Semantic Communication for Edge Intelligence: Theoretical Foundations and Implications on Protocols

Zoran Utkovski, Andrea Munari, Giuseppe Caire, Johannes Dommel, Pin-Hsun Lin, Max Franke, André C. Drummond, and Sławomir Stańczak

Abstract

Semantic communication has recently attracted considerable attention, mainly motivated by the trend of developing "task-oriented" communication solutions that tailor resource consumption to the task at hand. Despite the general intuition that semantic communication may contribute to more efficient system design, there have been only a few concrete attempts to implement aspects of it in practice. To help bridge this gap, in this paper, we revisit the theoretical foundations of semantic communication and address the possible implications on the protocol and system design. The focus is on two perspectives of semantic communication: (i) a goal-oriented perspective, which unifies aspects of traffic generation, communication, and control, with emphasis on the definition of appropriate semantic-aware metrics, and (ii) a semantic operability perspective, which extends the notion of data exchange through standardized interfaces (interoperability), to include the meaning or, more generally, the significance of data. We discuss applications of the concepts in scenarios such as robotic control and health monitoring.

Introduction

In an article on Shannon's "Mathematical Theory of Communication," Weaver noted [1] "In communication there seem to be problems at three levels: technical, semantic, and influential. The technical problems are concerned with the accuracy of transference of information from sender to receiver. [...] The semantic problems are concerned with the interpretation of meaning by the receiver, as compared with the intended meaning of the sender. [...] The problems of influence or effectiveness are concerned with the success with which the meaning conveyed to the receiver leads to the desired conduct on his part." As a matter of fact, Shannon's information theory deals principally with the "technical problem" of representing (source coding) and reliably delivering (channel coding) information, where the latter is defined regardless of meaning and/or effectiveness.

Following up on the initial remarks by Weaver, several researchers in the past have tried to bring "semantics" back into the picture. Early work by Carnap and Bar-Hillel [2] models semantic information in terms of logical probability. A different view is taken in [3], which argues that the notion of meaning, as well as the adequacy and truthfulness of models, can be explained using the toolbox of entropy-based theories of information (i.e., Shannon information theory, Gibbs' statistical mechanics, and Kolmogorov complexity), without the explicit need to develop a separate theory of semantic information.

More recently, "semantic information and communication" have re-emerged in the context of the next generation communication networks [4]. The general belief is that future information processing systems will be "task-oriented," and semantic communication may provide a conceptual framework for the unification of the process of data generation, transmission, and usage, with the aim of reducing overhead and energy consumption. However, despite the intuition that some major gains should be possible, putting these ideas into practice is still challenging.

Zoran Utkovski and Johannes Dommel are with the Fraunhofer Heinrich Hertz Institute (HHI), Germany.

Sławomir Stańczak is with Fraunhofer Heinrich Hertz Institute (HHI), Germany and the Technical University of Berlin, Germany.

Andrea Munari is with the German Aerospace Center (DLR), Oberpfaffenhofen-Wessling, Germany.

Giuseppe Caire and Max Franke are with the Technical University of Berlin, Germany.

Pin-Hsun Lin and Andre Drummond are with the Technical University of Braunschweig, Germany.

ISSN: 2576-3180

Digital Object Identifier: 10.1109/IOTM.001.2300167

Semantic Communication: the Search for Meaning

Probably the most concrete application of the "semantic communication" framework up to date is related to the introduction of new "semantic-aware" metrics to complement conventional ones such as throughput, reliability or latency. An example is provided by the notion of "information freshness" that has emerged as a proxy to capture the performance of wireless systems that monitor the state of one or multiple sources, e.g., sensing a physical process. In this context, established approaches deal with age of information (AoI) and its generalizations (see [5] for an overview). Relevant examples are offered by IoT applications, e.g., for industrial and environmental monitoring, by asset tracking, as well as by cyber-physical systems, where proper actuation and control decisions shall be made relying on these newly-introduced metrics. Another direction has recently appeared in the context of robust data transmission, where deep learning is used to map the source data into feature space. At the decoder side, the channel output is mapped back into the feature space such that the semantic content of the data is preserved. This approach is, in fact, an instance of joint source-channel coding (JSCC), with the novelty that the source-channel maps are learned by training suitably designed neural networks. While JSCC may indeed provide some advantage in terms of end-to-end performance for specific cases of sources and channels, we note that the JSCC paradigm breaks the fundamental principle of "separation" between system protocol layers, which is the driving principle according to which legacy communication systems are designed.

Criticism and Added Value of the Semantic Framework

Based on these remarks, two critical questions arise, which inspire the remainder of our discussion:

C1: *Is there more to semantic communication than "defining semantic-aware metrics" that are different from conventional metrics such as packet loss ratio, delay, and similar?*

C2: *Is the predication of large system gains based on a revival of joint source-channel coding, although re-interpreted with the preservation of semantic content as the main objective, novel and transformative?*

In light of these critical questions, we will discuss two perspectives on semantic communication.

In relation to the first question (C1), the focus will be on the *goal-oriented* interpretation of semantic communication, roughly defined as "the provisioning of the right and significant piece of information to the right point of computation (or actuation) at the right point in time" [6]. The rationale is that the goal-oriented approaches that incorporate semantic metrics can be more effective than traditional system-design approaches based on the optimization of conventional performance metrics, as they are concerned with the relevance of the information content being transmitted for the purpose of achieving the goal (for example meeting a certain control objective), rather than with the utilization of the wireless channel.

With respect to the second question (C2), we will focus on *semantic interoperability*, which extends the notion of interoperability (i.e. data exchange through standardized interfaces), to include the meaning or, more generally, the significance of data. This interpretation includes aspects of *generative modelling* and, respectively, the communication of generative models over wireless channels. From an information-theoretic perspective, some of these aspects are closely related to the problem of *remote source coding* (see, e.g., [7]), which also incorporates the well-known information bottleneck (IB) formulation (Tishby *et al.* [8]) as a special instance. We note that the IB principle has already been applied in the context of semantic communication (see, e.g., [9]).

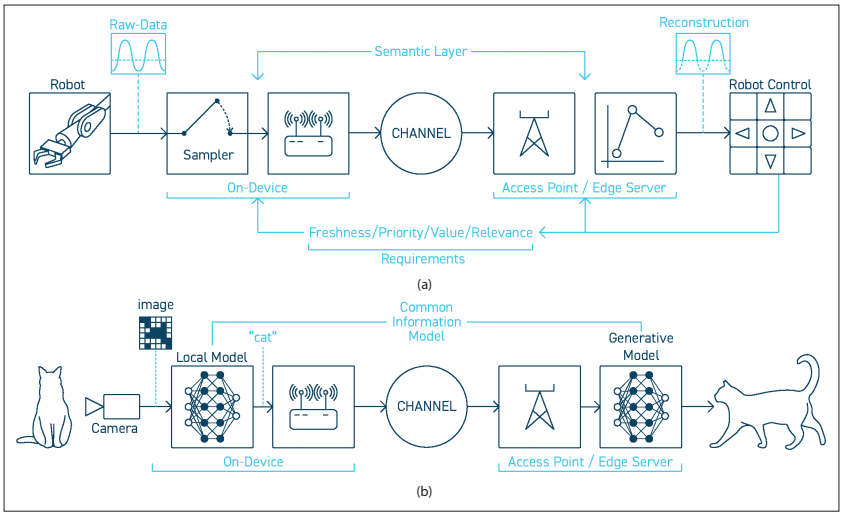

FIGURE 1. Two perspectives on semantic communication: a) goal-oriented communication: a semantic-aware architecture for robotic control that considers a unification of the processes of information generation, communication, and usage; b) semantic operability: communicating a semantic concept corresponds to communicating the corresponding generative model. Once a concept is communicated, the destination has the ability to generate instances of that concept.

Two Perspectives on "Semantic Communication" (and Their Implications)

Goal-Oriented Communication

In the spirit of this discussion, goal-oriented communication approaches the communication problem by focusing on efficient and effective exchange of information contributing to the realization of a desired goal. Besides a proper definition of a communication goal, this approach also requires related metrics that capture the notion of data significance.

The first steps in this direction were taken in the domain of vehicular communications, with the introduction of the AoI. The metric is simply defined as the difference between the current time and the time-stamp of the last update from the source of interest that is available at the destination and is thus agnostic of the actual information being transferred. However, AoI may capture some fundamental trade-offs and provide optimization criteria for goal-oriented communication. For instance, in a tracking problem, the deviation in position can be shown to be a simple linear function of the AoI. Moreover, in some control systems, the mean square error in state estimation, as well as the uncertainty at the receiver on the state of a tracked source, are, under proper conditions, non-decreasing penalty functions of age. Notably, the design insights that can be derived leaning on AoI may significantly deviate from those prompted by traditional metrics.

Going beyond AoI, an example of more advanced metric is offered by the age of incorrect information (AoII) (see [5]), evolving as a function of the AoI only when the monitored process is actually changing, and accounting for no penalty otherwise. Similarly, in the perspective of goal-oriented communication, the availability of fresh knowledge may only be important at times in which decisions are made, as captured by the query AoI (see [6] and references therein for an overview).

How to Integrate Control Aspects in the Goal-Oriented Communication Framework?: From an information-theoretic perspective, it is not clear how aspects of goal-oriented communication relate to networked control systems with feedback control loops over communication channels, as studied traditionally in control theory. Earlier works by Tatikonda, Sahai, and Mitter study the fundamental relationship between control and communication problems (see, e.g., [10])1 The main message is that Shannon's classical notion of capacity is not enough to characterize a noisy communication channel if the channel is intended to be used as part of a feedback loop to stabilize an unstable system. Instead, another sense of capacity (parametrized by reliability) called "anytime capacity" is necessary for the stabilization of the system.

We postulate that these aspects may be relevant to the establishment of a unified framework that integrates aspects of data generation, communication, and control. In this regard, a starting point may be provided by recently introduced metrics such as the value of information (VoI) [11], which systematically captures the semantics of data by estimating the relevance of the available data samples to the point of computation. From a control theory perspective, the VoI emerges as a solution to the rate-regulation tradeoff between the communication rate and the regulation cost, with an event trigger (collocated with the sensor) and a controller (collocated with the actuator) as two distributed decision makers. A practical example considering the problem of stabilization of an inverted pendulum on a cart was provided in [11], showing that the system under the triggering policy designed based on the VoI was able to achieve a much better regulation quality than a system with a conventional periodic triggering policy. While the above studies provide a good starting point, further research is needed that will expand the theoretical frameworks developed therein to more complex classes of systems, thus unifying aspects of semantic communication and control theory in more general settings. Another aspect that should be addressed is the potential impact that the adoption of these concepts has on the design of communication protocols. We discuss this issue later.

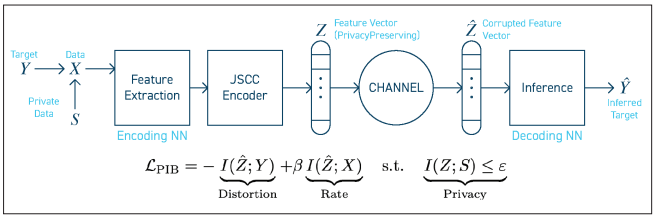

FIGURE 2. A semantic-aware architecture for device-edge co-inference based on the IB formulation with privacy constraints.

Semantic Interoperability

In addition to the goal-oriented perspective, a somewhat different perspective on semantic communication is conceivable — based on the concept of *semantic interoperability*, which can be defined as the ability of two or more agents to exchange and understand each other's data correctly. Interoperability is classically thought of solely in terms of data exchange (through standardized interfaces), but semantic interoperability extends this notion to include the meaning of data, and thus a data model translation becomes an essential part of the problem. In terms of interoperability, models and algorithms can be described using ontologies, which establish a formal and explicit digital specification (conceptualization) of a domain of interest. Within this framework, ontology is a means of defining semantically relevant descriptions, which can be used to interpret data, i.e., to match it to an appropriate information model.

From this general idea, one can postulate that communicating concepts corresponds to communicating models, and models are concretely defined as tools to reproduce an arbitrary number of instances of a given concept at the destination, i.e., sampling from a distribution defined on the manifold of the concept realization. As an example, suppose that both the sender and the receiver are familiar with the concept of "cat" (i.e. they share a common information model). According to the above said, to convey semantic information associated with the "object of interest" (i.e., cat), the sender can simply send the (suitably encoded) word `cat`. In particular, this enables the receiver to sample from a hypothetical distribution of existing cat images, and retrieve a random image of a cat. While the distortion (in terms of pixel-by-pixel distance, with respect to some suitable metric) may be very large, the semantic distortion is zero, since the receiver understands exactly what the transmitter wanted to convey. This, in essence, corresponds to the intuitive idea of human-to-human semantic communication.

Demystifying the Information Bottleneck: Under the assumption of a shared information model, semantic coding takes on the form of a remote source coding problem (see, e.g., [7]) where semantic information and the corresponding data realization are modeled as a pair of jointly distributed random variables (Y, X), where Y denotes the semantic label (e.g., "cat") and X denotes the corresponding sensory data (e.g., a corresponding image of a cat). In this context, the encoder produces a representation Z of the data X that aims to capture the most "relevant information" regarding X. This viewpoint in the context of semantic information has been recently proposed in [12]. It is apparent from the above that in the presence of a *common information model*, semantic information and communication may be interpreted as being an extreme form of data compression that aims at eliminating as much as possible the irrelevant features of the data, while preserving the "meaning" (in this case, the semantic class). In particular, when the representation variable Z is a probability distribution over the alphabet $\mathcal{Y}$ of the semantic label Y and the distortion measure between Y and Z is "log-loss," then remote source coding coincides with the information bottleneck problem [8].

In the *absence of a common information model*, on the other hand, the intended receiver (destination) is a blank slate that needs to be instructed. In that case, we need to convey the semantic concept such that the destination effectively learns the (new) concept. Hence, the idea is that communicating the semantic concept corresponds to communicating the corresponding generative model. The "ground-truth" generative model (unknown) can be approximated to a desired degree of accuracy by a deep neural network (DNN), referred to as generator. We note that this paradigm is intimately related to the concept of *universal coding*, which is central in information theory. In fact, there is a strong correspondence between learning the generative model from a data set and the universal compression of the data set itself.

Edge/On-Device Intelligence and Semantic Communication: A Perfect Match?

The principle of semantic interoperability is fairly general and can be applied to more complex ontologies, not just formed by sets of semantic classes, but also by their causal and spatial relationships. The ultimate goal is to define semantic data encoding targeted and optimized for given tasks. For example, if the machine's task consists of identifying the shape of objects or generating annotation of images, there is no reason to use standard image coding (optimized for human perception). Instead, the encoder at the source can be optimized according to the task and achieve the same end-to-end performance with a much smaller transmission or storage data rate. These aspects are of particular relevance in relation to the trend to migrate certain AI functionalities from the cloud to the wireless edge, with the aim of performing inference tasks closer to the end users. In the following, we provide an example of a semantic-aware architecture for edge inference.

A Semantic Architecture for Edge Intelligence

Consider a device-edge co-inference system as illustrated in Fig. 2. In this architecture, the device and the edge server cooperate to perform a certain task (e.g. image classification/object detection) relying on neural networks. In Fig. 2, X and Y correspond to the input data and the target random variable, respectively. The encoded feature vector, the received feature vector, and the inference result are respectively represented by random variables $Z, \hat{Z}$ and $\hat{Y}$.

The on-device neural network (NN) learns in a joint fashion how to extract the *task-relevant feature* from the raw input X, and to map the feature values to the channel input symbols Z. The server-based NN processes the received and corrupted (by the communication channel) feature $\hat{Z}$, and outputs the inference result $\hat{Y}$. In this context, the IB framework can be applied to decrease the communication overhead by retaining in Z only the "most relevant" information for the task in question. As such, IB formalizes a rate-distortion trade-off between the informativeness of the encoded feature and the inference performance. As in practice the mutual information terms in the IB formulation are intractable for DNNs with high-dimensional features, one can leverage the variational approximation, known as variational information bottleneck (VIB), to devise a tractable upper bound on the objective function, which can be then minimized via Monte Carlo gradient estimation.

In this context, we also emphasize the potential relevance of *neuromorphic computing* for efficient semantic-aware signal processing (see, e.g., [13]). Neuromorphic computing is particularly attractive for edge inference applications due to its energy efficiency and native support for event-driven processing. As a proof of concept, we present some initial experiments for a classification task with the Neuromorphic-MNIST (N-MNIST) dataset, a spiking version of the original MNIST dataset. In the context of the architecture in Fig. 2, we consider a neuromorphic implementation where the on-device semantic encoding is performed by a spiking neural network (SNN), and the communication is performed by using impulse-radio-based transmis-

sion. As a result, in the optimization framework, we resort to the *directed information bottleneck* formulation (DIB) introduced in [13]. Considering a binary symmetric channel with cross-over probability p, we test the classification accuracy for a fixed communication budget, quantified by the bit-length of the latent representation $\hat{Z}$. Table 1 summarizes the performance for various channel cross-over probabilities, under a communication budget of $k = 256$ and $k = 128$ bits respectively.

N-MNIST	$p = 0$	$p = 0.05$	$p = 0.1$	$p = 0.15$	$p = 0.2$	$p = 0.25$
256-bit SSCC	98.03	98.02	97.05	80.9	58.89	36.71
256-bit JSCC	98.27	**98.16**	97.97	97.58	96.26	88.75
256-bit S-VDIB	**98.4**	98.13	**97.98**	**97.69**	**97.02**	**96.01**
128-bit SSCC	96.9	96.8	96.3	87.7	72.54	56.22
128-bit JSCC	98.14	98	97.54	96.7	95.03	85.83
128-bit S-VDIB	**98.15**	**98.1**	**97.7**	**97.34**	**96.79**	**95.48**

TABLE 1. Comparison of models' accuracy (in %) for different channel parameters.

As a validation of the concept, we observe that the semantic-aware scheme implemented with neuromorphic processing, with learning based on the variational DIB (denoted as S-VDIB in Table I), provides competitive performance compared with a state-of-the-art joint source-channel coding scheme (JSCC), as well with and a conventional separate source-channel coding (SSCC), in the investigated regime.

PHY-Layer Security/Privacy in Semantic Communication

Notably, physical layer security and privacy can be integrated within the IB framework, relying on the formulations of privacy funnel (PF), and latent variable secrecy (LVS) as elaborated in the following examples.

Consider first the Markov chain $S - X - \hat{Z} - \hat{Y}$ in Fig. 2. From the PF viewpoint, we can interpret S, X, and $\hat{Z}$ as the private data, observed data, and displayed data, respectively. The problem of PF is to maximize the mutual information between X and $\hat{Z}$ such that the privacy leakage between S and $\hat{Z}$ is below a threshold value which represents the privacy constraint while keeping the original Markov chain constraints. Therefore the goal of PF is to share as much information between the observed data X and the displayed (extracted) data $\hat{Z}$ as possible, while bounding the leakage of the private data S to $\hat{Z}$. More formally, as illustrated in Fig. 2, the minimization of the functional $\mathcal{L}_{\text{IB}}$ in the original IB framework is now performed subject to privacy constraints (therefore $\mathcal{L}_{\text{PIB}}$).

A related problem is the LVS, which is based on the Markov chain $S - X - Z - (\hat{Y}, W)$ where W represents an additional channel output observed by an eavesdropper.

In the LVS context, the private data S has to be kept secret from the eavesdropper observing W. Formulated in the IB paradigm, this means that the mutual information between X and $\hat{Y}$ is maximized under the constraint that the information leakage between S and W is smaller than a prescribed threshold.

In general, the joint design of semantic communications and physical layer security/privacy must include an investigation of operational implications of the underlying privacy and confidentiality measures that quantify the resilience to specific attacks. In case the level of protection is insufficient, the joint design should be based on privacy and confidentiality measures that offer stronger protection, such as differential privacy, respectively semantic security.

An Application Example

Consider a health monitoring system based on the architecture in Fig. 2 (as illustrated in Fig. 3), designed to remotely assess a patient's health status (Y) by means of vital parameter measurements (X) obtained via biomedical sensors. Due to the high requirements on privacy and data protection, parts of the patient data (e.g. sex, age) is required to be private (S). As a specific example consider a continuous electrocardiography (ECG) measurement to detect the risk of heart attacks at an early stage. For this, a general indicator is the occurrence of arrhythmias in the ECG signal. Thus, instead of transmitting the raw data (X) via the wireless channel, a semantic-aware encoder jointly incorporates pre-processing (e.g., feature extraction/arrhythmia detection) and channel coding according to the specific monitoring task. Thus, the encoder produces transmit symbols (Z) that contain sufficient information for the receiver to infer the task-relevant information, which can be, e.g., a statistical measure related to the occurrence of abnormalities of the heart's rhythm. Hence, the receiver infers the health status while preserving the private part of the message (formulated as a predefined privacy leakage constraint).

Implications on Protocol/Network Design

In the context of the above discussions, an important question is how current communication protocols can be adapted to incorporate the various aspects of semantic communication. To provide an example, consider a scenario similar to the one depicted in Fig. 1a), where a sender conveys information to the destination about a physical process, e.g. by sending status updates. Of interest is the availability of a certain service, defined as the fraction of time over which the AoI is below a predefined threshold, as illustrated in Fig. 4.[2] The objective is to jointly design the sampling of the physical process (data generation) with the communication over the wireless channel (including channel access in the multi-user setting) such as to increase the service availability.

The joint optimization, however, would require abandoning the conventional assumption of *exogenous data arrivals* on which legacy systems are built. Since the traffic generation process is taken care of by the application layer, a form of *"semantic interface"* needs to be established between the application layer and the lower layers of the protocol stack. While this can be perceived as an unnecessary complication of protocol design, the approach can be justified if significant savings are demonstrated. Another aspect that should be considered is that today most traffic on the internet is end-to-end encrypted through (D)TLS, meaning that application data can not be accessed by on-path elements such as routers or switches. While it is feasible that transport layer encryption can be forgone in private (campus) networks, some form of encryption, e.g., on the physical layer, may be desirable (as discussed earlier).

Toward A Semantic Cross-Layer Networking

Going beyond the joint optimization of data generation and communication, various aspects of semantic communication can have an impact on the design of the different layers of the protocol stack. In this perspective, integrating aspects of semantic communication in the protocol stack would effectively require a form of cross-layer optimization, which would break the classical assumption of the separation between the layers.

One example of such semantic architecture is provided by the introduction of a *semantic-effectiveness plane*, as envisioned in [14]. The challenges involved in the implementation of this new "semantic plane" pass through some of the mechanisms already addressed by the QoS community, with the addition of a new perspective where the figure of merit is data significance (e.g., freshness, relevance, and value), rather than accurate data reconstruction at the destination.

When it comes to the actual semantic plane implementation, one has to consider where the semantic information is contained. As every layer of the protocol stack is supposed to be able to access and modify the semantic information-related parameters, these can not be stored in the header of any individual layer as that layer will strip the header along with the information (meaning higher layers will no longer have access

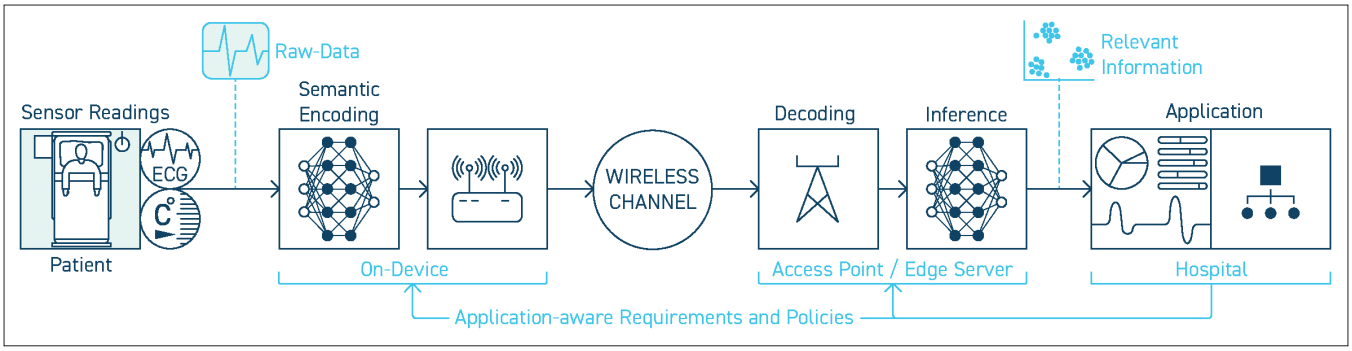

FIGURE 3. An example of a semantic-aware architecture for device-edge co-inference in a health monitoring application.

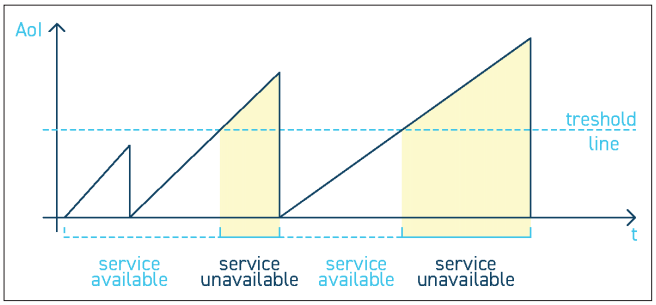

FIGURE 4. AoI and service availability.

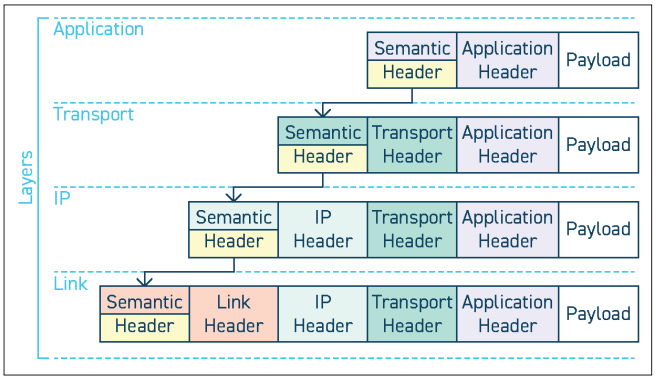

FIGURE 5. Depiction of an additional semantic header that floats through the layers to allow all of them to access it. Each layer removes the semantic header, adds its own header, and then reattaches it. The application itself is the origin of this header.

to it). In addition, encapsulation is also not a viable option as it would require knowledge of which network components do not support the semantic plane. This knowledge is needed to ensure that packets get decapsulated in advance in order to keep backward compatibility intact. While it seems challenging to integrate aspects of semantic communication into the existing Internet ecosystem, the situation might be different in private/campus networks (e.g., for industrial automation) where more flexibility and application awareness is desirable. In such smaller, single-domain networks, it can be assured that every network component supports the semantic plane, and a "floating" header can be used. This header would be read, potentially changed, and then moved to the next layer. Depending on the relevance of information in the semantic header, e.g., freshness, the packet could be dropped, as illustrated in Fig. 5.

Examples and Implications

We will now provide two further examples of how the semantic header can be used to allow information to flow between layers. This can happen in a *bottom-up* approach, where the lower layers (PHY and MAC) provide information about their current state that is then considered by the application layer. Alternatively, the application layer might provide certain parameters or requirements it has for any data it sends that can then be acted upon by the lower layers (*op-down*).

Bottom-Up: As an example, consider a robot swarm application where a group of robots is coordinated in a centralized fashion via a wireless infrastructure. Here, each robot can either pursue an individual task (e.g., logistics within a warehouse) or a common task (e.g., surveillance of an industrial plant). To keep formation and/or cooperation between individual agents, the centralized controller requires accurate and fresh updates on the positions and status of the individual robots' trajectories. Continuous transmission of updates from all robots via the shared wireless channel is challenging, especially under stringent latency requirements, as the wireless channel can be degraded or congested (due to the channel access mechanism). A goal-oriented approach jointly considers traffic generation and channel access and communication according to the value of the information encapsulated in a data packet for the specific goal. Freshness and/or *novelty* can be used as a measure for prioritizing data packets within the communication network, where novelty can be, e.g., derived from apriori knowledge of the physical process behind the sensor readings. Transmission of relevant information (e.g., changes in the direction of movement) can relieve the radio communication while still providing sufficient information for the control task.

Top-Down: In a top-down approach, the application layer adds additional information about preferences or requirements it poses to the connection. For example, an application might indicate that it has a preference for lower latency, even if it comes at the cost of reliability. For example, in a multi-connectivity scenario, a base station that could transmit to a user either directly or indirectly (e.g., via an intelligent reflective surface) might make the decision to take the direct link (even though it might be less reliable), as it induces lower latency. The advantage of a top-down approach is that there is no need for verification on the side of the network that the supplied information is accurate (a correctly functioning application would never provide parameters that go against its own interests). Hence, there is only one set of parameters that needs to be considered and what matters is that the demands and requests of the application will remain stable during the lifetime of the connection.

Concluding Remarks

This paper revisited the theoretical foundations of semantic communication and discussed some implications on system design. Two perspectives on semantic communication were addressed:

1. A goal-oriented perspective, with emphasis on the need for joint optimization of the processes of data acquisition, communication, and control
2. A semantic operability perspective, with emphasis on the information-theoretic foundations and the relevance to edge intelligence applications.

In the context of (1), a major and novel research challenge is the integration of networked control aspects. In this respect,

a more fundamental understanding of the relationship between control and communication problems needs to be developed, where concepts such as the anytime capacity and the VoI can serve as starting points. We postulate that resolving these challenges would eventually lead to a unified framework that integrates the aspects of data acquisition, communication, and control, with relevance to application fields such as collaborative robotics and autonomous systems.

In the context of (2), we introduced the principle of semantic interoperability and addressed the related information-theoretic concepts. This perspective is of particular relevance to edge intelligence applications, in light of the migration of AI functionalities from the cloud to the wireless edge. To illustrate the potential of the considered framework, we provided an example of a generic semantic-aware architecture for edge inference. In a novel contribution, we sketched how aspects of privacy and physical layers security can be natively integrated in an end-to-end learning framework based on the information bottleneck principle. In this context, we foresee an increased relevance of neuromorphic computing, where processing is performed by spiking neural networks (SNNs). Initial experiments illustrate the potential of neuromorphic, semantic-aware architectures, to provide support for energy-efficient, event-driven processing in such scenarios.

Acknowledgement

The authors would like to thank Gianluigi Liva, Igor Bjelakovic, Eduard Jorswieck, and Federico Clazzer for the valuable feedback. The authors also thank Risto Avramovski for the help with the illustrations. This work was supported by the Federal Ministry of Education and Research of Germany in the program "Souverän. Digital. Vernetzt." Joint project 6G Research and Innovation Cluster (6G-RIC), project identification numbers: 16KISK020K, 16KISK030, 16KISK031, and 16KISK022.

References

[1] W. Weaver, "The Mathematics of Communication," *Scientific American*, vol. 181, no. 1, 1949, pp. 11–15.
[2] Y. Bar-Hillel and R. Carnap, "Semantic Information," *The British Journal for the Philosophy of Science*, vol. 4, no. 14, 1953, pp. 147–57.
[3] P. Adriaans, "A Critical Analysis of Floridi's Theory of Semantic Information," *Knowledge, Technology & Policy*, vol. 23, no. 1-2, 2010, pp. 41–56.
[4] M. Kountouris and N. Pappas, "Semantics-Empowered Communication for Networked Intelligent Systems," *IEEE Commun. Mag.*, vol. 59, no. 6, 2021, pp. 96–102.
[5] R. D. Yates *et al.*, "Age of Information: An Introduction and Survey," *IEEE JSAC*, vol. 39, no. 5, 2021, pp. 1183–1210.
[6] E. Uysal *et al.*, "Semantic Communications in Networked Systems: A Data Significance Perspective," *IEEE Network*, vol. 36, no. 4, 2022, pp. 233–40.
[7] R. Dobrushin and B. Tsybakov, "Information Transmission with additional Noise," *IRE Trans. Inf. Theory*, vol. 8, no. 5, pp. 293–304, 1962.
[8] N. Tishby, F. C. Pereira, and W. Bialek, "The Information Bottleneck Method," arXiv preprint physics/0004057, 2000.
[9] E. Beck, C. Bockelmann, and A. Dekorsy, "Semantic Information Recovery in Wireless Networks," Sensors, vol. 23, no. 14, 2023.
[10] A. Sahai and S. Mitter, "The Necessity and Sufficiency of Anytime Capacity for Stabilization of a Linear System Over a Noisy Communication Link—Part I: Scalar Systems," *IEEE Trans. Info. Theory*, vol. 52, no. 8, 2006, pp. 3369–95.
[11] T. Soleymani, J. S. Baras, and S. Hirche, "Value of Information in Feedback Control: Quantification," *IEEE Trans. Autom. Contro*l, vol. 67, no. 7, 2022, pp. 3730–37.
[12] J. Liu, W. Zhang, and H. V. Poor, "A Rate-Distortion Framework for Characterizing Semantic Information," *2021 IEEE Int'l. Symp. Info. Theory (ISIT)*, 2021, pp. 2894–99.
[13] N. Skatchkovsky, O. Simeone, and H. Jang, "Learning to Time-Decode in Spiking Neural Networks Through the Information Bottleneck," *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 17,049–59.
[14] P. Popovski *et al.*, "Semantic-Effectiveness Filtering and Control for Post-5G Wireless Connectivity," *J. Indian Institute of Science*, vol. 100, 2020, p. 435–43.

Biographies

Zoran Utkovski is a senior researcher at the Fraunhofer Heinrich Hertz Institute in Berlin, Germany, where he heads the Smart Wireless Connectivity Group. He received a Dipl.-Ing. in electrical engineering (2000) from Sts. Cyril and Methodius University, Skopje, Macedonia, an M.Sc. degree (with distinction) from Chalmers University of Technology in Gothenburg, Sweden (2004), and a Dr. Ing. degree (summa cum laude) from Ulm University, Germany (2010). His research interests are in communication theory, machine learning, communication security, and complex systems theory.

Andrea Munari [M'10, SM'19] received the M.Sc. and the Ph.D. degrees in Telecommunications Engineering from the University of Padova, Italy, in 2006 and 2010, respectively. From 2007 to 2010 he was with IBM Research in Zurich, Switzerland, and in 2011 he joined the Corp. R&D division of Qualcomm Inc. in San Diego, California. Currently, he is with the Institute of Communications and Navigation of the German Aerospace Center (DLR). His main research interests include, among others, the design and modeling of medium access techniques, with special attention to Internet of Things applications.

Giuseppe Caire received a M.Sc. in Electrical Engineering from Princeton University in 1992, and a Ph.D. from Politecnico di Torino in 1994. He has been Assistant Professor in Telecommunications at the Politecnico di Torino, Associate Professor at the University of Parma, Italy, Professor with the Department of Mobile Communications at the Eurecom Institute, Sophia-Antipolis, France, a Professor of Electrical Engineering with the Viterbi School of Engineering, University of Southern California, Los Angeles, and he is currently an Alexander von Humboldt Professor with the Faculty of Electrical Engineering and Computer Science at the Technical University of Berlin, Germany.

Johannes Dommel received the Dipl.-Ing. in electrical engineering from the Ilmenau University of Technology, Germany, in 2010 and the Dr.-Ing. in the field of network information theory from the Technical University of Berlin, Germany, in 2022. Since 2010 he works as a research associate in the Wireless Communication and Networks Department at the Fraunhofer Heinrich Hertz Institute, Berlin. His research includes the application of information theory, statistical signal processing and machine learning with a focus on wireless communications.

Pin-Hsun Lin [S'05, M'10, SM'22] received the B.S. and Ph.D. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, in 2000 and 2010, respectively. From 2011 to 2012, he was with the INTEL-NTU LAB. From 2012 to 2014, he has been a Researcher with the Smart Wireless Lab, National Institute of Information and Communications Technology, Yokosuka, Kanagawa, Japan. From 2014 to 2019, he has been a Postdoctoral research associate with the Technische Universität Dresden, Germany. Since 2019, he is a Postdoctoral research fellow with the Information Theory and Communications System Department, Technische Universität Braunschweig, Germany.

Max Franke received his Bachelor and Master degree in Computer Science from Technical University of Berlin, Germany. Currently, he is a Ph.D. student, and his research centers on internet architecture, transport and network layer protocols, multicast, and 6G technology.

André C. Drummond [M] received a B.S. degree in computer engineering from the Pontifical Catholic University of Campinas in 2002 and an M.Sc. and Ph.D. in computer science from the State University of Campinas, Brazil, in 2005 and 2011, respectively. He is an Associate Professor with the Department of Computer Science, University of Brasilia, Brazil. He is on a three-year leave working as a senior researcher at the Technische Universität Braunschweig, Germany. His research interests include computer networks, optical networks, and traffic engineering for optical and 6G networks.

Sławomir Stańczak [SM] is Professor of Network Information Theory at the Technical University of Berlin and Head of the Wireless Communications and Networks Department at the Fraunhofer Heinrich Hertz Institute. He is the co-author of two books and more than 200 peer-reviewed journal articles and conference papers in the field of information theory, wireless communications, signal processing, and machine learning. Since 2020, he is chairman of the 5G Berlin association and since 2021 he is coordinator of the projects 6G Research and Innovation Cluster (6G-RIC) and CampusOS.

Footnotes

[1] As pointed in [10], Shannon himself had suggested looking to control problems for more insight into reliable communication.
[2] As discussed, in spite of its simplicity, AoI can capture some key trade-offs also in terms of data relevance and is chosen here for ease of discussion.