的语义相似度的 4 种方法中前 3 种方法可以在字词层次上比较发送和恢复信息的语义距离, 第 4 种方法可以在句子/段落层次上比较发送和恢复信息的语义距离.

5.2.1 基于边的相似度值 (edge-based similarity score)

基于边的相似度计算方法通过概念拓扑来衡量不同字词之间的距离. 它是衡量词的语义相似度的最基本和最直观的概念拓扑类方法之一 [33, 34]. 该方法通过测量两个单词在 WordNet 分类法的概念层

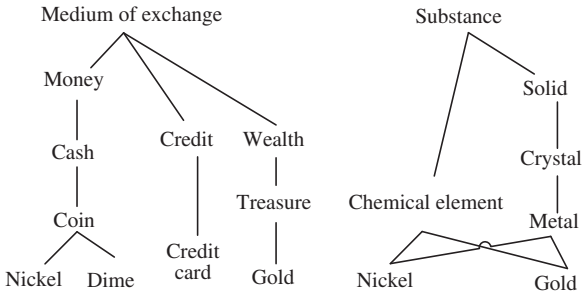

图 5　一个概念层次图的例子 [35]

Figure 5　An example of a taxonomy [35]

次图中的最短距离, 获得 0 到 1 范围内的语义相似度值. 在概念层次图中 (如图 5 [35] 所示), 两个词的词义相似度与两个词之间的最短边距离成反比. 两个词 w_1 与 w_2 的语义相似度可用下式计算:

$$\mathrm{sim}_e\left(w_1, w_2\right) = 2d_{\max} - \left[\min\left(c_1, c_2\right) \mathrm{len}\left(c_1, c_2\right)\right], \tag{7}$$

其中 $d_{\max}$ 代表着该概念层次图中最大的深度, c_1 和 c_2 分别取值为 $s(w_1)$ 和 $s(w_2)$. len 为 c_1 和 c_2 间最小路径长度, 且 $s(w)$ 在层次图中为词 w 词义的一组概念, 将距离度量转换为语义相似度度量的更详细信息请参考文献 [33].

5.2.2　基于词向量的相似度值 (word2vec similarity score)

由于概念层次图是在主观语法语义规则下建立的, 上述基于边的相似度计算方法高度依赖于主观知识或信息. 并且, 不在同一个概念层次图上或者不属于任何层次结构的两个字词无法衡量它们的相似性. 基于词向量的相似度计算方法与上述方法不同, 可以客观比较任何两个字词的相似性. 该方法通过使用预先训练的浅层神经网络来创建每个字词的特征向量 [36]. 最常见的基于词向量的相似度计算方法是计算字词特征向量之间的余弦相似度或软余弦相似度, 如下所示:

$$\mathrm{sim}_w\left(w_1, w_2\right) = \frac{\sum_{i,j}^{K} s_{ij} \boldsymbol{v}_i^{w_1} \boldsymbol{v}_j^{w_2}}{\sqrt{\sum_{i,j}^{K} s_{ij} \boldsymbol{v}_i^{w_1} \boldsymbol{v}_j^{w_1}} \sqrt{\sum_{i,j}^{K} s_{ij} \boldsymbol{v}_i^{w_2} \boldsymbol{v}_j^{w_2}}}, \tag{8}$$

其中 $\boldsymbol{v}^{w_1}$ 和 $\boldsymbol{v}^{w_2}$ 是 K 维特征向量, s_{ij} 表示特征 i 和特征 j 之间的相似程度. 更详细的信息请参考文献 [36].

5.2.3　混合相似度值 (hybrid-based similarity score)

基于词向量的相似度计算方法不依赖于主观认识, 但是基于边的相似度计算方法更能准确反映反义词和上下文字词间的关系. 因此, 综合两种方法的优点, 混合两种相似度计算方法提取语义相似度. 具体地, 首先利用基于边的相似度计算方法比较字词的语义相似度, 而对于那些基于边的相似度计算方法无法处理的字词例如动词和副词等, 则使用基于词向量的相似度计算方法生成特征向量来度量语义相似度.

5.2.4　METEOR 值

上述几种方法仅从字词层面上度量它们的语义相似度, 而有一些度量方法可以比较两个句子的语义相似度. BLEU (bilingual evaluation understudy) [37] 和 METEOR (metric for evaluation of translation

表 1 多层 LSTM 神经网络的设置

Table 1 The setting of the proposed LSTM neural network

Layer name	Units	Activation
Embedding	$\mathcal{N}$	Linear
LSTM I	256	Sigmoid Sigmoid Tanh Sigmoid
LSTM II	256	Sigmoid Sigmoid Tanh Sigmoid
Dense I	256	Relu
Dense II	$\mathcal{N}$	Softmax

表 2 CBOW 神经网络的设置

Table 2 The setting of the CBOW neural network

Layer name	Units	Activation
Embedding	$\mathcal{N}$	Linear
Projection	300	Linear
Dense	$\mathcal{N}$	Softmax

with explicit ordering) [38] 这两个评价指标在机器翻译和图像文本化描述中经常用来衡量两个句子的语义相似度. BLEU 通过综合比较两段文字中 n 元组词共同出现的占比来计算两段文字中字词的相似性, 反映出解译结果的忠实度和流畅度. 而 METEOR 在 BLEU 的基础上还考虑了同义词、单词的词形 (单复数、时态变化) 等.

6 仿真结果

该范例的仿真评估采用了根据系统性原则采集样本的标准语料库 Brown 语料库 [39], 该数据库收集超过 500 个文本文章涉及 15 个领域类别, 每个文本都超过 2000 字. 在实验中, 语义译码算法中涉及到的 CBOW 网络和多层 LSTM 神经网络的参数设置如表 1 和 2 所示.

6.1 有效性

仿真实验过程中按照词性将字符集内分成 4 类 $P = 4$, 用于训练的语料库中有 30632 个名词、10392 个动词、8054 个形容词和 4331 个其他类的字词. 表 3 和图 6 分别从平均码长和动态码长两个指标可以验证该范例的有效性. 平均码长 $\bar{\ell}$ 是指语义编码字符集内所有字符的码长的平均值定义, 反映了语义通信所能达到的平均压缩效果. 当字符集内所有字符个数为 $\mathcal{N}$, 码字 w_i 的长度为 $\ell(w_i)$, 码字 w_i 出现的概率为 $\Pr(w_i)$, 平均码长长度的公式可以表示为

$$\bar{\ell} = \sum_{j=1}^{\mathcal{N}} \Pr(w_j)\ell(w_j). \tag{9}$$

根据前文所提出的语义编码策略, 需要编码的状态数从 53409 下降到 30632. 将本文编码方法的码字平均码长计算结果列入表 3, 将其与 Huffman 编码方法进行比较. 对同一个语料库中字词编码的条件下, 本文编码方法比 Huffman 编码的码字平均码长长度减少了约 17.82%. 图 6 从动态平均码长 $\bar{\ell}^d$ 的角度验证了算法的有效性, 本文编码方法会比 Huffman 编码的动态平均码长长度更短, 而且差距会随着需要编码的字符数的增加而增大.

表 3　平均码长长度

Table 3　The average length of codeword per word

	The proposed coding	Huffman coding	Reduction
$\bar{\ell}$ (bits)	8.51907	10.36632	17.82%

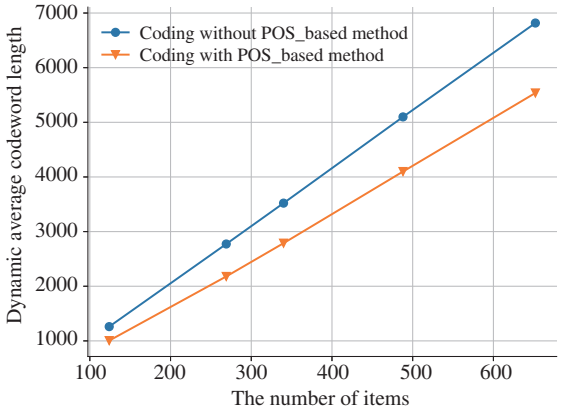

图 6　(网络版彩图) 动态平均码字长度

Figure 6　(Color online) Dynamic average codeword length

此外, 可从信息论角度证明本文编码方法较 Huffman 编码方法有更小的信源熵. Huffman 编码后信源熵为 $H_{\mathcal{N}} = -\sum_{j=1}^{\mathcal{N}} \Pr(w_j)\log\Pr(w_j)$. 范例方法编码后信源熵为

$$\begin{aligned} H_M &= -\sum_{j=1}^{M} \Pr(A_j)\log\Pr(A_j) = -\sum_{j=1}^{M}\left(\sum_{k=1}^{P}\Pr(w_j^k)\right)\log\left(\sum_{k=1}^{P}\Pr(w_j^k)\right) \\ &= -\sum_{j=1}^{M}\left(\Pr(w_j^1)\log\left(\sum_{k=1}^{P}\Pr(w_j^k)\right) + \cdots + \Pr(w_j^P)\log\left(\sum_{k=1}^{P}\Pr(w_j^k)\right)\right) \\ &\leqslant -\sum_{j=1}^{M}\sum_{k=1}^{P}\left(\Pr(w_j^k)\log\Pr(w_j^k)\right) = H_{\mathcal{N}}, \end{aligned} \tag{10}$$

其中 $M = \lceil N/P \rceil$. 当 $0 < x < 1$ 时 $\log(x)$ 单调递增, 且 $-\log(\sum_{k=1}^{P}\Pr(w_j^k)) \leqslant -\log(\Pr(w_j^k))$.

6.2　可靠性

在上下文语义译码过程中, 影响译码可靠性的两个关键参数为: (1) 上下文窗口大小 N, 即动态规划算法求解过程中最小子问题的大小, 用于基于上下文信息的译码过程; (2) 特征窗口大小 L, 即提取上下文特征的学习算法在训练过程中的训练窗口大小, 用于学习 L 个上下文间的语义关联特征.

如图 7 所示, 只有当上下文窗口大小 N 等于或大于学习算法中特征窗口大小 L 时, 语义相似度得分才可以达到峰值并保持稳定. 随着上下文窗口 N 的增加, 语义相似度得分会增加. 随着特征窗口大小 L 的增加, 语义相似度得分也会提高. 在 4 种评价指标中, 基于字词语义的评价指标的得分一般比基于句子的评价指标 METEOR 的分值高. 因为评价句子的相似性除了需要计算单词同义词匹配外, 还需要考虑句子的连续有序问题. 此外, 基于字词语义的评价指标中, 混合相似度方法和基于词向量的相似度方法的评价指标得分比基于边的评价指标得分高. 原因是基于边的方法高度依赖于在主观语法语义规则下建立的概念层次图, 而没纳入层次结构图的字词无法计算其相似性被认为译码错误.

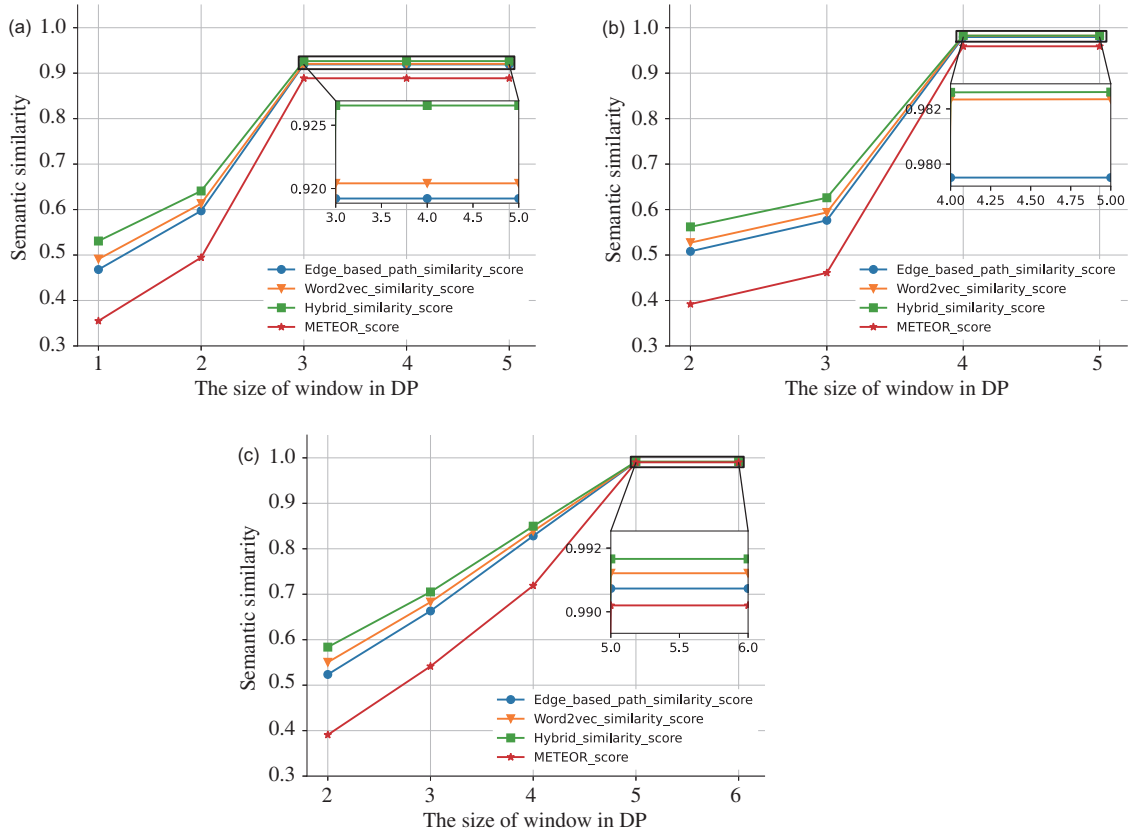

图 7　(网络版彩图) 特征窗口大小 (a) $L=3$, (b) $L=4$, (c) $L=5$

Figure 7　(Color online) The size of the feature window is (a) $L=3$, (b) $L=4$, and (c) $L=5$

7　小结

语义通信基于通信双方的上下文知识和对语义的理解力实现信息含义精准传输, 可显著提升通信的效率, 特别适用于智能化程度较高的设备之间进行通信, 也极有可能对以香农信息论为基础的现代通信系统带来革命性的变化, 可以使通信系统跃升到更高层次. 在本文中, 我们提出了一种语义通信的架构并给出了一个范例, 提出了能反映语义通信有效性和可靠性的几个指标计算方法. 整体而言, 未来语义通信具有很多理论和技术上的问题需要突破. 在根本性基础理论方面包括: 语义信息、语义噪声的表征和度量方法、类似香农限的语义熵定义以及容量计算方法、跨模态下的语义信息的压缩极限定义等 (文字、音频和图像模态等); 亟需解决的技术问题主要包括: 设计信源、信道语义编译码方案提高在带宽有限、信噪比较低或误码率、误比特率较高的不理想通信环境中传输文本、图片、声音等多模态信息的有效性、准确性和安全隐蔽性; 设计基于数据重要性的语义协议或对协议栈的所有层进行语义控制减少跨模态信息传输以及面向任务的服务中的协议及语义开销, 使其更有效地利用通信和计算资源且更适用于物联网、智能体通信等新兴应用等. 不过随着通信技术和人工智能技术研究的不断推进, 人们对通信本质的认识愈加深刻, 我们有理由期待语义通信将会有更广阔的发展前景.

参考文献

1　Shannon C E. A mathematical theory of communication. Bell Syst Tech J, 1948, 27: 379–423

2 Shi G M, Li Y Y, Xie X M. Semantic communications: outcome of the intelligence era. Pattern Recogn Artif Intell, 2018, 31: 91–99
3 Weaver W. Recent contributions to the mathematical theory of communication. ETC Rev Gen Semant, 1953, 10: 261–281
4 Sudan M. Communication amid uncertainty. In: Proceedings of IEEE Information Theory Workshop, Lausanne, 2012. 158–161
5 Bao J, Basu P, Dean M, et al. Towards a theory of semantic communication. In: Proceedings of the 1st International Workshop on Network Science, West Point, 2011. 110–117
6 Xie H Q, Qin Z J. A lite distributed semantic communication system for Internet of Things. IEEE J Sel Areas Commun, 2021, 39: 142–153
7 Goldreich O, Juba B, Sudan M. A theory of goal-oriented communication. J ACM, 2012, 59: 1–65
8 Shi G M, Xiao Y, Li Y, et al. Semantic networking for the intelligence of everything. Chinese J Int Thing, 2021, 5: 1–11
9 Xylomenos G, Ververidis C N, Siris V A, et al. A survey of information-centric networking research. IEEE Commun Surv Tut, 2014, 16: 1024–1049
10 Koponen T, Chawla M, Chun B, et al. A data-oriented (and beyond) network architecture. In: Proceedings of ACM SIGCOMM, Kyoto, 2007. 181–192
11 Gong S M, Wang P, Huang J W. Robust performance of spectrum sensing in cognitive radio networks. IEEE Trans Wirel Commun, 2013, 12: 2217–2227
12 Chen R L, Park J M, Reed J H. Defense against primary user emulation attacks in cognitive radio networks. IEEE J Sel Areas Commun, 2008, 26: 25–37
13 Musolesi M, Mascolo C. CAR: context-aware adaptive routing for delay-tolerant mobile networks. IEEE Trans Mobile Comput, 2009, 8: 246–260
14 Yau K A, Komisarczuk P, Teal P D. Context-awareness and intelligence in distributed cognitive radio networks: a reinforcement learning approach. In: Proceedings of Australian Communications Theory Workshop (AusCTW), Canberra, 2010. 35–42
15 Yuan Z, Han Z, Sun Y L, et al. Routing-toward-primary-user attack and belief propagation-based defense in cognitive radio networks. IEEE Trans Mobile Comput, 2013, 12: 1750–1760
16 Wang W, Kwasinski A, Han Z. A routing game in cognitive radio networks against routing-toward-primary-user attacks. In: Proceedings of IEEE Wireless Communications and Networking Conference (WCNC), Istanbul, 2014. 2510–2515
17 Lorenzo B, Kovacevic I, Gonzalez-Castano F J, et al. Exploiting context-awareness for secure spectrum trading in multi-hop cognitive cellular networks. In: Proceedings of IEEE Globecom Workshops, San Diego, 2015. 1–7
18 Yau K A, Komisarczuk P, Teal P D. A context-aware and intelligent dynamic channel selection scheme for cognitive radio networks. In: Proceedings of the 4th International Conference on Cognitive Radio Oriented Wireless Networks and Communications, Hannover, 2009. 1–6
19 Yau K A, Komisarczuk P, Teal P D. Achieving Context Awareness and Intelligence in Cognitive Radio Networks using Reinforcement Learning for Stateful Applications. Technical Report ECSTR10-01, 2010
20 Yau K A, Komisarczuk P, Teal P D. Achieving context awareness and intelligence in distributed cognitive radio networks: a payoff propagation approach. In: Proceedings of IEEE Workshops of International Conference on Advanced Information Networking and Applications, Singapore, 2011. 210–215
21 Yau K A, Komisarczuk P, Teal P D. Learning mechanisms for achieving context awareness and intelligence in cognitive radio network. In: Proceedings of the 36th Conference on Local Computer Networks, Bonn, 2011. 738–745
22 Vosoughi A, Cavallaro J R, Marshall A. A context-aware trust framework for resilient distributed cooperative spectrum sensing in dynamic settings. IEEE Trans Veh Technol, 2017, 66: 9177–9191
23 Clark D, Partridge C, Ramming J, et al. A knowledge plane for the internet. In: Proceedings of ACM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, Germany, 2003. 3–10
24 Wu D P, Li Z J, Wang J P, et al. Vision and challenges for knowledge centric networking. IEEE Wirel Commun, 2019, 26: 117–123
25 Carnap R, Bar-Hillel Y, et al. An Outline of A Theory of Semantic Information. RLE Technical Reports 247, 1952
26 Juba B, Sudan M. Universal semantic communication I. In: Proceedings of ACM International Symposium on Theory

of Computing, Victoria, 2008. 123–132
27 Güler B, Yener A, Swami A. The semantic communication game. IEEE Trans Cogn Commun Netw, 2018, 4: 787–802
28 Bourtsoulatze E, Burth Kurka D, Gunduz D. Deep joint source-channel coding for wireless image transmission. IEEE Trans Cogn Commun Netw, 2019, 5: 567–579
29 Weng Z Z, Qin Z J. Semantic communication systems for speech transmission. 2021. ArXiv:2102.12605
30 Shi G M, Gao D H, Song X D, et al. A new communication paradigm: from bit accuracy to semantic fidelity. 2021. ArXiv:2101.12649
31 Popovski P, Simeone O. Start making sense: semantic plane filtering and control for post-5G connectivity. 2019. ArXiv:1901.06337
32 Rong X. word2vec parameter learning explained. 2014. ArXiv:1411.2738
33 Majumder G, Pakray P, Gelbukh A, et al. Semantic textual similarity methods, tools, and applications: a survey. Comput Sist, 2016, 20: 647–665
34 Jeong S, Yim J, Lee H, et al. Semantic similarity calculation method using information contents-based edge weighting. J Int Serv Inform Secur, 2017, 7: 40–53
35 Gomaa W, Fahmy A. A survey of text similarity approaches. Int J Comput Appl, 2013, 68: 13–18
36 Lastra-Díaz J J, Goikoetxea J, Taieb M A H, et al. A reproducible survey on word embeddings and ontology-based methods for word similarity: linear combinations outperform the state of the art. Eng Appl Artif Intell, 2019, 85: 645–665
37 Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, 2002. 311–318
38 Banerjee S, Lavie A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Ann Arbor, 2005. 65–72
39 Lv C Y, Liu H H, Dong Y X. An efficient corpus based part-of-speech tagging with GEP. In: Proceedings of the 6th International Conference on Semantics, Knowledge and Grids, Beijing, 2010. 289–292

Semantic communication for intelligent devices: architectures and a paradigm

Yichi ZHANG[1], Ping ZHANG[2], Jibo WEI[1], Haitao ZHAO[1*], Jun XIONG[1] & Jiao ZHANG[1]

1. *College of Electronic Science and Technology, National University of Defense Technology, Changsha* 410073*, China*;
2. *State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing* 100876*, China*
* Corresponding author. E-mail: haitaozhao@nudt.edu.cn

Abstract Semantic communication focuses on the accurate transmission of meanings rather than data or symbols. It can achieve more efficient information interaction with less traffic, which is a promising communication mechanism for future intelligent devices. In this paper, we summarize existing research closely related to semantic communication. And then we propose a context-aware semantic communication architecture, explaining the functions of each module within it. We also introduce six metrics to evaluate the performance of semantic communication from the aspects of effectiveness and reliability. Finally, we present an example to illustrate the realization of semantic communication, which is proven effective in reducing data traffic compared to traditional communication mechanisms.

Keywords semantic communication, intelligent communication, semantic similarity, semantic coding, semantic decoding