Ref. Ares(2025)8332347 - 02/10/2025



Secured and Intelligent Massive Machine-to-Machine Communications for 6G (SCION)

D4.1: Design of Semantic Encoder and Decoder

Version: v2.0

Deliverable type:	Report
Dissemination level:	P (Public)
Work package:	WP4 (Task4.1)
Lead editor:	UESSEX
Keywords	6G, DeepSC, Hybrid NOMA, Semantic Communcation, Semantic Similarity, Spectral Efficiency, Successive Interference Cancellation

ABSTRACT

This deliverable of WP4 (D4.1) outlines foundational concepts and advanced methodologies for semantic encoding and decoding, enabling efficient and robust Semantic Communication (SemCom) within the emerging 6G wireless systems. The report includes key principles of SemCom and presents state-of-the-art machine learning-based techniques for semantic extraction, semantic similarity approximation, and multi-modal SemCom. The key contribution of this deliverable, is a hybrid Non-Orthogonal Multiple Access (NOMA) framework for resource allocation in scenarios with both semantic and bit users based on their channel gains, proposing optimized power allocation and adaptive decoding strategies for Spectral Efficiency (SE) enhancement. In slots including heterogeneous users, the framework allows bit user to first decode its signal by treating the semantic signal as interference and subsequently semantic signal is decoded without interference. Some key challenges and potential future avenues in SemCom are also discussed.

Disclaimer: Funded by the European Union's Horizon Europe programme under the Marie Skłodowska-Curie Actions Doctoral Networks Grant Agreement No. 101072375. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them. The UK institutions were funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant numbers EP/X027201/1 and EP/X027244/1.]





UK Research and Innovation

© 2023-2027 SCION Consortium. All rights reserved.

Executive Summary

This report provides the first deliverable of Work Package 4 (WP4) "Novel Design of Semantic Encoder and Decoder" and presents a comprehensive report on the design and principles underlying semantic encoding and decoding, for efficient integration into evolving 6G wireless systems.

The document begins by establishing the foundations of SemCom, outlining its theoretical concepts such as semantic channel capacity and the nature of semantic noise and errors, which significantly differ from conventional bit-level communication paradigms. A clear classification of SemCom systems is presented, highlighting its applicability in multi-modal Task-Oriented Communication (TOC) systems and satellite networks for real-world scenarios relevant to next-generation communications. The advantages of employing TOC in SemCom are summarized. A framework of TOC multi-modal system is also designed. The reason why SemCom can help improve the performance of satellite communication is clearly illustrated. What is more important, a potentially practical satellite communication system based on SemCom technique is presented.

The state-of-the-art techniques based on machine learning for semantic features extraction, encoding, transmission, and decoding are widely discussed. A detailed discussion about Deep Learning (DL) based SemCom system, namely DeepSC is done with particular focus on performance metrics of semantic similarity. For estimating the semantic similarity between the transmitted and reconstructed messages, a generalized logistic function approximation is adopted via data regression on DeepSC outputs for practical deployment in multi-modal SemCom systems. These methodologies are shown to greatly improve Spectral Efficiency (SE) through optimal resource allocation and minimal semantic distortion. Furthermore, the superiority of conventional Bit-based Communication (BitCom) in recovering the signal with exact accuracy at higher transmit power values is discussed, calling for a mechanism supporting heterogeneous semantic and bit users for performance enhancement in next-generation communication networks. Non-Orthogonal Multiple Access (NOMA) is also regarded as a promising approach to enable such networks offering higher flexibility and overall performance.

One of the significant contributions of this deliverable is the proposed integration of SemCom into a hybrid NOMA system, enabling a common radio resource to support both semantic and conventional bit-based users simultaneously. Semantic users are modelled using DeepSC equipped with pre-trained neural networks. For the proposed framework, at most two users share the same time-frequency resources dynamically for each time slot based on real-time channel conditions, merging stronger and weaker users without any fixed ordering. For this, the document provides a detailed system model, specifying innovative power allocation strategies and user detection techniques. More specifically, an optimization problem constraining on the available power is formulated to optimally allocate power among the heterogeneous users for maximizing the ergodic semantic SE following an innovative pairing strategy and bit-to-semantic decoding order in time slots with both types of users. In such slots, the bit user first decodes its own signal while treating the semantic

© 2023-2027 SCION Consortium. All rights reserved.

signal as interference. Afterward, the receiver decodes the semantic signal without residual interference. For the scenario when both semantic users are paired, channel reconstruction-based Successive Interference Cancellation (SIC) is deployed where the stronger signal is reconstructed by reversing the channel impairments until a target semantic similarity is achieved. Subsequently, this reconstructed signal can be subtracted from the received signal to decode the weaker semantic user's signal in an interference free manner. Simulations confirm the effectiveness of the proposed hybrid NOMA heterogeneous users setup under dynamic channel conditions.

The document refers to substantial challenges moving ahead in SemCom, such as handling diverse data types in multi-modal communication, semantic-aware resource allocation, and integration into resource-constrained environments. It further outlines future research directions, such as semantically enhanced pinching antennas, AI-native air interfaces, and energy-efficient system designs to improve SemCom's overall performance towards the transformative realization of 6G networks.

© 2023-2027 SCION Consortium. All rights reserved.

Document revision history

Date	Authors (partner's name)	Contribution
07.07.2025	Ishtiaque Ahmed (UESSEX), Yingzhuo Sun (KTH), Sicong Ye (KTH), Leila Musavian (UESSEX), Ming Xiao (KTH)	v1.0: First version
23.07.2025	S. Mumtaz (NTU), A. Köse (IYTE), B. Özbek (IYTE)	First Review
26.08.2025	All Authors	V1.8 Revision
08.09.2025	S. Mumtaz (NTU), A. Köse (IYTE), B. Özbek (IYTE)	Second Review
15.09.2025	All Authors	V1.9 Revision
01.10.2025	B. Özbek (IYTE)	V2.0 Final Version

© 2023-2027 SCION Consortium. All rights reserved.

Table of Contents

Executive Summary 3
Table of Contents 6
List of Figures 7
List of Tables 8
Abbreviations 9
1 Introduction 10
1.1 Foundations of SemCom 12
1.2 Classification of SemCom 13
1.3 Semantic Channel Capacity 14
1.4 Semantic Noise and Errors 15
1.5 SemCom in TOC 16
1.6 SemCom in Satellite Networks 17
2 Advances in Semantic Encoding and Decoding 19
2.1 Machine Learning for Semantic Extraction and Transmission 20
2.2 Semantic Similarity Function and its Approximation 21
2.3 Multi-modal SemCom 22
2.4 Performance Enhancement in Next-Generation Networks 23
2.5 Hybrid NOMA Heterogeneous Users Network 26
2.5.1 System Model 27
2.5.2 Optimal Power Allocation and Users Detection 30
3 Challenges and Future Research Perspectives in SemCom 37
4 Conclusions 39
5 References 40

© 2023-2027 SCION Consortium. All rights reserved.

List of Figures

Figure 1: Illustration of the semantic satellite relay communication system. 17
Figure 2: DL based SemCom using DeepSC model. 20
Figure 3: Illustration of the multi-modal SemCom system. 22
Figure 4: Secondary user ergodic semantic rate versus minimum bit rate requirement at the primary user in two-user uplink NOMA. 24
Figure 5: Secondary user ergodic semantic rate $P0$ in two-user uplink NOMA. 25
Figure 6: Time proportion distribution for the secondary user under different cases. 26
Figure 7: Users allocation in hybrid NOMA. 28
Figure 8: Downlink hybrid NOMA transmission and decoding strategy. 28
Figure 9: Users ergodic activity indicator versus $P_{\max}$. 33
Figure 10: Ergodic semantic SE versus $P_{\max}$ with $\epsilon_{\text{th}} = 0.9$. 33
Figure 11: Ergodic semantic SE versus ϵ_{th} for different $P_{\max}$. 35
Figure 12: Bit user ergodic SE versus $P_{\max}$ with $\epsilon_{\text{th}} = 0.9$. 35
Figure 13: Bit user equivalent ergodic semantic SE versus $P_{\max}$ for different ϵ_{th}. 36

© 2023-2027 SCION Consortium. All rights reserved.

List of Tables

Table 1: Explicit versus implicit semantic reasoning and their deployment trade-offs 12

Table 2: Received sentences via different encoding schemes over Rayleigh fading channels at SNR of 18 dB 14

Table 3: Comparative advantages and limitations of recent semantic encoding schemes 18

© 2023-2027 SCION Consortium. All rights reserved.

Abbreviations

AP	Access Point
AWGN	Additive white Gaussian noise
AoI	Age of Information
AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
BLEU	Bilingual Evaluation Understudy
BitCom	Bit-based Communication
BER	Bit Error Rate
CSI	Channel State Information
CIT	Classical Information Theory
CNNs	Convolutional Neural Networks
DL	Deep Learning
DLJSC	DL-based Joint Source-Channel Coding
IoT	Internet of Things
JSCC	Joint Source-Channel Coding
KKT	Karush-Kuhn-Tucker
KBs	Knowledge Bases
LoS	Line-of-Sight
LEO	Low Earth Orbit
MFMSC	Multi-modal Fusion based Multi-task SemCom
NLP	Natural Language Processing
NOMA	Non-Orthogonal Multiple Access
OMA	Orthogonal Multiple Access
OTFS	Orthogonal Time Frequency Space
QoS	Quality-of-Service
RIB	Robustness of the Information Bottleneck
SemCom	Semantic Communication
SINR	Signal-to-Interference-plus-Noise Ratio
SNR	Signal-to-Noise Ratio
SE	Spectral Efficiency
SIC	Successive Interference Cancellation
TOC	Task-Oriented Communication
TL	Transfer Learning
U-DeepSC	Unified multi-task SemCom
URLLC	Ultra Reliable and Low Latency Communication
VIB	Variational Information Bottleneck
VQA	Visual Question Answering

© 2023-2027 SCION Consortium. All rights reserved.

1 Introduction

This deliverable presents a progression towards the state-of-the-art work on semantic encoders and decoders. Specifically, we outline the key developments, persistent challenges, and potential research opportunities in the evolution of Semantic Communication (SemCom). Future general plan on how to further efficiently deploy SemCom in next-generation wireless networks is also provided in this deliverable.

Over the past few decades, mobile communication technology has evolved significantly, progressing from 1G in the 1980s to the recently commercialized 5G, with each generation introducing transformative advancements [1]. As wireless technologies continue to improve, the number of applications, connected users, and intelligent devices has grown substantially and is expected to rise even further in the coming years. Moreover, the current 5G cellular technology fails to meet the latency, reliability, and scalability requirements in 6G relevant scenarios. To meet the higher demands, there is an increasing need for more intelligent and efficient communication systems that can accommodate diverse Quality-of-Service (QoS) requirements while effectively managing limited network resources and infrastructure constraints [2]. With ongoing technological advancements, modern communication devices are becoming more context-aware and capable of intelligent decision-making. Projections suggest that by 2030, more than 125 billion reasoning capable devices will be connected through the Internet [1]. Researchers have also corroborated that the current communication systems are approaching the physical-layer limits proposed in Shannon's Classical Information Theory (CIT) [1]. With this, the conventional techniques would no longer be beneficial to meet such surging and diverse QoS needs within the limited communication resources. This necessitates a major transformation from basic digital connectivity to intelligent, meaning-driven network for facilitating the extensive exchange of data, and ultimately paving the way for SemCom. The underlying meaning of transmitted information is primarily the focus of SemCom [3], differing from the mere raw data exchange in traditional communication. With the aid of certain neural network architectures, advanced SemCom techniques work by extracting only the most important features of any message for its semantic reconstruction, and avoid the rest of details for an enhanced SE.

Researchers have now shifted focus towards attaining real-world performance demands for the 6G standard with the aim of transmitting contextual meaning instead of raw bits through the air. This meaningful data will be crucial for intelligent and reasoning-driven devices to operate effectively for the task required. For instance, the emerging satellite-terahertz links for broadcasting immersive streams to viewers [4], remote robotic surgery where the tactile feedback streams [5] must traverse within Ultra Reliable and Low Latency Communication (URLLC) precision, and the vehicle-to-everything communication while avoiding channel congestion [6]. Each of these use cases highlights the importance of transmitting only the task-critical meaning of the data, rather than its complete bit-level representation, to keep spectral usage, latency, and energy within acceptable budgets. These examples demonstrate the overarching demands of URLLC and congested device densities surpassing existing spectral provisions [7]. Classical link-level improvements have started to saturate with the approaching Shannon limit. Instead of stretching the physical layer, SemCom tackles 6G bottlenecks by eliminating symbols that are of no utility to the end task, reducing contention and allowing for more users. Essentially, semantic encoding reframes the communications problem to this resource-aware viewpoint by judiciously selecting few bits necessary for accurate semantic reconstruction [8].

© 2023-2027 SCION Consortium. All rights reserved.

Historically, communication was categorized into three levels, namely the technical level focusing on transmission of symbols (covered widely in traditional communication), semantic level, and the semantic-effectiveness level [8]. Traditional communication models, based on Shannon's probabilistic model, emphasize Bit Error Rate (BER) and channel capacity. Shannon's CIT relies on the mutual information in the entropy domain which is linked to the technical level, while overlooking the semantic and effectiveness levels. However, in many contemporary applications, especially those involving human–machine interaction and context-aware systems, the meaning behind the data is of paramount importance. This shift marks a departure from traditional communication models to meet the emerging demands of next-generation technologies, such as 6G and intelligent networked systems [1], [9]. The shift towards meaning-centric transmission is not completely new as in 1948, Shannon deliberately isolated the "technical" problem in his formulation of 1948. Weaver soon indicated semantic and effectiveness levels, although they remained hard to quantify [10]. A few years later, Bar-Hillel and Carnap presented the first logic-based quantification of semantic information, incorporating meaning into communication models [11]. Today's work on semantic encoders and decoders can be viewed as the follow-up of that tradition SemCom model, combining both link-level constraints and contextual value for developing heterogeneous-user communication schemes [12]. Researchers have made significant efforts incorporating the semantic and effectiveness level withing the classical Shannon's theory. One such efforts yielded a language model for extracting and estimating the semantic content from a collection of sentences [13], [11]. This provided the initial approach to quantify the semantic content, but lacked the accuracy of truthfulness or intent behind the transmission of data. For accurate semantic communication to occur between the sender and receiver, it was therefore inevitable to include the actual meaning and context in semantic transmission. This was achieved by the strong semantic information theory [14] which quantified the degree of discrepancy to better incorporate truth values in SemCom. Further advancements in SemCom models resulted in a collective framework comprising of a world model, background knowledge, inference, and message interpretation. With this shift towards the contextual meaning, SemCom is becoming crucial for designing intelligent systems.

With the shift towards semantic encoding and decoding, communication protocols are becoming more efficient in allocating resources through extracted semantic features for an improved overall performance even under unfavourable network conditions [15]. Moreover, studies suggest that SemCom offers superior performance in low Signal-to-Noise Ratio (SNR) conditions [14], where traditional Bit-based Communication (BitCom) becomes susceptible to noise. At higher SNR regimes, SemCom yields diminishing performance suggesting that it should be complemented with traditional Shannon communication for an enhanced overall performance [16]. The integration of semantic and bit-based communications within the same network defines heterogeneous users communication, where different users benefit from different approaches. But this integration of both types of communication, however will come with new design challenges. Particularly, implementation of interference mitigation strategies like SIC, becomes more complicated in scenarios with joint semantic and bit-based signals sharing the same communication resource. This is mainly due to the fact that semantic signals require pre-trained neural networks at both the transmitters and receivers for ensuring reliable reconstruction of the intended message [17].

© 2023-2027 SCION Consortium. All rights reserved.

1.1 Foundations of SemCom

Although Shannon first used the word "semantic" in his theory, but it lacked the interest of researchers during that time when the focus was on attaining high data rates. For that, CIT defines information based on entropy which is a measure of the uncertainty reduction in communication. As the Shannon's model was not taking into consideration the meaning of information, inefficiencies were recorded for the said model in context-aware, Artificial Intelligence (AI)-driven, and task-specific communication scenarios. Bar-Hillel in 1952 addressed this gap, by introducing a theoretical perspective on semantic information that focuses on the significance of conveyed meaning rather than the exact reproduction of raw symbols at the receiver [13]. In other words, semantic encoding transmits data based on its contextual intended meaning, thereby compressing the original message comprising of bit patterns. Consider the spoken command "Turn on the living-room light". In classical BitCom, every word (or text character) needs to be decoded perfectly at the decoder. However, in a semantic link, the receiver is only concerned with the recovery of the intent (action = activate, device = light, location = living-room). If background noise corrupts *"living-room"* into *"l·ving-room"* the semantic decoder can still map the token to the same intent, keeping semantic uncertainty low while bit-level uncertainty grows. Related to this example, [18] illustrates that SemCom is more resilient to symbol noise.

Building on Shannon's model, Weaver proposed an extension that highlighted the importance of semantic accuracy in communication, underscoring that the interpretation of messages can be just as critical as their error-free delivery [11]. Over time, researchers have developed additional metrics focusing on freshness of data at the receiver and are becoming indispensable in cyber-physical networks. Most prominent one is the Age of Information (AoI), which prioritizes time-sensitive and contextually relevant data over raw data rates [19]. Unlike traditional approaches prioritizing throughput, AoI emphasizes the prompt delivery of updates pertinent to specific tasks (e.g., sensor readings or haptic feedback). SemCom strives to reduce the AoI of meaningful symbols rather than mere bits. This alignment supports the objectives of 6G, including sub-millisecond URLLC and remote actuation. The AoI provides a measure of the freshness of data at any particular network node by measuring the amount of time that has elapsed since the last update was generated and received. By focusing on the meaning and timeliness of information, early research laid the foundational strategies for modern SemCom that aim to achieve a balance between reliability and efficiency within increasingly complex wireless networks.

Understanding the concepts of meaning and relevance of transmitted data become useful in building upon the CIT for SemCom and defining semantic entropy, Task-Oriented Communication (TOC), and semantic effectiveness. SemCom at the semantic level broadly considers two major aspects, the content semantics and channel semantics [1]. The former deals with the detection and extraction of pertinent information from the transmitted data and involves two reasoning approaches, namely explicit reasoning and implicit reasoning, as given in Table 1. The first one relies on structured Knowledge Bases (KBs) and predetermined set of rules, while the latter utilizes Deep Learning (DL) models and Transformer-based architectures for extracting and interpreting the semantics in data.

© 2023-2027 SCION Consortium. All rights reserved.

Table 1: Explicit versus implicit semantic reasoning and their deployment trade-offs

Aspect	Explicit Reasoning	Implicit Reasoning
Typical knowledge source	Curated ontologies and formal rules (rule-based)	Data-driven latent representations (DL-based)
Inference mechanism	Symbolic logic, graph traversal, rule chaining	Gradient-based neural inference, learned vectors
Strengths	High interpretability and consistency, low data demand	Automatic feature learning, scalability to unstructured data
Limitations	Brittle to unseen patterns, manual rule upkeep	Opaque decisions, large training data and compute
Example application	Regulated industrial IoT where safety rules must be auditable	Multimodal XR streaming with real-time context adaptation

The common background or context shared between the transmitter and receiver in any SemCom system is referred to as the KB. Consider, for example, a smart factory floor with hundreds of power-constrained sensors reporting the status of the machines. By pre-installing the same lightweight ontology in each node and at the edge gateway, each sensor reports a minimal semantic tuple conveying its state. Using the shared KB, the gateway instantly triggers the appropriate control action when a predefined condition is satisfied. Without the KB, each packet would require a lengthy self-description, inflating payload size and latency. This shows why tightly aligned KBs are pivotal for low-latency SemCom in dense Internet of Things (IoT) settings [20]. The second key component which is channel semantics, largely focuses on the characteristics influencing the semantic encoding process, like the wireless channel conditions and environmental factors, including Channel State Information (CSI) and the effects of surrounding noise. This requires adaptive coding schemes that can adjust under varying network conditions. Together, these components form the foundation SemCom strategies that not only compress data by transmitting only its core meaning but also adapt to network dynamics, making them well-suited for practical, large-scale applications.

1.2 Classification of SemCom

Authors in [21] classified SemCom into two types based on network topology, namely paired SemCom, and network SemCom. For attaining improved reliability, ultra-high transmission efficiency, and compatibility, paired SemCom is deployed which involves direct interaction between any two participating agents. The two sharing agents mutually learn through the KBs and achieve an "understand-before-transmit" capability, extracting and sending only the core semantic content through semantic-level reasoning. Networked SemCom, on the other

© 2023-2027 SCION Consortium. All rights reserved.

side involves more than two agents constituting a network with shared objectives and where each agent collaboratively shares to and learns from multiple agents. This flexible approach enables several intelligent modern era applications, like digital twins, virtual reality, autonomous vehicles, and the Metaverse.

1.3 Semantic Channel Capacity

The semantic channel capacity represents the limit of SemCom in the presence of noise, allowing for a transmission rate with minimal semantic errors within this limit. Analogous to bits in traditional communication, the semantic units (suts) quantifies the amount of conveyed meaning by a message for serving specific task (goal of communication) at the receiver [22] and acts as the cornerstone of SemCom. The semantic information is understood prior to transmission, and the semantic encoder generates messages based on defined strategies (common KB) by extracting features based on the required task. For transmission over the channel, the encoded messages in the forms of words are typically mapped to symbols carrying features rather than raw bits. According to the Semantic Channel Coding Theorem [22], minimizing semantic errors in a discrete memoryless channel requires both a defined channel capacity and an appropriate coding strategy. The effective capacity is shaped by factors such as inherent semantic ambiguity, the receiver's ability to interpret the message, and the common background knowledge shared between the sender and receiver. In essence, by carefully balancing these factors and using specialized algorithms for semantic encoding and decoding, one can ensure that the communication system minimizes semantic errors and maintains high efficiency, even when the underlying engineering channel operates at a low capacity.

The rate at which semantic information is transmitted with a prescribed accuracy is referred to as the semantic rate, which depends on the semantic similarity function $\epsilon(K,\gamma)$ with value ranging between zero and one [16]. This function quantifies the resemblance between original signal at the transmitter and reconstructed one at the receiver. For $\epsilon(K,\gamma)\rightarrow 1$, the signal will increasingly resemble the original bit-based signal rather than its reconstructed semantic counterpart [3]. As expressed in [16], the semantic rate for any semantic user is given as

$$R_S(K,L) = \frac{WI}{KL}\epsilon(K,\gamma), \tag{1}$$

where W is the channel bandwidth, I is the amount of semantic information in any transmitted message in units of semantic (suts), K represents the average number of semantic symbols transmitted for each word through deep learning based DeepSC system [3], γ represents the received SNR, and L denotes the number of words. The resulting $R_S(K,L)$ is therefore measured in units of suts/s/Hz. The DeepSC system consists of semantic encoder/decoder and channel encoder/decoder, equipped with neural networks for semantic features extraction from the transmitted and received text [3] The reason for deploying DeepSC is that it offers a clear robustness for accurate semantic reconstruction when compared with other benchmarks for sample sentence transmission over fading channels as in Table 2. Moreover, the sentence similarity offered by DeepSC is more human-friendly to

© 2023-2027 SCION Consortium. All rights reserved.

perceive as it is based on extracted meaning for a specified task [3]. On these grounds, we adopt DeepSC in the design approach for heterogeneous multi-user networks in the following section.

Table 2: Received sentences via different encoding schemes over Rayleigh fading channels at SNR of 18 dB

Transmitted sentence	it is an important step towards equal rights for all passengers.
DeepSC	it is an important step towards equal rights for all passengers.
JSCC	it is an essential way towards our principles for democracy.
Huffman + Turbo coding	rt is a imeomant step tomdrt equal rights for atp passurerrs.
Huffman + RS coding	it is an important step towards ewiral rlrsuo for all passengess.
Bit5 + Turbo coding	it is an yoportbnt ssep sowart euual qighd fkr ill passeneers.
Bit5 + RS coding	it iw an ymp!rdbnd stgo to!latds eq.al ryghts dkr alk pasengers.

Unlike Shannon's capacity, the semantic channel capacity is dependent on the shared knowledge base (KB) between the transmitter and the receiver which ultimately reduces the communication resources [22]. This forms an upper-bound based on the accurate features interpretation (meaning) at the receiver, instead of mere bit throughput as in Shannon's capacity which depends on the transmit SNR. This difference reflects a shift in the communication objective from reproducing the exact or approximate message to that of accurate interpretation of that message at the receiver, aligning meaning with the sender's intent [22]. Consequently, semantic capacity saturates once the intended meaning is recoverable under a given KB.

1.4 Semantic Noise and Errors

Sematic noise refers to the disturbance in the intended meaning of any message according to its context, which presents further challenges differing from conventional engineering communication. Such noise can stem due to the mistranslations between languages, cultural misunderstandings, or mistakes in unit conversions. Unlike traditional bit-level errors, discrepancies in meaning are more complex to detect and resolve. This "semantic noise" can lead to errors that are challenging to quantify and correct [19]. Semantic errors can be categorized into two types: a) unsoundness which arises when the message that is transmitted contains the truth but is misunderstood by the recipient, thus signifying a departure from the intended significance; b) incompleteness which results from a failure to

© 2023-2027 SCION Consortium. All rights reserved.

accurately convey the intended meaning when the transmitted message is incorrect, but the received message is genuine.

1.5 SemCom in TOC

Conventional communication systems have typically concentrated on the first level of communication [8], with separate design and optimization of components such as transmitters, receivers, encoders, and decoders based on various transmission goals. A significant breakthrough occurred when researchers parameterized encoders and decoders using Convolutional Neural Networks (CNNs) and implemented joint training approaches [23]. This development marked a paradigm shift toward DL-enabled communication models. Building upon this foundation, DeepSC emerged as a pioneering DL-based semantic communication system that specifically targets the second level of communication by focusing on semantic meaning recovery [3]. TOC represents an even more advanced level of communication, where the primary objective is to convey only task-relevant information while minimizing redundancy [24]. This approach requires establishing a shared KB between communicating parties and implementing joint training methods. The benefits of TOC are manifold, including improved communication efficiency through reduced data transmission and enhanced privacy protection by transmitting representative features instead of raw data. However, practical implementation of TOC systems faces significant challenges, particularly regarding computational resource limitations in edge devices [25]. While devices with powerful graphics processing units can leverage large neural networks for optimal performance, resource-constrained devices such as unmanned aerial vehicles, mobile phones, and wearable sensors require more efficient computational methods. Another critical challenge in TOC systems involves handling diverse task complexities [26]. Different tasks, and even different instances within the same task category, can vary substantially in their computational demands. Current TOC models generally lack the capability to automatically recognize and adapt to varying task complexities, resulting in inefficient computation resource allocation. This limitation becomes particularly apparent in applications like image classification, where some samples may be straightforward to process while others require significantly more computational effort.

TOC inherently involves a trade-off between correctness and compactness [24]. Several theoretical frameworks have been developed to understand and optimize this trade-off. The Variational Information Bottleneck (VIB) has been proposed as a theoretical upper bound for analyzing the rate-distortion trade-off in DL-enabled communication methods [27]. Subsequent research introduced the Robustness of the Information Bottleneck (RIB) concept to improve communication robustness against channel variations [28]. The theoretical framework has been further extended to multi-task [29] and multi-device scenarios [30], demonstrating the versatility of the approach. These diverse approaches highlight the ongoing innovation in developing TOC systems.

Resource allocation optimization has attracted significant research interest for improving semantic and TOC systems. Several innovative metrics and approaches have been proposed, including semantic SE for optimizing channel assignment and semantic symbol transmission [16]. Joint optimization of compression ratios and resource allocation has been investigated to maximize task success probability [31], while semantic-bit quantization has been introduced to combat semantic noise and enable non-guided adaptive resource allocation [32]. However, these studies primarily focus on allocating channel resources among multiple users at the semantic level, with computation resource allocation in neural network-based

© 2023-2027 SCION Consortium. All rights reserved.

SemCom and TOC models receiving comparatively less attention. Dynamic neural networks have emerged as a promising solution, offering robust representational capabilities and enhanced interpretability [33]. These networks can be broadly categorized based on their dynamic architectures, with dynamic depth being one prominent example. The flexibility and adaptability of dynamic neural networks make them particularly suitable for the evolving demands of semantic and TOC systems.

1.6 SemCom in Satellite Networks

The convergence of SemCom and AI is driving a paradigm shift in satellite communications, offering transformative solutions to long-standing challenges. Traditional satellite systems struggle with constrained bandwidth, harsh propagation conditions, and limited onboard processing capabilities, particularly in emerging 6G non-terrestrial networks. SemCom addresses these limitations through intelligent information prioritization, where only the most relevant data features are extracted and transmitted rather than raw bitstreams. This fundamental shift from syntactic to semantic transmission yields two key advantages: significantly reduced bandwidth requirements and improved resilience to channel impairments, particularly in low Signal-to-Interference-plus-Noise Ratio (SINR) scenarios common in satellite links. In addition, SemCom frameworks can define tailored performance metrics for various data formats and fuse multimodal information into compact, high-level representations. Critical enabling technologies include Joint Source-Channel Coding (JSCC) for integrated and efficient encoding, as well as generative AI techniques such as diffusion models that not only produce data ideally suited for semantic-driven transmission but also offer strong resistance to channel impairments. Furthermore, these diffusion-based models can adaptively equalize the channel more effectively than traditional neural networks. Finally, adaptive deep semantic strategies enable real-time adjustments to compression and transmission processes, optimizing resource usage even under stringent satellite link constraints. Taken together, these innovations point toward a more robust, bandwidth-efficient, and intelligent satellite communication ecosystem.

Inspired by the existing literature, we propose a satellite communication system based on the semantic relay and semantic encoder/decoder. The shared KB design is vital for the SemCom system. However, it is difficult for the transmitter and the receiver to own the exactly same KB in the satellite communication scenario because of the wide coverage area and the large amount of nodes. Another challenge brought by the giant network is the severe cascaded fading. To overcome these two difficulties, the semantic relay satellite nodes and KB coordinator satellites [34] can be used as illustrated in Figure 1. The JSCC encoder and the semantic encoder are employed at the transmitter. The transmitted signal can be enhanced semantically based on the shared KB at the relay satellite.

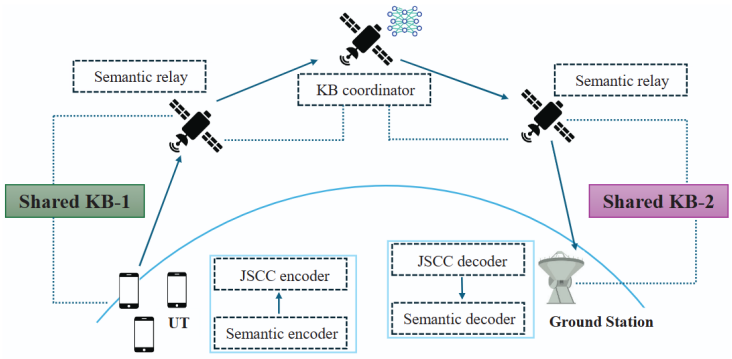


Figure 1: Illustration of the semantic satellite relay communication system [29].

© 2023-2027 SCION Consortium. All rights reserved.

A challenge brought by this SemCom-based satellite systems lies in ensuring KB consistency across spatially distributed and mobile nodes, due to the wide coverage area and the large number of satellites and terminals. To address this, dedicated KB coordinator satellites are introduced into the architecture. These satellites act as control nodes that manage and synchronize knowledge updates across the network. Specifically, they maintain a global KB version and periodically broadcast KB update vectors or deltas to relay nodes and user terminals. Potentially and technically, this can be achieved through a combination of semantic hash indexing and federated learning techniques. Semantic hash indexing enable lightweight consistency checks to verify KB alignment across nodes. Federated learning can assist this by local updates from satellites or ground terminals are aggregated and harmonized at the KB coordinator before redistribution.

Recent advances demonstrate the practical implementation of SemCom principles across various satellite architectures. For Earth observation systems, semantic-aware frameworks incorporating quality-of-service constraints have achieved 35% reductions in transmission latency while maintaining critical data integrity [35]. These systems employ hybrid optimization techniques combining metaheuristic algorithms with matching theory to dynamically allocate limited satellite resources. In Low Earth Orbit (LEO) constellations, semantic routing protocols have shown particular promise by incorporating KB compatibility into forwarding decisions [36]. Unlike conventional approaches where the focus is solely on link metrics, these semantic-aware strategies reduce end-to-end delays by 40% through intelligent data prioritization and redundant transmission elimination.

The integration of semantic techniques with advanced physical layer technologies further enhances system performance. The combination with Orthogonal Time Frequency Space (OTFS) modulation, for instance, addresses the severe Doppler effects characteristic of LEO satellite channels while maintaining semantic transmission efficiency [37]. Experimental results demonstrate this hybrid approach delivers 2.1 times the SE of traditional OTFS systems. Generative AI represents another breakthrough, with foundation models enabling content-aware compression and reconstruction that reduces bandwidth requirements by up to 60% compared to conventional methods [38]. These models employ adaptive encoding mechanisms that dynamically optimize transmission parameters based on both semantic content and real-time channel conditions.

Implementation challenges remain, particularly regarding KB synchronization across distributed satellite networks and standardization of semantic performance metrics. The proposed semantic relay architecture, incorporating dedicated coordinator satellites, offers a promising solution for maintaining consistency across wide coverage areas. Future developments are expected to focus on lightweight implementations suitable for resource-constrained small satellites and cross-layer optimization strategies that harmonize semantic processing with existing network protocols. As these technologies mature, SemCom is poised to become a cornerstone of next-generation satellite networks, enabling more efficient and intelligent global connectivity solutions.

© 2023-2027 SCION Consortium. All rights reserved.

2 Advances in Semantic Encoding and Decoding

Semantic encoding and decoding have evolved rapidly from the early transformer-based designs toward more adaptable and efficient models for specific tasks. DeepSC stands out to be the most well-validated model where the newer architectures widely relies on its core principles for enhancing efficiency. DeepSC preserves semantic fidelity even under low-SNR for the fading channels conditions and outperforms the conventional separate source-channel designs [39]. Recent approaches aim at strengthening the neural network architectures for compact sentence-embeddings that lowers down the overall symbol usage without compromising on the semantic fidelity at the receiver. Multi-modal transformers extend the principles of semantic encoding to cross-domain settings for text, image, and other data types [39]. Parallel research explores generative AI models as semantic priors where large language models to improve semantic reconstruction at moderate SNRs with exceptional generalization ability to multi-modal SemCom [40]. Diffusion-based encoders, initially introduced for generating high-quality images, are now utilized in SemCom which enables progressive refinement when symbol budgets are small, or packets are partially decoded [41]. The iterative removal of noise due to the training process adopted in diffusion-based model effectively maintains the semantic fidelity at the receiver. Another promising approach is the KB or ontology-assisted SemCom, where both the transmitter and receiver exploit the shared knowledge graphs to mitigate errors at the symbolic level as discussed in [42]. More specifically, by representing domain-specific knowledge through ontologies, SemCom enables the receiver to reconstruct the corrupted information. Moreover, techniques like pruning, quantization, distillation, and unified decoders make SemCom feasible choice for lightweight devices with low-computing and memory [43]. These advances result in higher semantic accuracy making them directly relevant to the hybrid NOMA framework, where DeepSC assisted network supports heterogeneous users pairing and power allocation for enhancing semantic SE. Table 3 highlights some key recent advancements in semantic encoding.

Table 3: Comparative advantages and limitations of recent semantic encoding schemes

Model	Advantages	Typical Limitations
DeepSC (Transformer JSCC) [39]	High sentence-level and semantic similarity; robust under low SNR; end-to-end training	Requires bulk of training data; large model size
Sentence-BERT–style JSCC [39]	Smaller payload; high semantic fidelity	Context loss in extreme compression
LLM-driven SemCom [40]	Strong generalization; improved reconstruction at moderate SNR; multimodal support	High computational cost; model adaptation needed
Diffusion-based [41]	Robust to packets loss; perceptual quality preservation	High latency for iterative decoding
KB-based (Ontology-assisted) [42]	Task relevance semantic accuracy; reduces ambiguity	Requires domain-specific KB

© 2023-2027 SCION Consortium. All rights reserved.

Lightweight device [42]	Low memory requirements; suitable for IoT and wearable devices	Reduced capacity

2.1 Machine Learning for Semantic Extraction and Transmission

In recent years, DL has driven significant progress in SemCom strategies, particularly through DL-based Joint Source-Channel Coding (DLJSC). One prominent development is the DeepSC model, which leverages Bidirectional Encoder Representations from Transformers (BERT) and other Transformer-based architectures to capture and reconstruct semantic information more effectively [44]. Transformers process input sequence in parallel (requiring less time), making it highly efficient for training. BERT is a large pre-trained model trained on billions of sentences to extract semantic information. This enables its deployment suitability across various contextual scenarios by generating distinct semantic vectors for each. Techniques such as Sentence-BERT and DeepSC integrate encoders and decoders in a single training framework. By optimizing both components together, these methods achieve robust semantic transmission in challenging network conditions [9].

Figure 1 shows the jointly trained DL based transmitter and receiver in the DeepSC model. The transmitter comprises of semantic encoder and channel encoder, respectively responsible for extracting semantic content from the input data, and ensuring reliable transmission over the physical channel. Accordingly, the receiver consists of channel and semantic decoders for reconstructing the transmitted symbols and sentences, respectively. The training of transmitter and receiver is done with KB, which differs for various applications and data requirements.

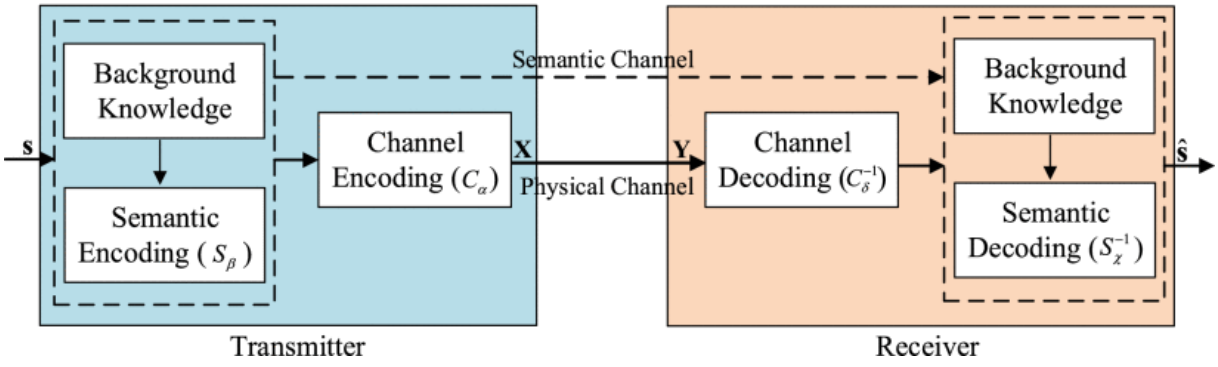


Figure 2: DL based SemCom using DeepSC model [3].

The reason why DeepSC (based on BERT) is heavily deployed in characterizing the performance of SemCom is that most of the other DL enabled approaches to extract semantic information are designed for specific tasks and require redesign whenever the task changes. Moreover, Natural Language Processing (NLP) enables machines to comprehend human language by using joint probability models that consider context. Consequently, language models leverage this contextual information to differentiate among words and phrases with similar meanings, even though they may fail in fully capturing syntactic details. To ensure DeepSC is applicable to different communication environments, Transfer Learning (TL) is deployed which is a machine learning technique by which knowledge learned from one

© 2023-2027 SCION Consortium. All rights reserved.

(previous) task is used for a second (future) task [45]. For the above reason, the DeepSC model leveraging TL has been introduced to as a universal tool to generate word vectors for a broad array of NLP tasks without necessitating any further redesign.

Since its proposition for text communication [3] where it outperformed conventional methods, researchers have been putting efforts to utilize DL based SemCom for different scenarios. DeepSC has been extended for multi-user text transmission in [46] using dedicated flag vectors instead of CSI feedback, and with uniform time and energy consumption even for increasing number of users. Authors of [47] utilized DeepSC for speech transmission using text-related features which reduces data transmission, and the speech is synthesized back by feeding the text and sender information into neural network. Image transmission using SemCom has been studied in [48] with the aid of DL based classifier and diffusion model, and this approach is extended for video transmission by converting videos into frames [49].

2.2 Semantic Similarity Function and its Approximation

The basic model of DeepSC consists of semantic and channel encoders/decoders, which are empowered by neural networks [3]. Basically, these neural networks at the DeepSC transmitter extracts the semantic features from the sentence for transmission and maps them to semantic symbols. The channel decoder and semantic decoder at the DeepSC receiver recover the original sentence. Let us denote the transmitted and recovered sentences by $\mathbf{s}$ and $\hat{\mathbf{s}}$. Let us denote the semantic similarity between $\mathbf{s}$ and $\hat{\mathbf{s}}$ by $\eta(\mathbf{s}, \hat{\mathbf{s}})$, and use the cosine similarity to measure it which ranges from o to one, where a higher value means higher similarity. Cosine similarity is one of the NLP techniques for approximating how similar two word/sentence vectors are and defined as the dot product of the vectors divided by the product of their magnitudes.

In the context of text semantic transmission, traditional metrics like BER is unable to provide an accurate reflection of performance. Instead of relying solely on human judgment to gauge the similarity between sentences, the Bilingual Evaluation Understudy (BLEU) score is commonly used in machine translation. However, the BLEU primarily measures differences at the word level, rather than capturing the underlying semantic meaning of the sentences. Thus, we can compute $\eta(\mathbf{s}, \hat{\mathbf{s}})$ using the BERT model and estimate the match level or distance of semantic information between s and $\hat{s}$ using the cosine similarity metric.

$$\eta(\boldsymbol{s}, \hat{\boldsymbol{s}}) = \frac{\boldsymbol{B}(\boldsymbol{s})\boldsymbol{B}(\hat{\boldsymbol{s}})^T}{\|\boldsymbol{B}(\boldsymbol{s})\|\|\boldsymbol{B}(\hat{\boldsymbol{s}})\|}, \tag{2}$$

The sentence-BERT model is represented by $\mathbf{B}(.)$, and T denotes the transpose and is used to make the dimensions of s and $\hat{s}$ consistent for dot product. Note that $\eta(\mathbf{s}, \hat{\mathbf{s}})$ depends on the neural network structure of DeepSC and channel conditions. Therefore, it can be expressed as a function of K and γ.

$$\eta(\boldsymbol{s}, \hat{\boldsymbol{s}}) = \epsilon(K, \gamma), \tag{3}$$

In practice, the closed-form expression of $\epsilon(K, \gamma)$ is not available, so the generalized logistic approximation is adopted via data regression on DeepSC outputs. Specifically, for each K the DeepSC tool is run over a grid of γ values to obtain empirical $\epsilon(K, \gamma)$ samples. Running the DeepSC with varying values of K and γ, $\epsilon(K, \gamma)$ was found to be monotonically non-decreasing with γ [16]. Moreover, its gradient change increases first with γ and then decreases. This pattern suggests that the fitted curve for $\epsilon(K, \gamma)$ should look like the sigmoid curve bounded in [0, 1]. A sigmoid function is a bounded, differentiable, real function that is

© 2023-2027 SCION Consortium. All rights reserved.

defined for all real input values and has a non-negative derivative at each point and exactly one inflection point. Authors in [12] deployed the data-regression method to tractably approximate the values of $\epsilon(K,\gamma)$ by following the criterion of minimum mean square error for fitting the values with a generalized logistic function (common form of sigmoid function), as expressed by

$$\epsilon(K,\gamma) \triangleq A_{K,1} + \frac{A_{K,2}-A_{K,1}}{1+e^{-\left(C_{K,1}\gamma+C_{K,2}\right)}}, \tag{4}$$

The lower (left) asymptote, upper (right) asymptote, growth rate, and the mid-point parameters of the logistic function are respectively denoted by $A_{K,1}$, $A_{K,2}$, $C_{K,1}$, and $C_{K,2}$ for different values of K. The logistic function provides a differentiable and close approximation for tractable analysis and optimization as backed by the results in [12]. To ensure that semantic communication achieves desired accuracy at the receiver, $\epsilon(K,\gamma)$ must meet a minimum threshold ϵ_{th}, below which the transmission will be treated as invalid. This is represented as

$$\epsilon_K(\gamma) = \begin{cases} \epsilon(K,\gamma), & \text{if } \epsilon(K,\gamma) \geq \epsilon_{\text{th}} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

The approximated generalized logistic function depends explicitly on γ, while the effect of K is captured through the parameters $A_{K,1}$, $A_{K,2}$, $C_{K,1}$, and $C_{K,2}$.

2.3 Multi-modal SemCom

The evolution of SemCom has progressed from transmitting isolated words or images to integrating multiple modalities, such as text, images, audio, and video. This advancement has given rise to multi-modal SemCom systems, which are designed to handle and process diverse types of data and tasks. As illustrated in Figure 3, a typical multi-modal SemCom system involves several key steps: first, the collection of data from different modalities; next, a central fusion process that integrates these heterogeneous inputs into a unified semantic representation; and finally, the encoded semantic message is distributed to multiple users.

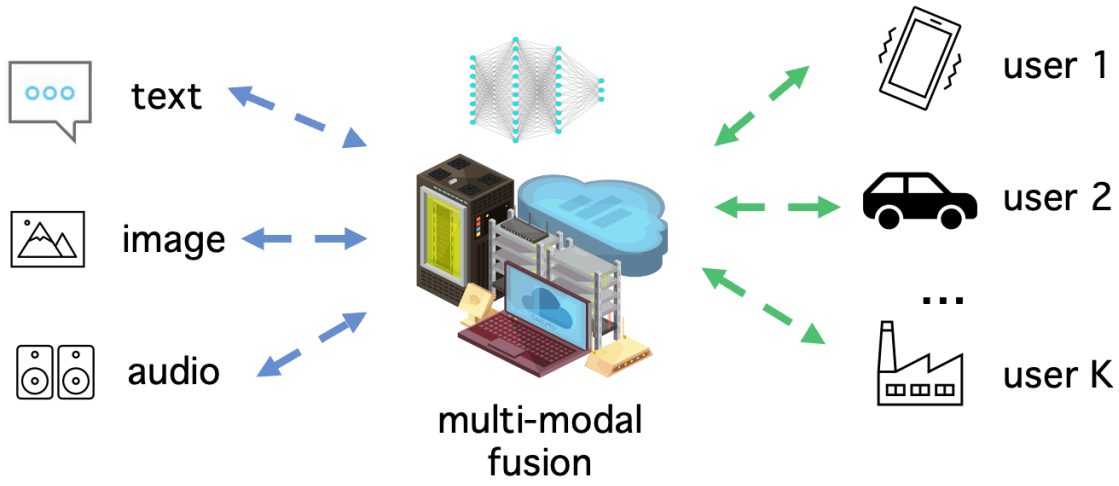


Figure 3: Illustration of the multi-modal SemCom system.

One such system is the task-oriented Multi-User SemCom for Visual Question Answering (VQA) framework, known as MU-DeepSC, which integrates multi-modal data from multiple users to enhance VQA performance [50]. By jointly optimizing transmitters and receivers, MU-DeepSC effectively captures correlated features across different modalities,

© 2023-2027 SCION Consortium. All rights reserved.

demonstrating resilience against channel variations, particularly in low SNR scenarios. Similarly, the Unified multi-task SemCom (U-DeepSC) system offers a flexible framework for handling multiple tasks across various data modalities [51]. It employs a dynamic feature adaptation mechanism that adjusts transmitted data based on specific tasks and channel conditions, thereby improving transmission efficiency. Additionally, the Multi-modal Fusion based Multi-task SemCom (MFMSC) framework utilizes BERT to effectively integrate semantic information from different modalities [52]. This approach not only enhances performance but also reduces communication overhead in complex multi-task environments.

2.4 Performance Enhancement in Next-Generation Networks

The advancements in wireless communication networks has resulted in an exponential growth of wireless subscribers which poses a challenge to the limited bandwidth. To overcome the limitations associated with conventional communication systems, particularly that of approaching the Shannon's limit, SemCom paradigm has been examined. Based on the transmission of semantic content rather than raw bits, SemCom enhances the SE by mitigating the transmission overhead. This paradigm, however, is best suited than the conventional BitCom at lower SNR requirements, as BitCom can recover the data with nearly exact precision at higher SNR regimes. It is therefore viable to deploy both BitCom and AI-based methods like SemCom for next-generation communication networks [53].

To enable massive connectivity, multiple access techniques play a critical role, especially to enable next-generation intelligent communication. Through the combination of semantic processing and advanced multiple access techniques, heterogeneous networks stand to benefit from greater flexibility and improved performance. One of the promising techniques enhancing the SE is the Non-Orthogonal Multiple Access (NOMA) where users share the same resource either through code or power-domain multiplexing. Although concerns relating to the fairness issues and increased complexity in implementation of SIC pose challenges for NOMA, integrating it with SemCom can assist in the design of heterogeneous users communication networks [19].

We compare the performance of SemCom and BitCom for the two users uplink NOMA setup, with single antenna antenna each at the Access Point (AP) and users. Specifically, the user nearest to the AP communicates using traditional BitCom and is named primary user. The farther user termed as secondary, reuses the same resource block used by the primary user and can flexibly utilize either BitCom or SemCom based on the channel conditions for optimal communication policy. We assume the primary and secondary users are at the distance of 15 meters and 45 meters from the AP and are sharing the resource block via NOMA. These distances are typical for a conventional near-far user scenario in NOMA, where the closer one enjoys better channel. This establishes the NOMA trade-off and benefits the deployment of SemCom at the secondary user. The uplink NOMA system holds for different distances with the same overall behavior, specifically the superiority of SemCom at lower SNR requirements.

For DeepSC with $K = 5$, the logistic approximation parameters are $A_{K,1} = 0.37$, $A_{K,2} = 0.98$, $C_{K,1} = 0.2525$, and $C_{K,2} = -0.7895$, providing high accuracy and a tractable form. To ensure high-quality semantic transmission, we set $\epsilon_{\text{th}} = 0.9$ for the semantic similarity-based reconstruction. The channel from each user to the AP is modelled as quasi static block-fading, where it remains constant over a fading block and varies independently between blocks. Mathematically, the independently distributed Rayleigh fading channel coefficients

© 2023-2027 SCION Consortium. All rights reserved.

are determined based on the distance-dependent path-loss model $\rho = \rho_0 \left(\frac{1}{d}\right)^{\beta}$. We assume a reference path-loss of $\rho_0 = -30$ dB at 1 meter, and the path-loss exponent $\beta = 4$, with d denoting the link distance in meters. The AP experiences additive white Gaussian noise (AWGN) with zero-mean and -80 dBm variance. We want to maximize the ergodic semantic rate of the secondary user under the minimum ergodic rate constraint of the primary user. For BitCom, we set the average number of bits per word to $\mu = 40$ and its semantic similarity as $\epsilon_C = 1$ to represent error-free reconstruction by the bit user. The ratio I/L is set to a constant value of 1 in our simulations. For a simpler analysis, we assume that the secondary user accesses the resources in an on-off approach by either transmitting with constant power or not accessing the resource at all. Here $P0$ is the fixed "on" transmit power of the secondary (far) user in the adopted access model whenever it is active. The primary user, however transmits with a constant power of 1 watt at all times. The results are averaged over 10^5 Monte Carlo channel realizations. For simulations, the transmission algorithm follows the approach discussed in [54].

Figure 4 illustrates the ergodic semantic rate achieved per unit bandwidth by the secondary user versus the minimum ergodic bit rate of the primary user within the NOMA uplink setup. We compare the opportunistic (BitCom or SemCom) scheme at the secondary user with SemCom-only and BitCom-only schemes in a two-user scenario. Notably, as the required minimum ergodic bit rate of the primary user increases, the ergodic semantic rate of the secondary user decreases across all the compared schemes due to the limiting available resources for the secondary user with the stricter rate requirements at the primary user. Among the compared schemes, the opportunistic scheme consistently offers the highest ergodic semantic rates. This is due to the adaptive nature of the opportunistic scheme by which the best communication method is selected for higher SE. Additionally, for transmit power $P0 = 1$ W at the secondary user, the SemCom-only scheme offers similar performance to the opportunistic scheme but significantly outperforms the BitCom-only scheme, reflecting the advantage of SemCom in low SNR regimes.

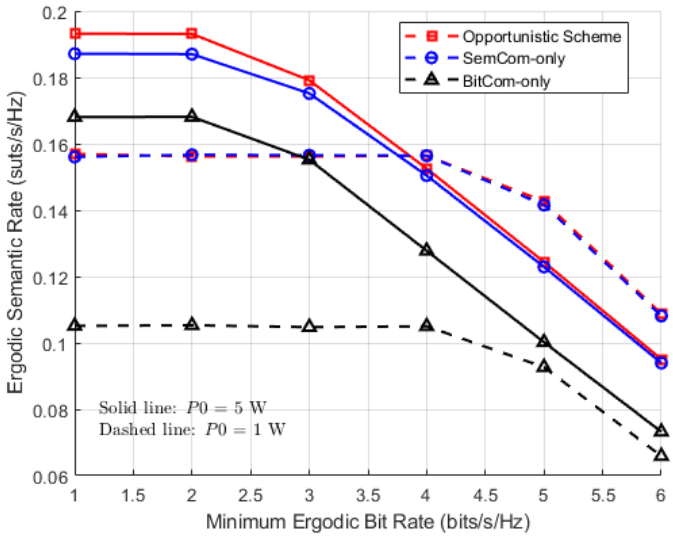


Figure 4: Secondary user ergodic semantic rate versus minimum bit rate requirement at the primary user in two-user uplink NOMA.

Figure 4 shows the ergodic semantic rate achieved at the secondary user versus $P0$ at the secondary user under minimum ergodic bit rate values ($\overline{R} = 2$ bps/Hz and $\overline{R} = 6$ bps/Hz) for the opportunistic, SemCom-only, and BitCom-only schemes. With increasing $P0$ values, the

© 2023-2027 SCION Consortium. All rights reserved.

ergodic semantic rate across all schemes improves for the secondary user at $\overline{R} = 2$ bps/Hz. The opportunistic scheme offers best performance at both $\overline{R} = 2$ bps/Hz and $\overline{R} = 6$ bps/Hz minimum ergodic bit rate requirements, effectively selecting either SemCom or BitCom for communication. At $\overline{R} = 2$ bps/Hz, the performance of SemCom-only closely matches the opportunistic scheme, suggesting the effectiveness of SemCom in low-to-moderate bit-rate requirements. However, at $\overline{R} = 6$ bps/Hz, SemCom-only stops offering improvement with an increase in $P0$. The BitCom-only scheme shows significantly underperformance compared to the opportunistic and SemCom-only schemes, particularly at lower $P0$.

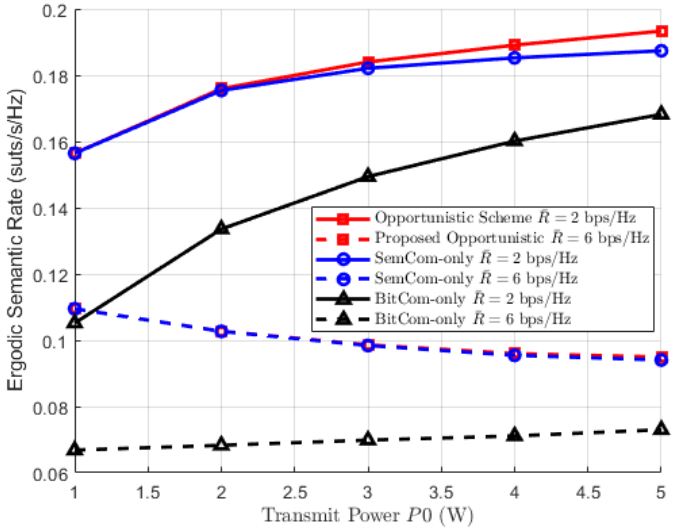


Figure 5: Secondary user ergodic semantic rate $P0$ in two-user uplink NOMA.

Figure 4 presents the time proportions for the secondary user during which it remains Off, deploys BitCom, or SemCom under different cases of $P0$ and $\overline{R}$ values. At $\overline{R} = 2$ bps/Hz and $P0 = 1$ W, the time portion for SemCom mode at the secondary user is higher than that at $P0 = 5$ W, highlighting the suitability of SemCom in low-power scenarios. Likewise, BitCom becomes more suitable at high-power values as in the case of $\overline{R} = 2$ bps/Hz and $P0 = 5$ W. This also calls for sensitivity of the primary user to interference from the secondary user, resulting in restricting the secondary user's activity at higher $P0$. Moreover, the secondary users have to remain Off for a majority portion of time as $\overline{R}$ of the primary user increases to mitigate the interference caused to the primary user, which limits the ergodic semantic rate of the secondary user. This suggests of a clear trade-off between interference mitigation and achieving higher ergodic semantic rates.

© 2023-2027 SCION Consortium. All rights reserved.

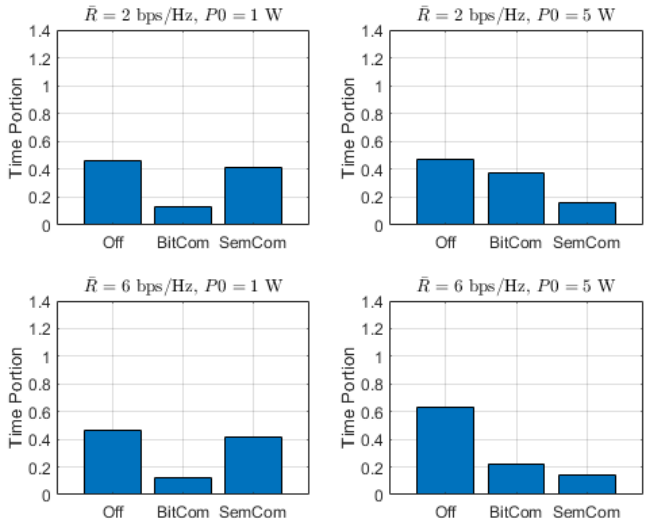


Figure 6: Time proportion distribution for the secondary user under different cases.

Although the results are for a two-user uplink NOMA, the same trade-offs between SemCom and BitCom extend to multi-user networks. The hybrid NOMA framework discussed afterwards schedules at most two users per time slot with optimal power split to maximize the equivalent semantic SE via heterogeneous users networks. Additionally, integrating semantic processing with flexible NOMA offers an approach for scalable deployment across diverse use cases related to IoT and other 6G services.

2.5 Hybrid NOMA Heterogeneous Users Network

To support massive connectivity, next-generation communication systems necessitate advanced multiple access techniques that push the boundaries of classical schemes such as Orthogonal Multiple Access (OMA) [55]. In this respect, NOMA is considered a promising technology for enhancing the SE of wireless systems by allowing multiple users to share the same time-frequency resources through power or code-domain multiplexing. However, NOMA also presents challenges which need to be carefully addressed, such as increased interference management complexity and fairness concerns [56], [57]. The increase in achievable SE using NOMA depends on the specific system configuration and use case, as NOMA does not always outperform orthogonal schemes, particularly in heterogeneous users communications [19]. In traditional NOMA, SIC is performed based on channel conditions, where the stronger user decodes the signal of the weaker user to mitigate interference. However, this approach cannot be directly applied in joint bit and semantic users communication systems. This is primarily due to the requirement of pre-trained neural networks at the semantic encoder and decoder for successful transmission and reception [3]. Consequently, the bit user is unable to decode the semantic signal from the superimposed signal, as traditional BitCom does not require any prior training at the transmitter or receiver. In contrast, the semantic user can decode the bit-based signal, resulting in a bit-to-semantic decoding order for joint bit and semantic users communication [58].

Given the differences in the signal processing requirements of BitCom and SemCom, a more flexible approach is needed to optimize users pairing and interference management [59]. Another technique that can efficiently be deployed to integrate the SemCom and BitCom is the hybrid NOMA which allows to capture the benefits of both OMA and traditional NOMA by allowing smooth transitions between these two schemes [60]. Essentially, the hybrid

© 2023-2027 SCION Consortium. All rights reserved.

NOMA enables multiple users to share the same resource block across different time slots, which ultimately enhances the overall system capacity. However, optimal power allocation and managing the complexities associated with SIC in hybrid NOMA framework remain important components for further investigation. Efficient power distribution can not only minimize interference between bit-based and semantic users but also ensures that both user types achieve their required performance targets. By strategically allocating power among bit-based and semantic users, the system not only achieves an improved semantic SE with minimal interference between users but also ensures the minimum rate requirements for BitCom.

The works in [58] and [61] have advanced the integration of semantic and bit-based communications and the deployment of hybrid NOMA. In [58], a semi-NOMA framework is proposed that investigates the trade-offs between semantic and bit-based communications in a shared transmission scheme. However, this approach does not exploit hybrid NOMA with user pairing based on channel conditions, limiting its adaptability in dynamic wireless environments. In [61], a multi-user hybrid NOMA scheme is introduced that optimizes power allocation and minimizes energy consumption under latency constraints. However, this work focuses solely on bit-based communication and does not incorporate semantic communication.

Inspired by the above works, we harness the benefits of heterogeneous users by adopting hybrid NOMA and a tailored decoding strategy that improves interference management. The key contributions of our work are:

i) We propose a hybrid NOMA framework for heterogeneous users that dynamically assigns bit and semantic users to time slots based on their channel conditions. To efficiently implement this, we develop a user pairing and decoding strategy, outlined in Algorithm 1 which enhances SE while ensuring adaptive resource allocation.

ii) Unlike previous works that assume a fixed decoding order, we deploy an adaptive decoding strategy that optimally manages interference in hybrid NOMA. In particular, our approach ensures a bit-to-semantic decoding order in time slots where both bit and semantic users coexist, while applying stronger-to-weaker SIC decoding in time slots containing only semantic users.

iii) We formulate and solve a power allocation optimization problem that maximizes the equivalent ergodic semantic SE with a constraint on the total available power.

iv) We demonstrate through simulations that the optimization prioritizes bit user in a heterogeneous users network by allocating more power, particularly at higher available power values.

2.5.1 System Model

We consider a downlink hybrid NOMA communication system consisting of a bit user (B) and two semantic users (S1 and S2), sharing the entire time-frequency resources. While this setup is chosen for analytical clarity, the proposed framework applies to any number of semantic users. In our proposed approach, we allow at most two users to share any slot, as shown in Figure 4.

© 2023-2027 SCION Consortium. All rights reserved.

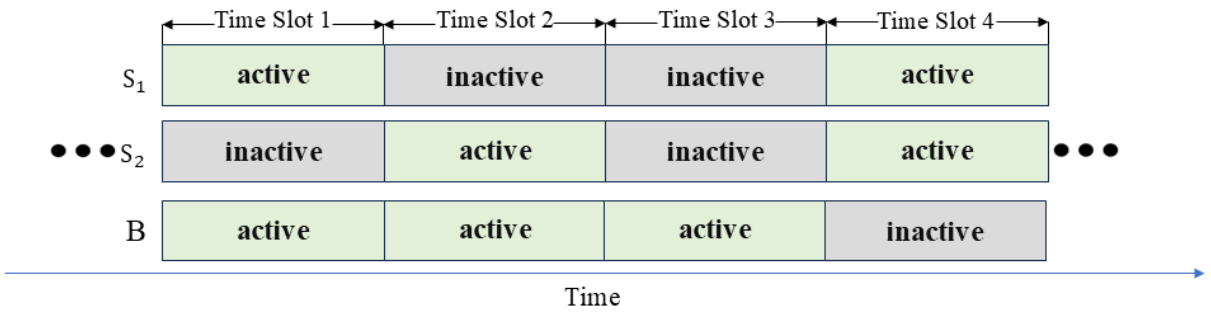


Figure 7: Users allocation in hybrid NOMA.

Communication between the AP and the users occurs over Rayleigh fading channels, where the channel gains remain constant within each time slot but may vary across time slots. At the receiver, AWGN with mean zero and variance σ^2 is assumed to impair the superimposed signal. We deploy the DeepSC model in our hybrid NOMA assisted system that can be generally applied to any semantic architecture. The considered DeepSC neglects potential errors arising from finite neural network capacity, limited training data, and variations in real-time inference accuracy.

Let $P_{\max}$ be the total available power budget at the AP. The channel gain coefficients from the AP to users B, S_1, and S_2 are denoted by h_B, h_{S1}, and h_{S2}, respectively, while the transmit powers allocated to them are P_B, P_{S1}, and P_{S2}. The semantic rate for any semantic user (in suts/s/Hz) is given as

$$R_{Si}(K,L) = \frac{WI}{KL}\epsilon(K,\gamma), i \in \{1,2\}. \tag{6}$$

We denote the rates of users S_1, S_2 and B by R_{S1}, R_{S2}, and R_B, respectively, where R_{S1} and R_{S2} are in suts/s/Hz, while R_B is in bits/s/Hz. The rate of User B is determined on the basis of Shannon's capacity formula. Perfect channel state information is assumed to be available at the AP for simplicity and tractability. However, achieving perfect CSI is difficult due to channel estimation error, feedback delay, or limited signaling overhead. To ensure efficient decoding of signals, the AP informs the receiver about the type of active users in each time slot via embedded control information in the headers of transmitted data packets. Specifically, this control information is encoded in the media access control sublayer using compact fields specifying user pairing, power allocation, and decoding order while resulting in an overhead of less than 5% relative to typical packet size, with minimal impact on SE [62]. Moreover, reusing the same decoding order across multiple successive slots further reduces the average overhead.

Our goal is to maximize the expected equivalent semantic SE of the three users while ensuring that the system obeys the AP power constraint. We transform the rate of bit user R_B (in bit/s/Hz) into its semantic counterpart R_{SB} (in suts/s/Hz), as in [16], yielding

$$R_{SB} = R_B \frac{I}{\mu L}\epsilon_C, \tag{7}$$

where μ represents the average number of bits per word, the semantic similarity function for User B is denoted by ϵ_C and taken as one to reflect error-free transmission. This transformation allows a meaningful comparison between both types of users by expressing their efficiencies in the same unit. Notably, $\epsilon_C = 1$ reflects the requirement for perfect bit reconstruction in conventional systems, where even a single bit error can render a message unusable. In contrast, for semantic users, $\epsilon_{th} = 0.9$ is imposed to ensure sufficiently accurate

© 2023-2027 SCION Consortium. All rights reserved.

reconstruction with the DeepSC [3], [16].

For the hybrid NOMA structure, in each time slot, up to two users are scheduled to be active and share the resources. Users' activity is governed by the indicator variables I_{B}, I_{S1}, and I_{S2} for users B, S_1, and S_2, respectively. The selection of users for pairing is done dynamically based on real-time channel conditions, ensuring that the strongest user pairs with the weakest one across different time slots without a fixed ordering. However, when a semantic user has stronger channel gain than the bit user, the slot is allocated exclusively to the bit user, effectively operating as OMA. In contrast, when the channel conditions allow simultaneous transmission, NOMA is employed, where both users share the same time-frequency resources while following a bit-to-semantic decoding order. This adaptive hybrid NOMA strategy dynamically transitions between OMA and NOMA while maintaining the bit-to-semantic decoding strategy.

As semantic and bit users can simultaneously be active in each slot, a new decoding mechanism will be required. Let us denote the corresponding transmit signals for users B, S_1, and S_2 by x_{B}, x_{S1}, and x_{S2}, respectively. In slots where any two users share the channel via NOMA, the received signal will be represented as a superposition of their transmitted signals, scaled by their respective channel gains and corrupted by AWGN. In such slots, decoded signal of one user will always be affected by interference from the other. For the OMA slots, there will be no decoding interference experienced. Moreover, traditional channel-condition based SIC will no longer be beneficial in joint bit and semantic users communication networks as the bit user cannot decode a semantic signal due to the absence of prior training at transmitter and receiver. Consequently, we adopt the bit-to-semantic decoding order in slots involving both types of users, where semantic signals are decoded interference-free after SIC. However, the bit user being decoded first will treat the semantic user's signal as interference.

The pairing approach and decoding strategy for users are outlined in Algorithm 1. As the algorithm pairs the strongest gain user with the weakest, the transmit powers assigned to any pair are constrained such that the power given to the weak user plus that for the strong user must remain within $P_{\max}$ for any slot. The activity indicator variable of non-selected user forces its power to zero in that particular slot. The pairing strategy therefore directly impacts which users are active in each slot and hence determines the power allocation variables for each possible case.

Algorithm 1: User Pairing and Decoding Strategy for the Hybrid NOMA Framework.

1. **Initialization:** Set $P_{\max}$ value, and $I_{\mathrm{B}} = 0$, $I_{\mathrm{S1}} = 0$, and $I_{\mathrm{S2}} = 0$.
2. **Users Pairing in Each Slot:** Determine h_{B}, h_{S1}, and h_{S2} and pair the strongest user with the weakest one. Set the indicators for the paired users to 1.
3. **if** slot contains B and either S_1 or S_2 **then**
4. **if** involved semantic user has a stronger channel power gain than User B **then**
5. Reset its indicators to o which allocates the slot to User B only (OMA).
6. **else**
7. follow **Bit-to-Semantic Decoding:**
8. **B-receiver** decodes its own signal treating semantic user's signal as noise.
9. **S-receiver** firstly decodes B user's signal and applies SIC to subtract it,

© 2023-2027 SCION Consortium. All rights reserved.

resulting in its own signal without interference.

10. **end if**

11. **else if** slot contains S_1 and S_2 **then**

12. **Stronger Semantic User** decodes its own signal treating the weaker user's signal as noise.

13. **Weaker Semantic User** firstly decodes the stronger user's signal and applies SIC to subtract it, resulting in its own signal without interference.

14. **end if**

Figure 8 illustrates the overall transmission and decoding strategy of the proposed hybrid NOMA heterogeneous users setup.

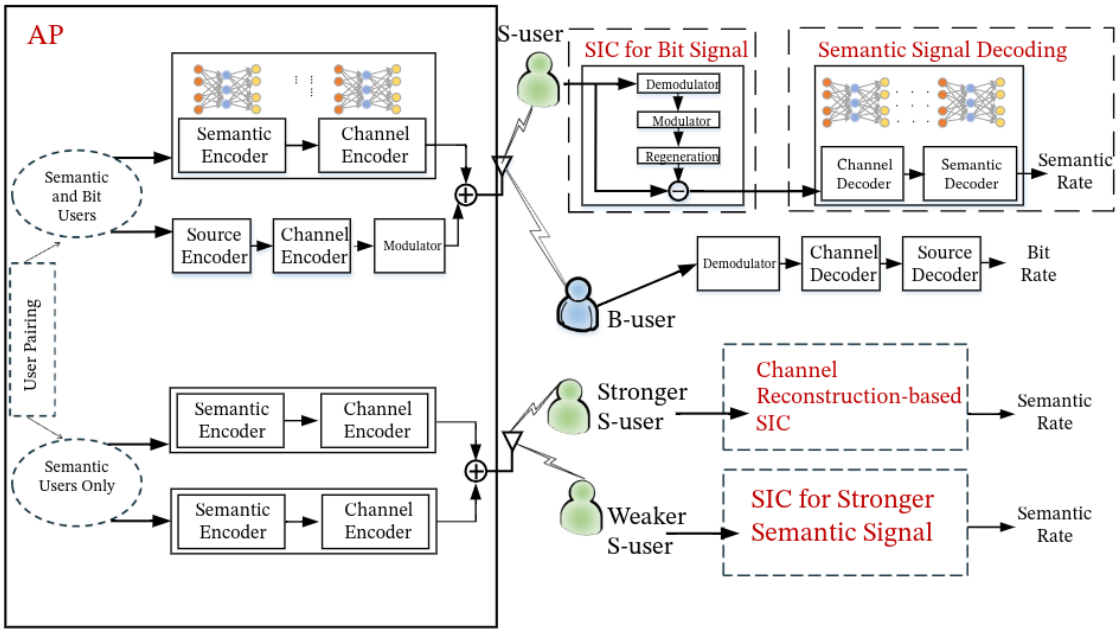


Figure 8: Downlink hybrid NOMA transmission and decoding strategy.

This framework is not limited to the illustrative three-user case and is scalable to larger networks and diverse application situations. In IoT deployments, low-power devices are able to carry task-specific information in addition to conventional bit streams due to low-overhead semantic encoders, enabling cost-effective coexistence of heterogeneous users. Hybrid NOMA can aid in satellite networks in balancing control and telemetry with semantically compressed data. In addition, semantic relays and coordinator satellites can improve shared KB consistency for non-terrestrial networks.

2.5.2 Optimal Power Allocation and Users Detection

After pairing users and establishing the decoding order in each time slot, we now formulate the power allocation optimization problem.

© 2023-2027 SCION Consortium. All rights reserved.

2.5.2.1 Power Allocation Optimization

Objective of our optimization problem is to maximize the expected value of the equivalent semantic SE, which can be formulated as

$$\begin{aligned} \max_{\{P_{\mathrm{B}},P_{\mathrm{S1}},P_{\mathrm{S2}}\}} \ & \mathbb{E}[I_{\mathrm{S1}}R_{\mathrm{S1}} + I_{\mathrm{S2}}R_{\mathrm{S2}} + I_{\mathrm{B}}R_{\mathrm{SB}}] \\ \text{s.t.} \quad & P_{\mathrm{B}} + P_{\mathrm{S1}} + P_{\mathrm{S2}} \le P_{\max}, \\ & P_{\mathrm{B}}, P_{\mathrm{S1}}, P_{\mathrm{S2}} \ge 0, \end{aligned} \tag{8}$$

where $\mathbb{E}[.]$ represents the expectation operator. Each of the considered rates in (8) can be generalized with a closed-form approximation of (4), exhibiting a logistic distribution model due to semantic similarity function [14], where $A_{K,1} = 0.37$, $A_{K,2} = 0.98$, $C_{K,1} = 0.25$, and $C_{K,2} = -0.7895$. The resulting rate curve (in suts/s/Hz) is resilient to moderate parameter variations due to the robustness of the neural network-based DeepSC.

Unlike conventional rate functions, the approximated semantic similarity function does not exhibit concavity with respect to γ which renders the optimization problem as non-convex. However, the problem satisfies the "time-sharing" condition following the proof in [54] and [63]. If any problem satisfies the "time-sharing" condition [63], then strong duality holds between the primal and Lagrangian dual problems, regardless of convexity of the original problem. Moreover, when strong duality exists, the Karush-Kuhn-Tucker (KKT) conditions are both sufficient and necessary for optimality [64].

2.5.2.2 Solution Method

We solve the power allocation problem in (5) by developing its Lagrangian and then using the KKT conditions in each time slot. The complementary slackness and primal feasibility conditions ensure that the solution respects the power constraint and precludes any negative power allocation. The KKT conditions are developed based on the Lagrangian formulation given by

$$\mathcal{L}(P_{\mathrm{B}}, P_{\mathrm{S1}}, P_{\mathrm{S2}}, \lambda) = \mathbb{E}[I_{\mathrm{S1}}R_{\mathrm{S1}} + I_{\mathrm{S2}}R_{\mathrm{S2}} + I_{\mathrm{B}}R_{\mathrm{SB}}] + \lambda(P_{\max} - P_{\mathrm{B}} - P_{\mathrm{S1}} - P_{\mathrm{S2}}), \tag{9}$$

where λ is the corresponding Lagrange multiplier for the power constraint.

2.5.2.3 Decoding Strategy in Slots with Semantic and Bit Users

Bit-to-semantic decoding order is followed where the rate of User B will always be adversely impacted by interference from the semantic user. The roots of (10) provide the solution for optimal power of the bit user.

$$\frac{\partial\left(\mathbb{E}\left[I_{\mathrm{B}}R_{\mathrm{SB}} + I_{Si}R_{Si}^{\mathrm{SB}}\right]\right)}{\partial P_{\mathrm{B}}} = 0, i \in \{1,2\} \tag{10}$$

Specifically, the corresponding rate for bit user in such time slots is defined by

$$R_{\mathrm{B}i} = \log_2\left(1 + \frac{P_{\mathrm{B}}|h_{\mathrm{B}}|^2}{\sigma^2 + P_{Si}|h_{\mathrm{B}}|^2}\right), i \in \{1,2\} \tag{11}$$

where P_{Si} denotes the power of semantic user paired with User B. Although we consider the equivalent semantic rate of bit user in our objective function, R_{SB} is merely a scalar multiple of conventional rate $R_{\mathrm{B}i}$ (in bits/s/Hz). User B first decodes its own signal, while the semantic

© 2023-2027 SCION Consortium. All rights reserved.

user extracts its desired signal without interference through SIC. The corresponding rate for semantic user involved in such slots is given by (12).

$$R_{\mathrm{S}i}^{\mathrm{SB}} = \frac{I}{KL}\epsilon_K\left(\frac{P_{\mathrm{S}i}|h_{\mathrm{S}i}|^2}{\sigma^2}\right), i \in \{1,2\} \tag{12}$$

Say, we solve for P_{B} when S_1 is active by putting $I_{\mathrm{B}} = 1$, $I_{\mathrm{S}1} = 1$, and $I_{\mathrm{S}2} = 0$. In the considered time slot with B and S_1 active, the closed-form solution for optimum P_{B} can be found from

$$\frac{I|h_{\mathrm{B}}|^2}{\mu\,\mathrm{Lln}(2)(\sigma^2+|h_{\mathrm{B}}|^2P_{\max}-|h_{\mathrm{B}}|^2P_{\mathrm{B}})} - \frac{I|h_{\mathrm{S}1}|^2(A_{K,2}-A_{K,1})C_{K,1}e^{\left(-C_{K,1}\frac{(P_{\max}-P_{\mathrm{B}})|h_{\mathrm{S}1}|^2}{\sigma^2}+C_{K,2}\right)}}{KL\sigma^2\left(1+e^{\left(-C_{K,1}\frac{(P_{\max}-P_{\mathrm{B}})|h_{\mathrm{S}1}|^2}{\sigma^2}+C_{K,2}\right)}\right)^2} = 0 \tag{13}$$

Similarly, taking the partial derivative with respect to $P_{\mathrm{S}1}$ in such slots and solving for P_{B} leads to a function with transcendental value [65] that is solved numerically for the optimal power P_{B}^* of bit user.

$$f(P_{\mathrm{B}}) = \frac{|h_{\mathrm{B}}|^2}{\mu\ln(2)(\sigma^2+|h_{\mathrm{B}}|^2P_{max}-|h_{\mathrm{B}}|^2P_{\mathrm{B}})} - \frac{|h_{\mathrm{S}1}|^2(A_{K,2}-A_{K,1})C_{K,1}e^{\left(-C_{K,1}\frac{(P_{max}-P_{\mathrm{B}})|h_{\mathrm{S}1}|^2}{\sigma^2}+C_{K,2}\right)}}{K\sigma^2\left(1+e^{\left(-C_{K,1}\frac{(P_{max}-P_{\mathrm{B}})|h_{\mathrm{S}1}|^2}{\sigma^2}+C_{K,2}\right)}\right)^2} \tag{14}$$

In such time slots, if the involved semantic user does not satisfy the set minimum threshold ϵ_{th}, the entire available power is allocated to the bit user. The optimal power allocation for the bit is thus given as

$$P_{\mathrm{B}}^* = \begin{cases} f(P_{\mathrm{B}}), & \text{if } \epsilon_K(\gamma) \geq \epsilon_{\mathrm{th}}, \\ P_{\max}, & \text{otherwise}. \end{cases} \tag{15}$$

2.5.2.4 Decoding Strategy in Slots with Semantic Users Only

In these slots, the Lagrangian is accordingly adjusted as

$$\mathcal{L}'(P_{\mathrm{S}1}, P_{\mathrm{S}2}, \lambda) = \mathbb{E}[I_{S1}R_{\mathrm{S}1} + I_{\mathrm{S}2}R_{\mathrm{S}2}] + \lambda(P_{\max} - P_{\mathrm{S}1} - P_{\mathrm{S}2}), \tag{16}$$

focusing solely on the rates of semantic users under the power constraint. Performing SIC becomes challenging in such slots due to the joint training of semantic encoders and decoders. To address this issue, we exploit an innovative channel reconstruction-based SIC method proposed in [17]. This method allows the semantic user with better channel condition to be decoded first while treating the signal from the weaker user as noise. The stronger user's signal is reconstructed by reversing the channel effects and mitigating interference until a target semantic similarity ϵ_{th} is attained. Once the stronger signal is successfully reconstructed, the weaker semantic user can subtract the decoded stronger user's signal from the superimposed signal at the receiver. This approach remains effective even when the semantic users have similar channel conditions and rate requirements, allowing either user to be decoded first.

Putting $I_{\mathrm{S}1} = 1$, $I_{\mathrm{S}2} = 1$, and assuming S_2 experiences higher channel gain than S_1, then the roots of (17) provide the optimal solution for $P_{\mathrm{S}2}$.

© 2023-2027 SCION Consortium. All rights reserved.

$$\frac{\partial(\mathbb{E}[R_{S2}+R_{S1}])}{\partial P_{S2}} = 0 \tag{17}$$

More generally, the achievable semantic rate of the first decoded user in such time slots is given by

$$R_{S2} = \frac{I}{KL}\epsilon_K\left(\frac{P_{S2}|h_{S2}|^2}{P_{S1}|h_{S2}|^2+\sigma^2}\right), \tag{18}$$

Expressing $P_{S1} = P_{\max} - P_{S2}$, the closed-form solution for the optimal power of stronger semantic user can be found from

$$\frac{I}{KL}\left[\frac{\gamma_2^2(\sigma^2+P_{\max}|h_{S1}|^2)(A_{K,2}-A_{K,1})C_{K,1}e^{-(C_{K,1}\gamma_2+C_{K,2})}}{P_{S2}^2|h_{S2}|^2\left(1+e^{-(C_{K,1}\gamma_2+C_{K,2})}\right)^2} - \frac{\gamma_1(A_{K,2}-A_{K,1})C_{K,1}e^{-(C_{K,1}\gamma_1+C_{K,2})}}{(P_{\max}-P_{S2})\left(1+e^{-(C_{K,1}\gamma_1+C_{K,2})}\right)^2}\right] = 0 \tag{19}$$

Where $\gamma_1 = \frac{(P_{\max}-P_{S2})|h_{S1}|^2}{\sigma^2}$, and $\gamma_2 = \frac{P_{S2}|h_{S2}|^2}{(P_{\max}-P_{S2})|h_{S2}|^2+\sigma^2}$ denote the received SNR at S_1 and S_2, respectively. Solving for P_{S2} after taking the partial derivative with respect to P_{S1} results in a transcendental value function for the optimal power of stronger semantic user.

2.5.2.5 Simulation Results

For the simulation figures of the proposed heterogeneous users hybrid NOMA communication setup, we use the parametric values of $\epsilon_{\text{th}} = 0.9$, $K = 5$, and $I/L = 1$. Users are randomly distributed for each Monte Carlo simulation within a 100 meters cell radius. Users channel coefficients for the independently distributed Rayleigh fading are determined based on the path-loss model $\rho = \rho_0\left(\frac{1}{d}\right)^{\beta}$, where $\rho_0 = -30$ dB and $\beta = 4$.

In Figure 4, we present the average activity of each user versus $P_{\max}$ in the heterogeneous users hybrid NOMA setup. Specifically, the figure shows the ergodic average of the indicators for each S1, S2 and B users when accessing the channel. It is evident that I_B remains nearly constant at a relatively high value compared to I_{S1} and I_{S2}, indicating that the bit-based user continuously enjoys higher resource allocation. More specifically, this trend reflects a clear prioritization of bit-based communication in the proposed system. The increase in available power does not translate into a significant change in the ergodic participation of semantic and bit users.

© 2023-2027 SCION Consortium. All rights reserved.

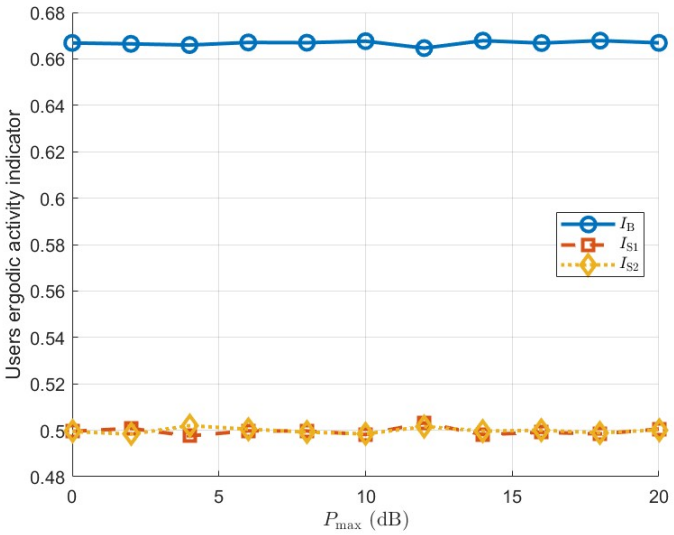


Figure 9: Ergodic activity indicator for each user versus $P_{\max}$.

We now compare the performance of the heterogeneous users hybrid NOMA communication with a heterogeneous users OMA setup, where the OMA time fractions are proportionally allocated to each user according to their participation ratio under hybrid NOMA. Figure 4 shows the equivalent ergodic semantic SE versus $P_{\max}$ for both setups. The results demonstrate that the proposed setup consistently outperforms its OMA counterpart. At $P_{\max} = 20$ dB, the hybrid NOMA attains roughly a four times higher equivalent ergodic semantic SE compared to the OMA setup. This indicates the effectiveness of hybrid NOMA for integrating both semantic and traditional bit-based communications, leveraging NOMA and OMA resource allocation strategies.

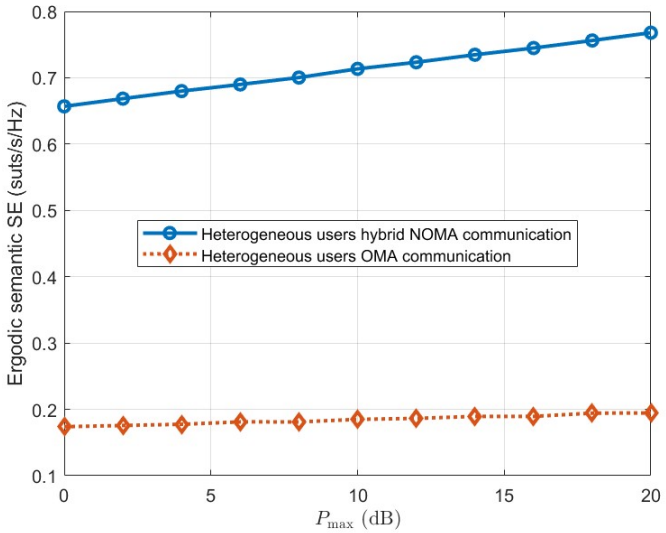


Figure 10: Ergodic semantic SE versus $P_{\max}$ with $\epsilon_{\text{th}} = 0.9$.

In Figure 4, the equivalent ergodic semantic SE is plotted versus ϵ_{th} for $P_{\max}$ values of 5 dB, 10 dB, and 20 dB. It is evident that the proposed setup achieves higher ergodic semantic SE across all values of ϵth for different $P_{\max}$. For both the setups as the value of ϵ_{th} approaches one, a drop in SE is recorded due to the inherent limitations in accurate semantic reconstruction at higher semantic similarity thresholds. However, the hybrid NOMA still maintains a higher SE due to the higher resource utilization by the bit user aiding in accurate

© 2023-2027 SCION Consortium. All rights reserved.

signal reconstruction. Across the entire considered ϵ_{th} sweep, the hybrid NOMA consistently offers roughly a three-to-four times higher equivalent ergodic semantic SE for different $P_{\max}$ values.

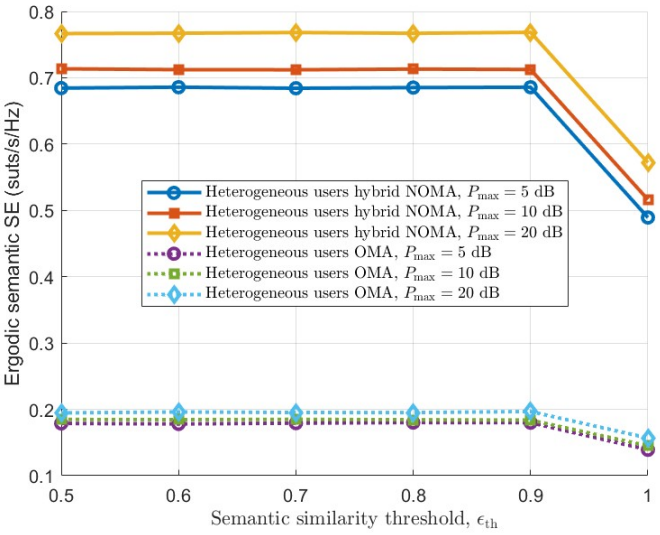


Figure 11: Ergodic semantic SE versus ϵ_{th} for different $P_{\max}$.

Figure shows the bit user ergodic SE versus $P_{\max}$ for hybrid NOMA and OMA setups. As $P_{\max}$ increases, the hybrid NOMA curve consistently lies above the OMA curve, demonstrating the benefits of its flexible resource allocation strategy. If the semantic user has stronger channel gain than the bit user in shared time slots, the entire power is redirected to the bit user, otherwise the total power splits between both the users. More specifically, at $P_{\max} = 20$ dB, the hybrid NOMA setup lifts the ergodic SE to about 23 bits/s/Hz, whereas its OMA counterpart barely exceeds 6 bits/s/Hz.

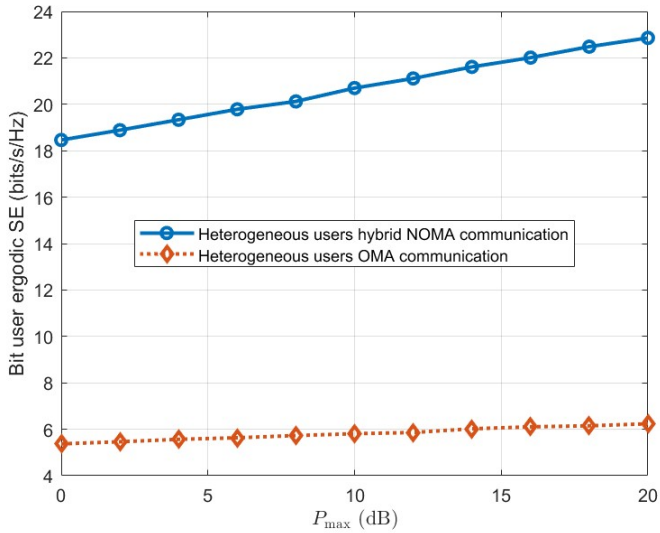


Figure 12: Bit user ergodic SE versus $P_{\max}$ with $\epsilon_{\text{th}} = 0.9$.

In Figure , the equivalent ergodic semantic SE of bit user in the heterogeneous users hybrid NOMA setup is shown versus $P_{\max}$ for ϵ_{th} values of 0.85, 0.9, 0.95. The three curves corresponding to different values of ϵ_{th} show a steadily growing trend with $P_{\max}$. The figure shows that the equivalent semantic SE of bit user consistently increases with $P_{\max}$. This behaviour can be attributed to the efficient resource management of the hybrid NOMA

© 2023-2027 SCION Consortium. All rights reserved.

system which prioritizes bit user in a heterogeneous users network for accurate signal reconstruction.

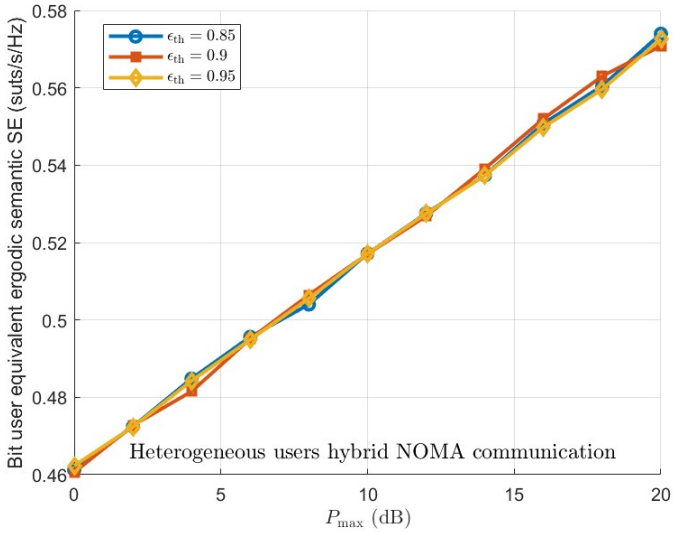


Figure 13: Bit user equivalent ergodic semantic SE versus P_{max} for different ϵ_{th}.

© 2023-2027 SCION Consortium. All rights reserved.

3 Challenges and Future Research Perspectives in SemCom

With the increasing number of power-intensive applications, it becomes crucial to evaluate efficient solutions for next-generation wireless networks. SemCom represents a transformative shift in 6G network design, addressing critical challenges in SE, heterogeneous service integration and resource-constrained environments. Despite the recent research advancements, SemCom still faces several significant challenges that need further explorations are:

i) Developing unified metrics for evaluating the performance of SemCom and its accuracy remain critical, complicating comparative analyses across wireless systems

ii) Communication bandwidth overhead remains a significant concern, as transmitting multi-modal data typically demands substantial bandwidth, potentially causing increased communication overhead and latency issues.

iii) The performance of semantic encoders is often highly dependent on the training data. Models trained on specific domains may fail in others Effectively achieving semantic fusion among heterogeneous data is complex due to the diverse information inherent to different modalities. Developing efficient mechanisms to extract and integrate semantic correlations across diverse and integrated networks (e.g., satellite, terrestrial, maritime) involve overcoming significant interoperability issues.

iv) Understanding and interpreting the decision-making processes of SemCom models capable of multi-tasking become increasingly challenging yet essential.

v) Achieving performance trade-offs and fairness pose a substantial issue to optimally manage communication and computational resources while maintaining semantic accuracy and QoS in a heterogeneous users setup.

vi) While semantic transmission reduces data size, it also introduces unique security and privacy vulnerabilities, addressing which will be of paramount importance against any potential semantic-level eavesdropping.

vii) Extending SemCom to multi-hop or relay-aided networks remains underexplored. Semantic representation, aggregation, and transformation in cooperative networks pose new complexity and delay trade-offs.

To maximize the potential of SemCom, future research should explore cross-layer optimization strategies that link the physical, medium access control, and application layers. Integrating semantic awareness into the physical layer through AI-native air interface designs could unlock new efficiencies in resource utilization in satellite communications. Additionally, such integration could enable protocol decisions that are responsive to semantic content and communication objectives under the resource-constrained devices, balancing computational overhead with performance gains.

One promising direction is the deployment of pinching antennas system with SemCom for performance exploration. Traditionally, network throughput has been limited by the inherent properties of fixed wireless channels imposing substantial challenges due to intense path loss and frequent Line-of-Sight (LoS) blockages, especially at high-frequency bands like millimeter-wave and terahertz. The pinching antennas system makes it possible to position

© 2023-2027 SCION Consortium. All rights reserved.

antennas anywhere over the arbitrarily long dielectric waveguide to create strong LoS links for each user based on their specific signal conditions. The approach is also cost-effective and simple to install, since it only involves the physical addition or removal of passive dielectric parts. Optimization problems formulating joint semantic performance metrics and antenna activation strategies through dynamic positioning along dielectric waveguides could lead to innovative theoretical frameworks and algorithms for next-generation networks.

As next-generation communication systems evolve, SemCom has the potential to become a cornerstone of intelligent, context-aware networks, reshaping how information is transmitted and processed in the wireless ecosystems of the future.

© 2023-2027 SCION Consortium. All rights reserved.

4 Conclusions

This deliverable has presented a detailed exploration into the design and implementation of semantic encoders and decoders for 6G communication systems, with the goal of enhancing efficiency, adaptability, and intelligence in wireless communication networks. By addressing both the theoretical and practical aspects of SemCom, we have demonstrated its potential in contributing to the SE improvements, especially in scenarios involving heterogeneous users sharing the same time-frequency resource. With the incorporation of machine learning based methods for semantic features extraction and signal reconstruction, SemCom offers great potential for intelligent and adaptive networks. A structured classification of SemCom systems is also provided, emphasizing their relevance in both multi-modal TOC scenarios and satellite network applications. The benefits of integrating TOC into SemCom systems have been demonstrated, particularly in terms of efficiency and relevance to user tasks. A comprehensive TOC-based multi-modal framework has been proposed, serving as a foundation for future development. Furthermore, the potential of SemCom to enhance satellite system performance has been thoroughly analyzed. Most importantly, a practical and forward-looking satellite communication architecture leveraging SemCom techniques has been introduced, paving the way for next-generation intelligent and efficient satellite networks.

A novel framework for enabling heterogeneous semantic and bit users network is proposed with the hybrid NOMA mechanism. Through this synergy, the formulated power allocation optimization problem optimally maximizes the equivalent ergodic semantic SE and allows more efficient use of available spectrum resources for both semantic and bit-based users. The bit-to-semantic decoding order is adopted for the detection of signals within the slots having both types of users, while channel reconstruction-based SIC is implemented in semantic users-only time slots. Comprehensive simulations have validated the effectiveness of the proposed framework in terms of improved semantic SE and decoding reliability of the semantic signal under diverse channel conditions. More specifically, the equivalent ergodic semantic SE using hybrid NOMA outperforms the OMA setup by roughly four times for the heterogeneous semantic and bit users at $P_{\max}$ value of 20 dB. Therefore, the proposed user pairing and decoding strategy confirms the effectiveness of our power allocation optimization problem for the semantic-aware heterogeneous users networks.

Looking ahead, further research is encouraged in areas such as multi-modal semantic fusion, energy-efficient system design for SemCom, and its integration with state-of-the-art pinching antennas architectures. These directions will be pivotal in the deployment of SemCom into next-generation intelligent and real-world applications, ultimately supporting SCION's broader vision of secure, scalable, and massive machine-to-machine connectivity.

© 2023-2027 SCION Consortium. All rights reserved.

5 References

[1] Z. Lu, R. Li, K. Lu, X. Chen, E. Hossain, Z. Zhao, and H. Zhang, "Semantics-empowered communications: A tutorial-cum-survey," *IEEE Commun. Surv. Tuts.*, vol. 26, no. 1, pp. 41–79, 2024.

[2] P. Zhang et al., "Towards intelligent and efficient 6G networks: JCSC enabled on-purpose machine communications," *IEEE Wireless Commun*., vol. 30, no. 1, pp. 150–157, Feb. 2023.

[3] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning enabled semantic communication systems," *IEEE Trans. Signal Process*., vol. 69, pp. 2663–2675, 2021.

[4] R. Song and N. Li, "High speed terahertz communication in the space and terrestrial integrated next generation wireless communication systems," *13th UK-Europe-China Workshop on Millimetre-Waves and Terahertz technologies (UCMMT)*, Tianjin, China, pp. 376-378, Aug. 2020.

[5] F. Boabang, A. Ebrahimzadeh, R. H. Glitho, H. Elbiaze, M. Maier, and F. Belqasmi, "A machine learning framework for handling delayed/lost packets in Tactile Internet remote robotic surgery," *IEEE Trans. Netw. Service Manag*., vol. 18, no. 4, pp. 4829–4845, Dec. 2021.

[6] S. Ren, Z. Lei, Z. Wang, M. Dianati, Y. Wang, S. Chen, and W. Zhang, "Interruption-aware cooperative perception for V2X communication aided autonomous driving," *IEEE Trans. Intell. Veh.*, vol. 9, no. 4, pp. 4698–4714, Apr. 2024. 2024.

[7] J. Choi and S. Sun, "Semantic-aware V2X messaging for congestion-aware safety," *IEEE Trans. Veh. Technol.*, vol. 72, no. 1, pp. 215-229, 2023.

[8] C. E. Shannon and W. Weaver, "The Mathematical Theory of Communication," University of Illinois Press, 1949.

[9] X. Luo, H.-H. Chen, and Q. Guo, "Semantic communications: Overview, open issues, and future research directions," *IEEE Wireless Commun*., vol. 29, no. 1, pp. 210–219, 2022.

[10] D. Gündüz et al., "Beyond transmitting bits: Context, semantics, and task-oriented communications," *IEEE J. Sel. Areas Commun*., vol. 41, no. 1, pp. 5–41, Jan. 2023. [11] Y. Bar-Hillel and R. Carnap, "Semantic information," *Brit. J. Philosophy Sci*., vol. 4, no. 14, pp. 147–157, 1953.

[12] X. Mu, Y. Liu, L. Guo, and N. Al-Dhahir, "Heterogeneous semantic and bit communications: A Semi-NOMA scheme," *IEEE J. Sel. Areas Commun*., vol. 41, no. 1, pp. 155-169, 2023.

[13] R. Carnap and Y. Bar-Hillel. "An outline of a theory of semantic information." Oct. 1952.

[14] W. Yang, H. Du, Z. Q. Liew, W. Y. B. Lim, Z. Xiong, D. Niyato, X. Chi, X. Shen, and C. Miao, "Semantic communications for future internet: Fundamentals, applications, and challenges," *IEEE Commun. Surv. Tuts.*, vol. 25, no. 1, pp. 213–250, 2022.

[15] T. Han, Q. Yang, Z. Shi, S. He, and Z. Zhang, "Semantic-Preserved Communication System for Highly Efficient Speech Transmission," *IEEE J. Sel. Areas Commun*., vol. 41, no. 1, pp. 245–259, 2023.

[16] L. Yan, Z. Qin, R. Zhang, Y. Li, and G. Y. Li, "Resource allocation for text semantic communications," *IEEE Wireless Commun. Lett*., vol. 11, no. 7, pp. 1394–1398, 2022.

[17] B. Chen, X. Wang, D. Li, R. Jiang, and Y. Xu, "Uplink NOMA semantic communications: Semantic reconstruction for SIC," in *IEEE/CIC Int. Conf. Commun. China (ICCC)*, Dalian, China,

© 2023-2027 SCION Consortium. All rights reserved.

2023, pp. 1–6.

[18] Z. Weng, Z. Qin, X. Tao, C. Pan, G. Liu, and G. Y. Li, "Deep learning enabled semantic communications with speech recognition and synthesis," *IEEE Trans. Wireless Commun*., early access, Feb. 6, 2023.

[19] R. D. Yates, Y. Sun, D. R. Brown, S. K. Kaul, E. Modiano, and S. Ulukus, "Age of information: An introduction and survey," *IEEE J. Sel. Areas Commun*., vol. 39, no. 5, pp. 1183–1210, May 2021.

[20] M. Arazzi, A. Nocera, and E. Storti, "The SemIoE ontology: A semantic model solution for an IoE-based industry," *IEEE Internet Things J*., vol. 11, no. 24, pp. 40376–40387, Dec. 2024.

[21] S. Guo, Y. Wang, N. Zhang, Z. Su, T. H. Luan, Z. Tian, and X. Shen, "A survey on semantic communication networks: Architecture, security, and privacy," *arXiv:2405.01221*, pp. 1–31, 2024.

[22] Bao, Jie, Prithwish Basu, Mike Dean, Craig Partridge, Ananthram Swami, Will Leland, and James A. Hendler. "Towards a theory of semantic communication (extended technical report)." In *Performing Organization: US Army Research Lab, Adelphy, Presented at the 2011 IEEE First International Workshop on Network Science*. 2011.

[23] E. Bourtsoulatze, D. B. Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," *IEEE Trans. Cogn. Commun. Netw*., vol. 5, no. 3, pp. 567-579, 2019.

[24] E. C. Strinati and S. Barbarossa, "6G networks: Beyond shannon towards semantic and goal-oriented communications," *Comput. Netw*., vol. 190, p. 107930, 2021.

[25] H. Xie and Z. Qin, "A lite distributed semantic communication system for internet of things," *IEEE J. Sel. Areas Commun*., vol. 39, no. 1, pp. 142-153, 2020.

[26] G. Huang, D. Chen, T. Li, F. Wu, L. Van Der Maaten, and K. Q. Weinberger, "Multi-scale dense networks for resource efficient image classification," *arXiv preprint arXiv:1703.09844*, 2017.

[27] J. Shao, Y. Mao, and J. Zhang, "Learning task-oriented communication for edge inference: An information bottleneck approach," *IEEE J. Sel. Areas Commun*., vol. 40, no. 1, pp. 197-211, 2021.

[28] S. Xie, S. Ma, M. Ding, Y. Shi, M. Tang, and Y. Wu, "Robust information bottleneck for task-oriented communication with digital modulation," *IEEE J. Sel. Areas Commun*., 2023.

[29] Y. Yang, Y. Wu, S. Ma, and Y. Shi, "Multi-task-oriented broadcast for edge ai inference via information bottleneck," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 2023, pp. 3142-3147.

[30] J. Shao, Y. Mao, and J. Zhang, "Task-oriented communication for multidevice cooperative edge inference," *IEEE Trans. Wirel. Commun*., vol. 22, no. 1, pp. 73-87, 2022.

[31] C. Liu, C. Guo, Y. Yang, and N. Jiang, "Adaptable semantic compression and resource allocation for task-oriented communications," *IEEE Trans. Cogn. Commun. Netw*., vol. 10, no. 3, pp. 769–782, Jun. 2024.

[32] L. Wang, W. Wu, F. Zhou, Z. Yang, Z. Qin, and Q. Wu, "Adaptive resource allocation for semantic communication networks," *IEEE Trans. Commun*., vol. 72, no. 11, pp. 6900–6916, 2024.

© 2023-2027 SCION Consortium. All rights reserved.

[33] Y. Han, G. Huang, S. Song, L. Yang, H. Wang, and Y. Wang, "Dynamic neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7436-7456, 2021.

[34] J. H. Huang, M. Y. Sun, S. J. Han et al., "Prospects of semantic communication technology for 6G satellite communication," *ZTE Technology Journal*, vol. 30, no. 5, pp. 3–8, 2024.

[35] S. H. Sheik, L. X. Nguyen, Y. K. Tun, Z. Han, and C. S. Hong, "Semantic enabled 6G LEO satellite communication for earth observation: A resource-constrained network optimization," in *IEEE Global Commun. Conf. (GLOBECOM) 2024, 8-12 Dec, Cape Town, South Africa*, 2024.

[36] B. Guo, Z. Xiong, B. Wang, T. Q. S. Quek, and Z. Han, "Semantic communication-aware end-to-end routing in large-scale LEO satellite networks," in 2024 *IEEE International Conference on Metaverse Computing, Networking, and Applications (MetaCom)*, Aug. 2024, pp. 137–142.

[37] W. Chen, S. Li, C. Ju, D. Wang, J. Li, and D. Wang, "Semantic-enhanced downlink LEO satellite communication system with OTFS modulation," *IEEE Commun. Lett.*, vol. 28, no. 6, pp. 1377–1381, Jun. 2024.

[38] P. Jiang, C.-K. Wen, X. Li, S. Jin, and G. Y. Li, "Semantic satellite communications based on generative foundation model," *arXiv preprint arXiv:2404.11941*, Apr. 2024.

[39] T. M. Getu, G. Kaddoum, and M. Bennis, "Semantic communication: A survey on research landscape, challenges, and future directions," *Proceedings of the IEEE*, vol. 112, no. 11, pp. 1649–1685, 2024.

[40] Y. Zhao, Y. Yue, S. Hou, B. Cheng, and Y. Huang, "LaMoSC: Large language model-driven semantic communication system for visual transmission," *IEEE Trans. Cogn. Commun. Netw.*, early access, May 16, 2024.

[41] L. X. Nguyen, A. D. Raha, P. S. Aung, D. Niyato, Z. Han, and C. S. Hong, "A contemporary survey on semantic communications: Theory of mind, generative AI, and deep joint source-channel coding," *arXiv preprint arXiv*:2502.16468, Feb. 2025.

[42] F. Zhou, Y. Li, X. Zhang, Q. Wu, X. Lei, and R. Q. Hu, "Cognitive semantic communication systems driven by knowledge graph," in *ICC 2022 - IEEE Int. Conf. Commun.*, 2022, pp. 4860–4865.

[43] P. Zhang, Y. Liu, Y. Song, and J. Zhang, "Advances and challenges in semantic communications: A systematic review," *Nat. Sci. Open*, vol. 3, no. 4, 2024, Art. no. 20230029.

[44] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," *arXiv:1908.10084*, Aug. 2019.

[45] J. Devlin, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, May 2019.

[46] W. Huang, J. Wang, X. Chen, Q. Peng, and Y. Zhu, "Flag vector assisted multi-user semantic communications for downlink text transmission," *IEEE Commun. Lett.*, 2024.

[47] Z. Weng, Z. Qin, X. Tao, C. Pan, G. Liu, and G. Y. Li, "Deep learning enabled semantic communications with speech recognition and synthesis," *IEEE Trans. Wirel. Commun.*, vol. 22, no. 9, pp. 6227-6240, 2023.

[48] C. Liang, D. Li, Z. Lin, and H. Cao, "Selection-based image generation for semantic

© 2023-2027 SCION Consortium. All rights reserved.

communication systems," *IEEE Commun. Lett.*, 2023.

[49] Z. Zhang, Q. Yang, S. He, and J. Chen, "Deep learning enabled semantic communication systems for video transmission," in *2023 IEEE 98th Vehicular Technology Conference (VTC2023-Fall)*. IEEE, 2023, pp. 1–5.

[50] H. Xie, Z. Qin, and G. Y. Li, "Task-oriented multi-user semantic communications for VQA," *IEEE Wireless Commun. Lett.*, vol. 11, no. 3, pp. 553–557, 2021.

[51] G. Zhang, Q. Hu, Z. Qin, Y. Cai, G. Yu, and X. Tao, "A unified multi-task semantic communication system for multimodal data," *IEEE Trans. Commun.*, 2024.

[52] Z. Zhu, R. Zhang, X. Cheng, and L. Yang, "Multi-modal fusion-based multi-task semantic communication system," arXiv preprint arXiv:2407.00964, 2024.

[53] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y.-J. A. Zhang, "The roadmap to 6G: AI empowered wireless networks," *IEEE Commun. Mag.*, vol. 57, no. 8, pp. 84–90, 2019.

[54] X. Mu and Y. Liu, "Exploiting semantic communication for non-orthogonal multiple access," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 8, pp. 2563–2576, 2023.

[55] J. Wang, C. Jiang, H. Zhang, Y. Ren, K.-C. Chen, and L. Hanzo, "Thirty years of machine learning: The road to Pareto-optimal wireless networks," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 1472–1514, 2020.

[56] X. Pei, Y. Chen, M. Wen, H. Yu, E. Panayirci, and H. V. Poor, "Next-generation multiple access based on NOMA with power level modulation," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 4, pp. 1072–1083, 2022.

[57] X. You, C.-X. Wang, J. Huang, X. Gao, Z. Zhang, M. Wang, Y. Huang, C. Zhang, Y. Jiang, J. Wang et al., "Towards 6G wireless communication networks: Vision, enabling technologies, and new paradigm shifts," *Sci. China Inf. Sci.*, vol. 64, pp. 1–74, 2021.

[58] X. Mu and Y. Liu, "Semantic communications in multi-user wireless networks," *arXiv preprint arXiv:2211.08932*, 2022.

[59] I. Ahmed, Y. Sun, J. Fu, A. K¨ose, L. Musavian, M. Xiao, and B. ¨Ozbek, "Semantic communications in 6G: Coexistence, multiple access, and satellite networks," *IEEE Commun. Stand. Mag.*, 2025.

[60] Z. Ding, R. Schober, P. Fan, and H. V. Poor, "Next generation multiple access for IMT towards 2030 and beyond," *Science China Information Sciences*, vol. 67, no. 6, p. 166301, 2024.

[61] Z. Ding, D. Xu, R. Schober, and H. V. Poor, "Hybrid NOMA offloading in multi-user MEC networks," *IEEE Trans. Wireless Commun.*, vol. 21, no. 7, pp. 5377–5391, 2022.

[62] S. R. Islam, M. Zeng, O. A. Dobre, and K.-S. Kwak, "Resource allocation for downlink NOMA systems: Key techniques and open issues," *IEEE Wireless Commun.*, vol. 25, no. 2, pp. 40–47, Apr. 2018.

[63] W. Yu and R. Lui, "Dual methods for nonconvex spectrum optimization of multicarrier systems," *IEEE Trans. Commun.*, vol. 54, no. 7, pp. 1310–1322, 2006.

[64] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge Univ. Press, 2004.

[65] S. Bernard, "Formalization of the Lindemann-Weierstrass theorem," in *Interactive Theorem Proving: 8th International Conference, ITP 2017, Brasilia, Brazil, September 26–29, 2017, Proceedings 8*. Springer, 2017, pp. 65–80.

© 2023-2027 SCION Consortium. All rights reserved.