Toward Goal-Oriented Semantic Communications: New Metrics, Framework, and Open Challenges

Aimin Li, Shaohua Wu, Siqi Meng, Rongxing Lu, Sumei Sun, and Qinyu Zhang

Abstract

Since Shannon's pioneering masterpiece which established the foundation of modern information theory, the design target of communications has long been promising bit-level message reconstruction and achieving Shannon capacity. However, this focus neglects the semantics and effectiveness aspects of information. Recently, the development of wireless technologies and the spurt of deep learning (DL) techniques allow us to reclaim the meaning/usefulness aspect in the design of 6G communication paradigms, where goal-oriented communication is becoming a trend. Age of information (AoI), a well-known metric that captures the significance of information by recording the time elapsed from the generation time slot, has been extended to various variants, such as value of information (VoI), urgency of information (UoI), age of incorrect information (AoII), and so on. While each of them proposes novel ways to measure the semantics/effectiveness aspect of information, there is not yet a unified framework encompassing all of them. To this end, we propose a novel tensor-based approach, the goal-oriented tensor (GoT), to unify them, which allows more flexible and fine-grained goal characterizations. Following the proposed GoT, we architect a holistic goal-oriented semantic communication framework, in which semantics perception, dissemination, and control-plane decision making collaboratively work toward realizing specific goals. Finally, we outline several open challenges to fulfill the vision of the GoT framework.

Introduction

Beyond Bit-Oriented Transmission

The modern prosperity of communication technologies is primarily standing on the cornerstone built by Shannon's Information Theory in 1948 [1]. Since its inception, a primary goal in the field of communication has been to approach Shannon's theoretical capacity for transmission. Today, the advanced 5G technology has nearly achieved this goal. Technically, advances like massive multiple-input multiple-output (MIMO), and millimeter wave communications have enabled a plethora of services with ultra reliability, low latency, and high data rate. At the physical layer, the combination of Low-Density Parity-Check (LDPC) codes and Polar codes also represents a breakthrough in realizing low-complexity near-Shannon-limit coding-decoding pairs. Such technologies aim at building perfect bit-by-bit information pipelines between transceivers, which follow Shannon's legacy precisely.

Looking forward, we could anticipate a massive data boom in the forthcoming data-driven world. On the one hand, non-terrestrial nodes are being deployed throughout the space to assure seamless coverage and ubiquitous connectivity. This expansion amplifies bit generation sources. Meanwhile, data types are becoming diversified and versatile, encompassing Virtual Reality (VR), high-resolution video streams, and multi-modal sensing data. As the raw data demands massive amounts of bits to characterize, the limited communication resources are further burdened. Inevitably, the traditional paradigm will suffer from a bottleneck due to limited communication resources that are not aligned with the requirements for bit-oriented ubiquitous data reconstruction.

To overcome the above challenges, Shannon's classical theory is in need of a paradigm shift, which can be traced back to Shannon's seminal work [2], where Shannon and Weaver categorized communications into three basic levels: *technical* level, *semantic* level, and *effectiveness* level. *Technical* level aims to accurately transmit data bits. *Semantic* level aims to precisely convey the desired meaning. *Effectiveness* level aims to transmit *relevant* symbols such that the receiver makes effective decisions to affect conducts in the desired way. Both the *semantic* level and the *effectiveness* level move forward to go beyond bit-oriented transmission, which have attracted extensive research interests recently.

Goal-oriented semantic communication represents an orchestration of the *semantic* level and the *effectiveness* level, and its key idea is prioritizing the sharing of goal-relevant semantic fragments, such that the decision making is accurate and the ultimate goal of the system is effectively achieved.

Aimin Li, Shaohua Wu (corresponding author), Siqi Meng, and Qinyu Zhang are with Harbin Institute of Technology, China; Rongxing Lu is with University of New Brunswick, Canada; Sumei Sun is with Agency for Science, Technology, and Research, Singapore, and also with National University of Singapore, Singapore.

Digital Object Identifier: 10.1109/MWC.014.2300492

1536-1284/24/$25.00 © 2024 IEEE

Overview of Semantic Communications

There are two primary avenues that lead the development of semantic communications. The first avenue applies deep-learning (DL) techniques to extract semantic features and filter semantics-irrelevant redundancy within a specific packet, thus enhancing communication efficiency. The second avenue interprets semantics as its etymological meaning, the *significance* of information (also known as *priority* of information or *relevance* of information).

Avenue 1: Semantics as Meaning: The first avenue focuses on the engineering fulfillment of semantic compression and extraction, in which DL techniques play indispensable roles to accurately extract semantic information within a specific packet. Such a promising avenue arises from the interest in applying DL at the physical layer [3], which facilitates a shift from bit reconstruction-oriented to semantic similarity-oriented transmission. An earlier End-to-End (E2E) prototype for DL-based semantic communications is the Deep Joint Channel Source Coding (Deep JSCC) [4], which is used to extract (at the transmitter) and reconstruct (at the receiver) the semantics of an image. Today, such a paradigm underpins a wide range of applications, such as text, video, and speech-based semantic communication. A more comprehensive view of such a track could be found in [5, Sec. IV; 6, Sec. III].

Avenue 2: Semantics as Significance: The second avenue interprets semantics as "*significance*," "*priority*," or "*relevance*" [7], which is also our focus in this article. From this perspective, intelligent networked systems can identify high-*priority* messages and adaptively allocate more resources. Notably, the determination of *priority* deviates from traditional, probability-centric entropy measures. Instead, it leans on the "*significance*" attributed to the goal-oriented usefulness of message. As noted in [8], *"Imagine two equally rare events, occurring with very low probability, one of which carries a major safety risk while the other is just a peculiarity. Although they provide the same high amount of information, the information conveyed by the first event is evidently of higher significance,"* it holds great potential to design a semantics/goal-aware *relevant filter* to slim down the information pipeline, where only packets that carry *relevant* meaning for accomplishing the goal are prioritized for transmission. Since the *significance* relates to the goal, the *significance* interpretation also bridges this avenue to the effectiveness level.

In the second avenue, the core challenge is to design metrics that could characterize the *significance* property. Age of Information (AoI), defined as the time elapsed since the generation of the most recent information update at the destination, is one typical metric that captures the *significance* property. In the AoI-oriented system, fresher messages are deemed more *relevant* to the end-user, and thus prioritized for transmission. However, AoI does not perceive the content of the packet, nor the task at the receiver. To address this issue, improved non-linear variants of AoI, such as VoI [9], AoII [10], UoI [11], Cost of Actuation Error [12], and so on, have been developed. However, there is not yet a unified framework that encompass all these metrics. Moreover, the existing metrics do not directly characterize the desired goal. These issues motivate us to propose a unified, flexible, and directly goal-oriented metric that could assist in accurately capturing the underlying goal, facilitating decision-making, and ultimately improving effectiveness.

There are two primary avenues that lead the development of semantic communications. The first avenue applies DL-learning techniques to extract semantic features and filter semantics-irrelevant redundancy within a specific packet, thus enhancing communication efficiency. The second avenue interprets semantics as its etymological meaning, the *significance* of information (also known as *priority* of information or *relevance* of information)

Toward Goal-Oriented Semantic Communications

This article addresses the limitations in the second avenue of semantic communication, focusing on the challenge of designing metrics for directly measuring information significance at the effectiveness/goal-oriented level. First, this article comprehensively reviews the existing metrics that capture the *significance* property in the following section. In this way, we reveal how these metrics manifest the effectiveness level. Then, upon examining the inherent relationships among these metrics, we are inspired to unify them in a more cohesive framework. This leads to the proposed Goal-oriented Tensor (GoT). We visualize examples to demonstrate how the GoT metric reduces to existing metrics that capture *significance* of information and how it enables more fine-grained and flexible goal characterizations. Consequently, abstract goals could be defined and quantified via GoT.

Furthermore, this article illustrates that information traverses a perception-actuation *life cycle* to achieve a certain goal. Within this *life cycle*, information initiates with its generation, goes through semantics extraction (the meaning-aware approach will enhance the effectiveness), coding and modulation, data dissemination, and ends with its transformations into effective decisions (or control) toward goals. In this way, the *significance* of information is not solely tied to the packet's inherent meaning, but the ultimate usage of the meaning for achieving particular goals.

Incorporating the *life cycle*, this article then envisions a holistic framework of goal-oriented semantic communications, whereby semantics perception, dissemination, and control-plane decisions are orchestrated in harmony with a shared goal characterized by GoT. We consider an easy-to-follow case study, a real-time wireless fire monitoring and rescue system, to illustrate the universality of our proposed framework. We illustrate how the GoT describes the goal in real scenarios. The idea of goal and semantics-aware filter (an open challenge proposed in [7] and [8]) is also explored and verified in this article.

Capturing the Significance of Information: Critical Metrics

Recent studies have explored various metrics to assess the *significance* of information. This section reviews these metrics, with an emphasis on how these metrics interconnect with the effectiveness of communications.

Age of Information (AoI)

Age of Information (AoI), proposed in 2012 [13], is the first concrete and quantitative metric to characterize the *significance* of information. Intuitively known as *freshness* of information, AoI was designed to fulfill the *freshness* requirements for the proliferated machine-type communications, such as intelligent vehicle networks, industrial internet of things, and many other ultra reliable-supported applications. Central to optimizations on AoI is a subtle yet vital consensus:

While previous metrics focus solely on the *significance* of the source, practical scenarios demonstrate that the surrounding environment also influences information *significance*.

fresher messages contain more valuable information. This consensus is rooted in the understanding that *fresh* information ensures more informed and goal-oriented decision-making.

Value of Information (VoI)

VoI introduced a non-linear penalty of AoI to capture the degree of "*dissatisfaction*" resulted by *staleness* of information [9]. A preliminary approach to characterize this non-linear penalty is to classify the applications into basic levels in terms of their sensitivity to *freshness* [9]. *Freshness*-critical applications are penalized exponentially (see Table 1 for an example), *freshness*-insensitive ones are penalized logarithmically, and those neutral ones fall into linear penalties. VoI bridges AoI with effectiveness by applying a non-linear utility function.

Mean Square Error (MSE)

MSE is a well-known metric that characterizes the accuracy aspect of information. MSE is defined as the long term average squared error (SE) between transceivers' statuses, which is used for timely reconstruction-oriented communications. In the MSE-oriented design, a status at the source is deemed significant if it considerably deviates from the estimated status at the monitor. When the sample policy is content-agnostic, the MSE can be expressed as a nonlinear function of AoI [14], making it a specialized form of VoI. MSE emphasizes the impact of the severity of E2E status mismatch on the effectiveness.

Age of Synchronization (AoS)

As its name suggests, AoS measures how long has passed since the last time the transceivers synchronized their statuses [15], addressing AoI's content-agnostic nature. Consider scenarios where the source undergoes frequent changes: even a "*fresh*" packet with a low AoI might become obsolete if the true status of the source has significantly evolved. Conversely, a "*stale*" packet can still offer accurate estimations if the source changes are minimal. As shown in Table 1, AoS posits that the duration of E2E mismatch adversely impacts effectiveness. Thus, minimizing AoS improves the effectiveness.

Age of Incorrect Information (AoII)

AoII is a new metric that orchestrates the strengths of MSE and AoS to measure *significance* property [10]. As shown in Table 1, AoII introduces three key innovations: First, AoII integrates AoS with MSE by multiplying the variants of them together, which incorporates both the duration and severity of E2E mismatch. Second, AoII incorporates a nonlinear penalty function predicated on AoS, drawing parallels to the penalty function in the Value of Information (VoI) relative to AoI. Third, AoII introduces a generalized error gap function, broadening the traditional Euclidean-distance-based MSE depictions. In this regard, AoII addresses that both severity and duration of E2E statues mismatch affect the effectiveness.

Cost of Actuation Error

Previous metrics typically measure the system effectiveness indirectly. To address this issue, *Cost of Actuation Error* emerges as the first metric to directly quantify the effectiveness at the point of actuation [12]. This metric highlights the role of the actuator to affect the goal-achieving effectiveness: an E2E status mismatch will trigger an actuation error, hence inducing associated cost. As Table 1 shows, this metric highlights the direct impacts of actuation error on effectiveness.

Urgency of Information (UoI)

While previous metrics focus solely on the *significance* of the source, practical scenarios demonstrate that the surrounding environment also influences information *significance*. For example, a self-driving car in complex situations like traffic jams would require more frequent status updates for safety, unlike when it's on a clear highway. Addressing this, UoI is the first metric to link environment factors with information *significance* [11]. Particularly, UoI introduces an environment-aware weighted coefficient based on the environment urgency level. The final UoI formula is a product of this coefficient and the error gap function, as detailed in Table 1.

One More Step Forward: The Goal-Oriented Tensor

The overview and previous discussions lead to two key insights: the existing metrics' assessment of effectiveness is indirect and they lack a unified approach. In this section, we introduce a unified metric to capture the significance of information in a goal-oriented manner and seek to bring coherence to these various metrics.

Existing Metrics: Inherent Connection

From the previous section and Table 1, the existing metrics share two fundamental components:

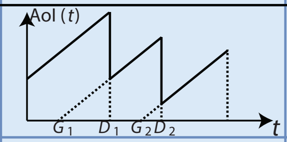
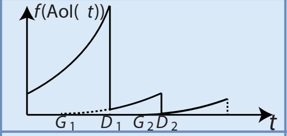
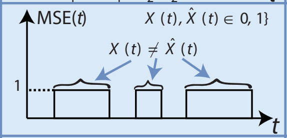
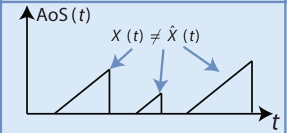
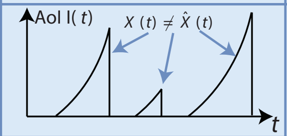
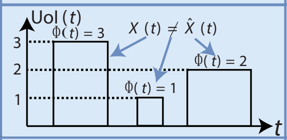
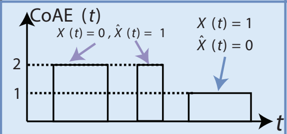
- A *content-aware* error cost function $g(X(t), \hat{X}(t))$
- A *content-independent* weighted coefficient $\Phi(t)$.

The former one describes the real-time E2E distance (like MSE or other error gap function) or mismatch cost (like *Cost of Actuation Error*). The latter one exerts a multiplicative effect on the error cost function. Thus, the inherent connection among the existing metrics is that they all involve a triple-tuple consisting of $X(t)$, $\hat{X}(t)$, and $\Phi(t)$. This connection motivates us to propose a tensor-based approach to unify the existing metrics.

Goal-Oriented Tensor: A Unified Metric

Given that all existing metrics are based on a triple-tuple, a unified approach is natural: *Unify the existing metrics as a 3-dimensional tensor, with each dimension corresponding to a tuple element*. In this subsection, We first show that a 3-dimension tensor could degenerate existing metrics, and then illustrate how a generalized GoT characterizes a specific goal.

Degeneration to Existing Metrics: Figure 1a visualizes an example of the tensor to characterize AoI. Since AoI is *content-independent*, the values in the tensor only depend on the *content-independent* weighted coefficient $\Phi(t)$, with $\Phi(t)$ equals to AoI. Figure 1b visualizes a tensor to characterize MSE where $\mathcal{S} = \{0, 1, 2, 3, 4\}$. The MSE focuses on the squared error between $X(t)$ and $\hat{X}(t)$, while neglecting the content-independent weighted coefficient $\Phi(t)$. In this case, different $\Phi(t)$ produce the same tensor

Metric	Definition	Evolution Sketch	Evolution Description	Illustrations	References
AoI	Time elapsed since the generation of the latest received packet.		The staleness of information is linearly penalized.	Fresher message provides higher value information.	[12]
VoI	Non-linear penalty of AoI.		The staleness of information is non-linearly penalized.	Information value and freshness share a non-linear relationship.	[8]
MSE	Average squared error between transcievers' statuses.		The E2E mismatch is penalized by the mismatch distance.	Timely reconstruction enhances the system performance.	[14]
AoS	Time elapsed since the last time the transcievers are syronized.		The E2E mismatch is penalized by the mismatch duration.	Reducing E2E mismatch duration enhances system performance.	[15]
AoII	Mulplication of MSE and the non-linear penalty of AoS.		The E2E mismatch is penalized by both mismatch duration and distance.	Minimized duration and distance in E2E mismatch enhances system performance.	[9]
UoI	Mulplication of the enviroment-aware coefficient $\Phi(t)$ and MSE.		The E2E mismatch is penalized by both environment urgency and mismatch distance.	Environment urgency influences the severity of E2E mismatch.	[10]
Cost of Actuation Error	The actuation error cost under a E2E mismatch $X(t) \neq \hat{X}(t)$.		The E2E mismatch is penalized by the mismatch category.	Different types of mismatch lead to diverse costs from imperfect decisions.	[11]

Notations

AoI: Age of Information
VoI: Value of Information
MSE: Mean Square Error
AoS: Age of Synchronization
AoII: Age of Incorrect Information

UoI: Urgency of Information
G_i : The time stamp of the i^{th} generated status update
D_i : The delivery time slot of the i^{th} status update

$X(t)$: The source status at time slot t
$\hat{X}(t)$: The estimated status at time slot t
$\Phi(t)$: The level of environmet emergency at time slot t

TABLE 1. Metrics for capturing the importance of information: a review.

slice characterized by the squared error. Figure 1c visualizes an example for AoII. Since AoII is characterized in a multiplicative manner, a base slice exists in the AoII-based tensor. The base slice is obtained by the error gap function $g(X(t), \hat{X}(t))$, which is the first slice in front of us in Fig. 1c. A linear multiplication of the base slice could represent the other slices. The examples are not exhaustive. A similar approach could also be implemented to obtain a tensor to characterize VoI, AoS, UoI, and *Cost of Actuation Error*.

Generalized GoT

A more generalized GoT is visualized in Fig. 1d. First, $X(t)$ is extended from content-agnostic raw status to semantics-aware status in the generalized GoT. This extension highlights the asymmetry of the mismatch cost between $X(t)$ and $\hat{X}(t)$, corroborating the principle of the *Cost of Actuation Error*. The rationale of this extension could be illustrated through an autonomous driving situation: a false positive detection — perceiving a person where there is no one — may cause unnecessary braking but remains safe. In contrast, a false negative detection — the failure to detect an actual person — may lead to severe consequences. This highlights the need for an asymmetrical error function as an extension of the symmetrical one used in MSE and UoI. Second, the *content-independent* weighted coefficient $\Phi(t)$ is extended from multiplicative coefficient to context-aware representation. The role it plays in determining the tensor value is versatile, moving beyond a mere multiplicative method. Depending on the context, different end-to-end semantic discrepancies will result in varying costs, as depicted in Fig. 1d.

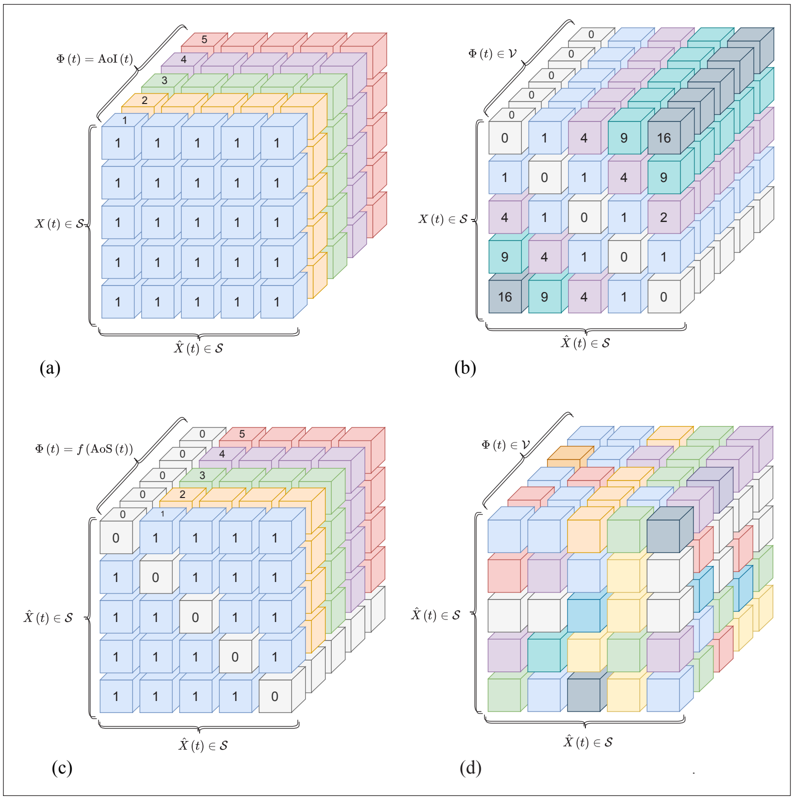


FIGURE 1. Examples of tensor-based metric visualizations: a) GoT characterizes AoI; b) GoT characterizes MSE; c) GoT characterizes AoII; d) The generalized GoT.

Constructing Generalized Goal-Oriented Tensor

The values in the generalized GoT manifest the cost of E2E statuses mismatch under a specific context status. Here a detailed method to tailor a GoT based on the goal is given.

Step 1: Clarifying the scenario and the goal. In wireless accident monitoring and rescue systems, for example, the goal is to achieve effective monitoring and rescue such that the long-term average cost resulted by accident damages and rescue resources is minimized.

Step 2: Defining the semantic status set $\mathcal{S}$ and the context status set $\mathcal{V}$. Both sets are constructed as finite collections of discrete semantic segments, with the semantic status $X(t) \in \mathcal{S}$ and $\Phi(t) \in \mathcal{V}$ determined through semantic quantization (refer to next section). As a specific example, in a fire monitoring and rescue system, the set $\mathcal{S}$ represents the fire intensity levels, defined as $\mathcal{S}$ = {No Fire, Fire}. Concurrently, the set $\mathcal{V}$ can encompass various weather conditions in the vicinity of the fire, specified as $\mathcal{V}$ = {Sunny, Rainy}. Obviously, the semantics hold varying significance across different contexts.

Step 3: Determining the decision strategy in terms of the observation at the receiver $\hat{X}(t)$, denoted by $\pi(\hat{X}(t)) : \mathcal{S} \rightarrow \mathcal{A}$, where $\mathcal{A}$ is the action space. This action will turn back to affect the goal-oriented effectiveness of the system. For instance, $\mathcal{A}$ can be {Rescue, Idle}, and the rescue action response to the received fire intensity level, given as π(No Fire) = Idle, π(Fire) = Rescue can be a specific decision strategy.

Step 4: Evaluating the specific cost. There are three types of costs:

- The status inherent cost $\mathcal{C}_1(X(t), \Phi(t))$, which represents the cost per time slot under $X(t)$ and $\Phi(t)$, for instance, $\mathcal{C}_1$(No Fire, Sunny) = 0 represents zero cost per time slot when there is no fire in a sunny day
- The decision gain $\mathcal{C}_2(\pi(\hat{X}(t)))$ is the cost reduction from an effective action $\pi(\hat{X}(t))$, for instance, $\mathcal{C}_2$(Rescue) = 5 represents that the rescue action can reduces 5 unit cost per time slot
- The action cost $\mathcal{C}_3(\pi(\hat{X}(t)))$ corresponds to the resource overhead due to the action $\pi(\hat{X}(t))$, for instance, $\mathcal{C}_3$(Rescue) = 3 represents that the rescue action incur 3 unit cost per time slot.

Step 5: Calculating the GoT. The tensor value, given a specific triple-tuple $\langle X(t), \hat{X}(t), \Phi(t)\rangle$ and a determined decision mapping approach $\pi(\cdot)$, is calculated by

$$[\mathcal{C}_1(X(t), \Phi(t)) - \mathcal{C}_2(\pi(\hat{X}(t)))]^+ + \mathcal{C}_3(\pi(\hat{X}(t))), \qquad (1)$$

where the ramp function $[\cdot]^+$ ensures that any decision gain $\mathcal{C}_3$ will not reduce the cost below 0. For example, under the ⟨No Fire, Fire, Sunny⟩ tuple, the cost per time slot is $[0 - 5]^+ + 3 = 3$, representing that if Fire recognized as No Fire will incur 3 unit cost per time slot. Through an exhaustive traversal of the tuple space $\mathcal{S} \times \mathcal{S} \times \mathcal{V}$, the tensor is formulated.

To conclude, the tensor value within GoT is determined by both the E2E semantic mismatch and the decision strategy,[1] representing the real-world cost per time slot due to these mismatches. Through GoT optimization, we enhance decision-making and mitigate significant E2E semantic mismatch impact. Therefore, the long-term real-world cost is effectively reduced, highlighting GoT's direct focus on effectiveness.

A Holistic Framework for Goal-Oriented Semantic Communication

This section introduces a comprehensive goal-oriented semantic framework, depicted in Fig. 2, with the GoT acting as the central metric. The modules are elaborated below.

Semantic Quantization

Status update packets are generated when the sensors collect continuous raw data. The types of metadata are various, for example, a piece of video/speech, a process of temperature variations, a fragment of moving track, or even their multi-modal combinations. Raw data are fed into a semantic quantizer for discrete semantic segments, producing semantic representation $X(t)$ and context representation $\Phi(t)$. The focus of the quantizer is to convert continuous, non-meaningful metadata into discrete semantic pieces, and thus we call this process as semantic quantization. This method of discretization offers dual benefits: it reduces the data volume per packet, thereby speeding up further processing and saving communication resources; and it enables the identification of semantics priorities in terms of goals, facilitating the sparse semantics-aware sampling design.

Sparse Semantics-Aware Sampling

Context status $\Phi(t)$, semantic status $X(t)$, and the feedback signal, that is, ACK or NACK are fed into the sampler to decide whether the current status should be sampled and transmitted. In our framework, $\Phi(t)$ can illustrate the context surrounding

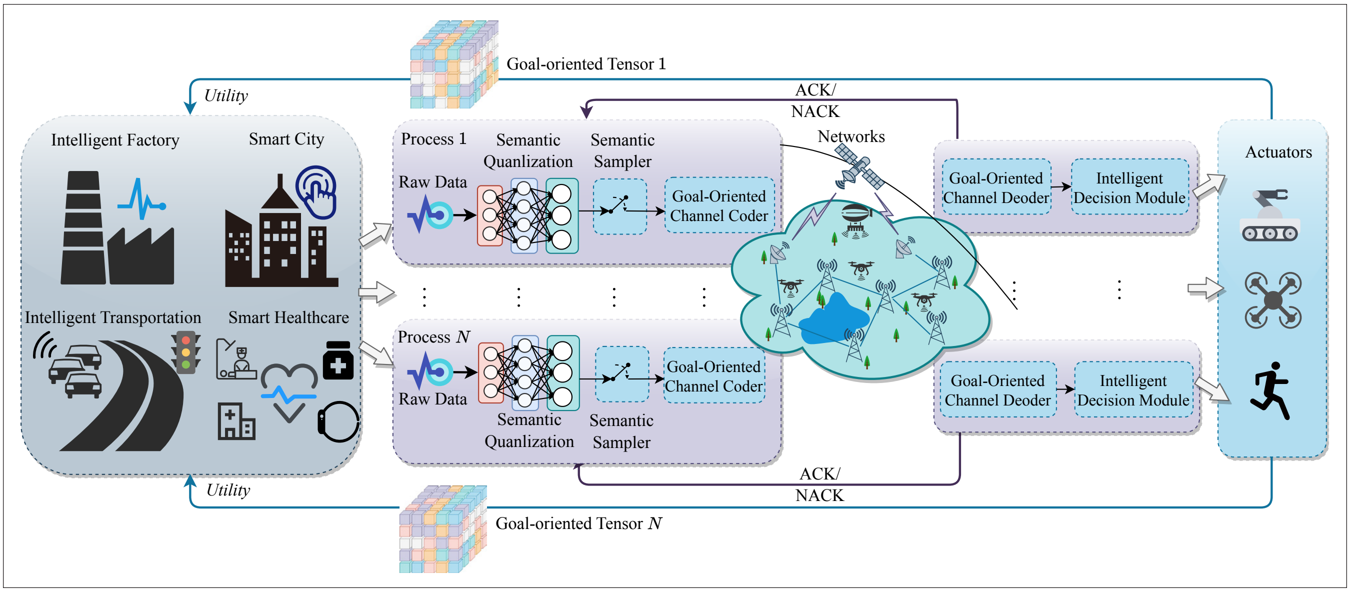


FIGURE 2. A sketch for the semantics-empowered goal-oriented networks.

the perceived source and is a sensor-derived meaningful perception. Only updates that facilitate goal achieving are transmitted, while others are filtered. In an ideal design, sparse semantics-aware sampling that fulfills the goal can be achieved.

Goal-Oriented Channel Coding

Traditionally, channel coding strategies aim for perfect reconstructions and lower average decoding error rate. However, this conventional approach treats all end-to-end (E2E) errors equivalently, disregarding the varying consequences associated with different types of errors. To address this issue, we can introduce goal-oriented channel coding to minimize the average cost associated with errors. In this context, E2E errors with potentially severe impacts are allocated larger code distances, thereby diminishing the likelihood of such errors, vice versa. This method ensures that more critical errors are less probable, thus enhancing system goal-achieving utility.

Intelligent Decision Making

This process indicates the end of the *life cycle* of information, which intelligently transforms an estimated status update into a particular utilizable decision update. The decision could be executed by actuators, such as an automobile, a remote-controlled UAV, a robotic arm, or well-trained people. The actuators transform the decisions into the ultimate *effectiveness*, which generally turns back to exert positive influences on the evolutionary process of the source.

Case Study: Goal-Oriented Sparse Semantic Sampling

This section considers the real-time wireless fire monitor and rescue system in Fig. 3 as a case study. The goal in Fig. 3 is to minimize the long-term average economic losses and resource consumption associated with fire and rescue induced by fire and rescue. As an initial exploration, our focus in this article is on the design of sparse semantics-aware sampler. The sampler should decide which semantics to sample and transmit, such that the rescue is timely and effective to minimize the long-term real-world costs induced by fire and rescue. This E2E status mismatch cost could be recorded by GoT, and thus to achieve the goal is equivalent to minimizing the long-term average E2E status mismatch costs.

As a case study, we evaluate a basic system where the context status $\Phi(t)$ remains unchanged. The semantic status captures three fire severity levels, with $\mathcal{S}$ = {No Fire, Small Fire, Big Fire}. We consider the 3 × 3 GoT visualized at the right-hand side of Fig. 3, signifying that the "No Fire" estimated as "No Fire" will consume zero real-world cost per time slot, while "Small Fire" estimated as "No Fire" will incur 20 unit real-world cost per time slot, and so on. We assume perfect semantic quantization. The semantic sampler decides when and what to sample and transmit a status through an e-erasure channel to the fire department. Firefighters (actuators) are dispatched based on the latest fire levels. The time is slotted and the source is modeled as a controlled Markov Source. The transition probabilities of the fire levels depend on the firefighters assigned. In such a system, we compare the following six sampling policies in terms of goal-oriented effectiveness and sampling rate:

- *Uniform:* Sampling is triggered periodically, which is independent of the packet's content.
- *Age-aware:* Sampling is conducted once the AoI reaches a pre-defined threshold, which is content-agnostic.
- *Change-aware:* Sampling is triggered when the source changes, which is content-aware.
- *Optimal MMSE:* Sampling is designed to minimize long-term average MSE, which is content-aware.
- *Optimal AoII (E2E-Semantics):* Sampling is designed to minimize long-term average AoII, which holds consistent with the E2E-semantics policy proposed in [8]. Specifically, sampling is triggered once an E2E status mismatch arises.
- *Optimal GoT:* Sampling is designed to minimize long-term average E2E mismatch costs record-

[1] Decision strategy can serve as an extended dimension of the GoT. It will change the value of the 3-dimensional GoT.

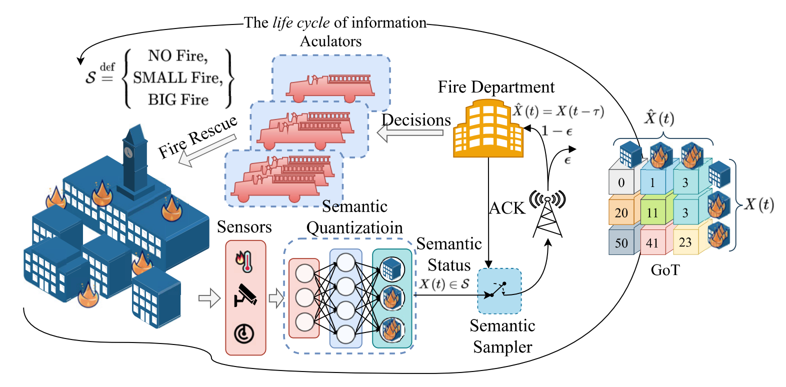


FIGURE 3. Case study: A real-time wireless fire monitor and rescue system.

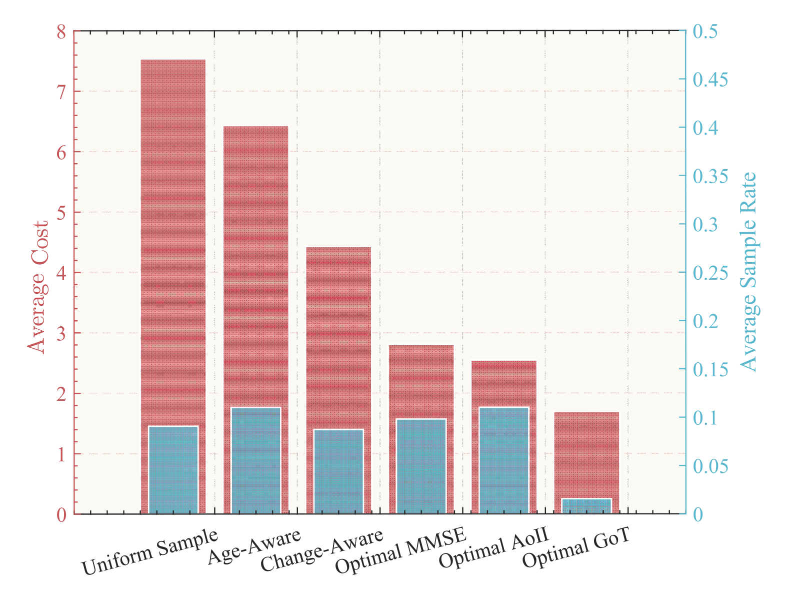


FIGURE 4. The performance comparisons among different sampling policies. Here the goal is to reduce the long-term average cost caused by fire and rescue.

ed by GoT, which is resulted by a combination of personnel casualties, equipment and building damages, as well as manpower or resources consumption in the real world. The sampling policy is solved resorting to the Markov Decision Process (MDP).

Figure 4 showcases the simulated results that compare the performance of the discussed sampling policies. The performance is evaluated by two indicators: long-term average cost, reflecting the cumulative real-world implications of fire and rescue, and average sample rate, indicating communication overhead. The primary goal is to reduce the long-term average cost, and as Fig. 4 reveals, the GoT-optimal policy excels in this, marking its inherent alignment with our goal. This policy also demonstrates an efficient balance between reduced communication overhead (sample rate) and effectiveness, suggesting a semantics-aware, goal-driven sampling method. *For example, the GoT-optimal policy can adaptively identify the semantics "Big Fire" as highly significant, thereby prioritizing its transmission to the receiver*. In summary, the GoT-optimal policy excels by focusing on the transmission of data that is most significant to achieve the goal. This method enhances the recognition of semantic significance.

Conclusions and Future Challenges

The *significance* of information plays a critical role in interpreting *"semantics."* Following this avenue, this work moves further by proposing a new goal-oriented metric and envisioning a holistic goal-oriented architecture. The proposed GoT provides a unified and extensible solution to characterize the goal of communications, which provides a solution to the effectiveness problem. The proposed goal-oriented network architecture demonstrates its great superiority in substantially alleviating the communication burden of the next-generation Internet of Everything (IoE) networks. A preliminary instantiated case study is demonstrated to address the challenge of sparse semantic sampling presented in [7, 8].

Toward the promising avenue of research, some interesting open challenges have been discussed in [7, 8], which also align with our proposed framework. Here we complement some open challenges that have not yet been presented.

Communication Networks with Heterogeneous Goals

The intelligent factory, smart cities, intelligent transportation, and smart healthcare are typical real-time applications for the future IoE networks. Sensors in this network collect terabytes of metadata every second, placing a great deal of pressure on the constrained communication resources. As such, it is imperative to slim down the data from its generation. An initial architecture to address this issue is shown in Fig. 2, where the E2E heterogeneous goals are characterized by different GoTs with diversified sizes and values, and the semantic sampling and coding are optimized from a system perspective.

Goal-Oriented Physical (PHY) Layer Techniques

To achieve a specific goal under constrained resources, the bit-by-bit reconstruction is so energy-intensive and low-efficiency that it could not perceive the *priority* of information in terms of the ultimate effectiveness to adaptively allocate the limited resources. In this regard, new goal-oriented paradigms at the PHY could be explored to further improve communication efficiency. In particular, the goal-oriented channel coding and decoding algorithms, the goal-oriented retransmission and feedback mechanism, the multi-user power allocation mechanism, the goal-oriented modulation and signal shaping are some promising candidates. By this means, the grand vision beyond the traditional paradigm is that the future communication is not designed to only reduce the error probability in an average manner but to circumvent the severe errors, where the severity and the cost of the errors are recorded by the GoT.

Goal-Oriented Perception, Communication, Computation, and Control Co-Design

The *life cycle* of information is closely affiliated with the processes of perception, communication, computation, caching, and control. Therefore, a true leap forward can be achieved by merging these modules together for the co-design optimizations. Particularly, the multi-modal sensors consecutively perceive high-resolution raw data, the computation leads to precise semantic extractions and effective intelligent decisions, the communication serves as a *priority*-aware information pipeline for information disseminating, and the actuators execute the control commands precisely, so as to transform the information into practical usage. To achieve this co-design, high-order GoTs might be designed to integrate the entire perception-action closed-loop process across various function modules.

Acknowledgment

This work has been supported in part by the National Key Research and Development Program of China under Grant no. 2020YFB1806403, and in part by the Guangdong Basic and Applied Basic Research Foundation under Grant no. 2022B1515120002, and in part by the China Scholarship Council under Grant no. 202306120310, and in part by the Fundamental Research Funds for the Central Universities under Grant no. HIT.DZJJ.2023116, and in part by the Shenzhen Science and Technology Program under Grant no. ZDSYS20210623091808025, and in part by the Major Key Project of PCL under Grant no. PCL2024A01.

References

[1] C. E. Shannon, "A Mathematical Theory of Communication," *Bell Syst. Tech. J*, vol. 27, no. 3, 1948, pp. 379–423.

[2] W. Weaver, "Recent Contributions to the Mathematical Theory of Communication," *ETC: A Review of General Semantics*, 1953, pp. 261–81.

[3] Z. Qin *et al*., "Deep Learning in Physical Layer Communications," *IEEE Wirel. Commun*., vol. 26, no. 2, 2019, pp. 93–99.

[4] E. Bourtsoulatze, D. Burth Kurka, and D. Gündüz, "Deep Joint Source-Channel Coding for Wireless Image Transmission," *IEEE Trans. Cogn. Commun. Netw*., vol. 5, no. 3, 2019, pp. 567–79.

[5] D. Gündüz *et al*., "Beyond Transmitting Bits: Context, Semantics, and Task-Oriented Communications," *IEEE JSAC*, vol. 41, no. 1, 2022, pp. 5–41.

[6] W. Yang *et al*., "Semantic Communications for Future Internet: Fundamentals, Applications, and Challenges," *IEEE Commun. Surv. Tutor*., vol. 25, no. 1, 2023, pp. 213–50.

[7] E. Uysal *et al*., "Semantic Communications in Networked Systems: A Data Significance Perspective," *IEEE Netw*., vol. 36, no. 4, 2022, pp. 233–40.

[8] M. Kountouris and N. Pappas, "Semantics-Empowered Communication for Networked Intelligent Systems," *IEEE Commun. Mag*., vol. 59, no. 6, 2021, pp. 96–102.

[9] A. Kosta et al., "Age and Value of Information: Non-Linear Age Case," *Proc. IEEE ISIT*, 2017, pp. 326–30.

[10] A. Maatouk *et al*., "The Age of Incorrect Information: A New Performance Metric for Status Updates," *IEEE/ACM Trans. Netw*., vol. 28, no. 5, 2020, pp. 2215–28.

[11] X. Zheng, S. Zhou, and Z. Niu, "Urgency of Information for Contextaware Timely Status Updates in Remote Control Systems," *IEEE Trans. Wirel. Commun*., vol. 19, no. 11, 2020, pp. 7237–50.

[12] N. Pappas and M. Kountouris, "Goal-Oriented Communication For Real-Time Tracking In Autonomous Systems," *Proc. IEEE ICAS*, 2021, pp. 1–5.

[13] Kaul *et al*., "Status Updates Through Queues," *Proc. Annual CISS*, 2012, pp. 1–6.

[14] Y. Sun, Y. Polyanskiy, and E. Uysal, "Sampling of the Wiener Process for Remote Estimation Over a Channel With Random Delay," *IEEE Trans. Inf. Theory*, vol. 66, no. 2, 2020, pp. 1118–35.

[15] J. Zhong *et al*., "Two Freshness Metrics for Local Cache Refresh," *Proc. IEEE ISIT*, 2018, pp. 1924–28.

Biographies

Aimin Li received his B.E. from Harbin Institute of Technology Shenzhen (HITSZ) in 2020, where he was awarded the highest honor of Undergraduate Thesis. He is currently a Ph.D student at HITSZ. He has served as a Reviewer for IEEE Trans. Wirel. Commun., IEEE Commun. Lett., and IEEE Trans. Neural Networks Learn. Syst.. His current research interests include age of information, channel coding, and goal-oriented semantic communications.

Shaohua Wu [M] received his Ph.D. degree in communication engineering from Harbin Institute of Technology, Harbin, China, in 2009. is a Full Professor at HITSZ and Peng Cheng Laboratory. He was also a Visiting Researcher at the University of Waterloo's BBCR in 2014-2015. His current research interests include wireless image/video transmission, space communications, advanced channel coding techniques, and B5G wireless transmission technologies. He has authored or co-authored over 50 Chinese patents and 200 articles, two of which have received the best paper awards.

Siqi Meng received his B.E. from Harbin Institute of Technology Shenzhen (HITSZ) in 2021. He is currently a Ph.D. student in HITSZ. His research interests include age of information and semantic communications.

Rongxing Lu [F] received his Ph.D. degree from the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada, in 2012. He is a Mastercard IoT Research Chair, a University Research Scholar, an Associate Professor with the Faculty of Computer Science (FCS), University of New Brunswick (UNB), Fredericton, NB, Canada. Before that, he worked as an Assistant Professor with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, from 2013 to 2016. He worked as a Postdoctoral Fellow with the University of Waterloo from 2012 to 2013. His research interests include applied cryptography, privacy enhancing technologies, and IoT-Big Data security and privacy. He won the 8th IEEE Communications Society (ComSoc) Asia–Pacific Outstanding Young Researcher Award in 2013. He currently serves as the Chair for the IEEE ComSoc Communications and Information Security Technical Committee, and the Founding Co-Chair for the IEEE TEMS Blockchain and Distributed Ledgers Technologies Technical Committee.

Sumei Sun [F] is Executive Director of the Institute for Infocomm Research (I2R), Agency for Science, Technology, and Research (A*STAR), Singapore. She also holds an adjunct appointment with the National University of Singapore, and joint appointment with the Singapore Institute of Technology, both as a full professor. Her current research interests include next-generation wireless communications, joint communication-sensing-computing-control design, industrial internet of things, applied deep learning and artificial intelligence. She is a member of the IEEE Vehicular Technology Society Board of Governors (2022-2024), Fellow of the IEEE and the Academy of Engineering Singapore.

Qinyu Zhang [SM] received his B.E. from Harbin Institute of Technology (HIT), Harbin, China, in 1994, and the Ph.D. degree in biomedical and electrical engineering from the University of Tokushima, Tokushima, Japan, in 2003. He was an Assistant Professor with the University of Tokushima from 1999 to 2003. He has been with HITSZ since 2003, where he is currently a Full Professor and serves as the Vice President. His research interests include aerospace communications and networks, and wireless communications and networks. He is the Founding Chair of the IEEE Communications Society Shenzhen Chapter. He has been awarded the National Science Fund for Distinguished Young Scholars, Young and Middle-Aged Leading Scientist of China and the Chinese New Century Excellent Talents in University, and obtained three scientific and technological awards from governments.