Semantic Communications: Overview, Open Issues, and Future Research Directions

Xuewen Luo, Hsiao-Hwa Chen, and Qing Guo

Abstract

With the deployment of the fifth generation (5G) in many countries, people start to think about what the next-generation of wireless communications will be. The current communication technologies are already approaching the Shannon physical capacity limit with advanced encoding (decoding) and modulation techniques. On the other hand, artificial intelligence (AI) plays an increasingly important role in the evolution from traditional communication technologies to the future. Semantic communication is one of the emerging communication paradigms, which works based on its innovative "semantic-meaning passing" concept. The core of semantic communication is to extract the "meanings" of sent information at a transmitter, and with the help of a matched knowledge base (KB) between a transmitter and a receiver, the semantic information can be "interpreted" successfully at a receiver. Therefore, semantic communication essentially is a communication scheme based largely on AI. In this article, an overview of the latest deep learning (DL) and end-to-end (E2E) communication based semantic communications will be given and open issues that need to be tackled will be discussed explicitly.

Introduction

From the first generation (1G) to the fifth generation (5G), the goals of communication systems have changed drastically from analog audio signal transmission to high speed and low-latency multimedia services. Especially in 5G, various advanced wireless communication technologies have been used, such as non-orthogonal multiple access (NOMA), massive multiple-input multiple-output (MIMO), millimeter wave communications, and so on. Despite the fact that 5G can meet most requirements of different services with a low-latency and a high data rate, the existing technologies may not be able to support many intelligent applications in beyond-5G (B5G) communications. The services in B5G networks, such as connected living, brain-to-computer interaction, virtual reality (VR), augmented reality (AR), and mixed-reality (MR), will be supported to enrich our future intelligent life. The technical requirements for these services are much higher than 5G networks, such as $1 \sim 10$ Gb/s/m^3 traffic density, 1 Tb/s uplink and downlink data rate, 0.1 ms latency, and so on [1]. In this context, 5G serves as only a transitional platform from traditional communications to futuristic artificial intelligence (AI) communications.

On the other hand, the existing communication technologies have nearly approached the Shannon physical-layer capacity limit. The goals and services provided by B5G prompt researchers to think about what the next-generation of wireless communications will be. A pioneering work done by Weaver [1, 2] revealed that communications can be categorized into three levels, as shown in Fig. 1. The lowest level is the technical level, which is defined by Shannon's classical information theory and focuses on how to transmit symbols (bits) accurately and effectively from a transmitter to a receiver. On the middle level, that is, the semantic level, semantic information of the data is extracted and transmitted via a semantic channel, whereas the upper level, that is, the effectiveness level, is responsible for providing the needed communication efficiency on the lower two levels.

Benefitting from the advancements in microelectronics and AI technologies, deep learning (DL) and end-to-end (E2E) communication technologies emerged recently to play an important role in the transformation of traditional communication technologies to the future. Semantic communication was proposed as an intelligent communication scheme, which concerns the meaning of transmitted messages rather than accurate bit stream transmission [3]. For instance, in a natural language system, if a source sends a message that "Bob's automobile was parked there," its destination may receives "Bob's car was parked there" in a semantic communication system, but "Bab's autmkobile was pbrked there" in a traditional communication system. In this example, semantic communications concern the meaning behind the transmitted symbols (bits). Even though the word phrases interpreted at the receiver have been changed a little bit, the receiver can still understand it. However, in a traditional communication system, the received message is confusing because the transmitted symbols (bits) have been distorted due to channel noise and interference. Therefore, although syntactic mis-matches may exist in semantic communication systems, there are no semantic errors. Moreover, it also suggests that when bandwidth is limited or signal-to-noise ratio (SNR) is relatively low, a semantic communication system may still perform well and likely consume less energy.

It should be noted that semantic communication is not a security communication scheme, but an intelligent way to exchange information. The biggest difference between semantic communication and encryption-based security communication

Digital Object Identifier:
10.1109/MWC.101.2100269

Xuewen Luo and Qing Guo are with Harbin Institute of Technology; Hsiao-Hwa Chen is with the National Cheng Kung University.

1536-1284/22/$25.00 © 2022 IEEE

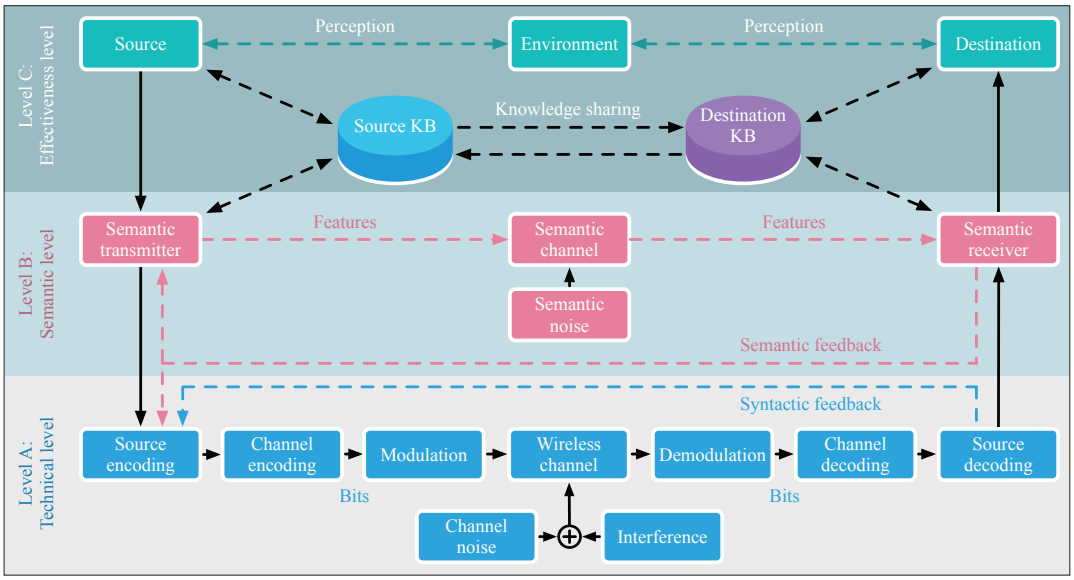

FIGURE 1. Three-layer model in semantic communications.

lies in their goals. Semantic communication works on the "semantic meaning passing" concept, but the goal of encryption-based security communication is to prevent unauthorized decoding at an attacker or an eavesdropper.

Normally, a semantic communication system should include all three levels in Fig. 1. As a matter of fact, semantic communication is not a completely new idea, which can be traced back to the seminal works done by Weaver in the 1940s [2]. Then in 1952, Carnap and Bar-Hillel introduced a semantic information theory (SIT) based on logical probability functions derived from the contents of a sentence [4]. Inspired by those pioneering works, Bao *et al*. [5] reviewed the existing works on quantifying semantic information theory, and then proposed a model-theoretical approach for semantic data compression and reliable semantic communications. Recently, enabled by DL technologies, various semantic communication systems were designed for the transmission of text [3, 6–8], image [9, 10], and speech signals [11]. The objective of this article is to provide an overview of the most recent works on DL and E2E communication based semantic communications.

Semantic communication is an interdisciplinary research topic, which involves linguistics, computer science, and wireless communications. We would like to give detailed explanations for several key terms used in the linguistics to help readers understand the contents:

- Syntax: A set of rules, principles, and processes (e.g., word order) that govern the structure of sentences in a given language.
- Polysemy: An individual word or phrase that can be used (in different contexts) to express two or more different meanings.
- Synonym: A word or phrase that gives exactly or nearly the same meaning as another word or phrase in the same language.
- Dialect: Refers to a variant of a language shared by a particular group of the speakers of the language.

The rest of this article can be outlined as follows. The next section focuses on a comparison between semantic communications and traditional communications. An overview of semantic communication systems will be given, followed by use cases and open issues on semantic communications. Finally, the conclusions will be given at the end of this article.

Differences Between Semantic and Traditional Communications

Traditional communication systems aim to offer a high data transmission rate and a low symbol (bit) error rate. However, the basic idea of semantic communications is to extract the "meanings" or "features" of sent information from a source, and "interpret" the semantic information at a destination. In this section, the similarities and differences between traditional and semantic communications will be discussed.

Semantic Source and Destination

In a traditional communication system, the entities at the source and destination are only electronic equipments, which work in a workflow of different communication blocks, such as source encoding (decoding), channel encoding (decoding), and so on. As shown in Fig. 2a, data in a traditional communication system are compressed by a source encoder and redundancy is added in the channel encoder to improve its robustness against interference/noise in the channels. At a destination, a reverse process proceeds to recover the originally sent data. In such a block-based structure, no intelligence is involved in signal transmission and reception, and the implicit meanings behind the messages are completely ignored at the transmitter and the receiver.

On the other hand, a semantic communication system is a complicated system. The Semantic source and destination are agents that need to perform not only the functions of traditional communication terminals, but also various highly intelligent algorithms. The agents in a semantic communication system can be humans, machines, or other devices with intelligence. Moreover, the semantic source and destination can perceive the environment and operate autonomously [1]. A semantic

Traditional communication systems aim to offer a high data transmission rate and a low symbol (bit) error rate. However, the basic idea of semantic communications is to extract the "meanings" or "features" of sent information from a source, and "interpret" the semantic information at a destination.

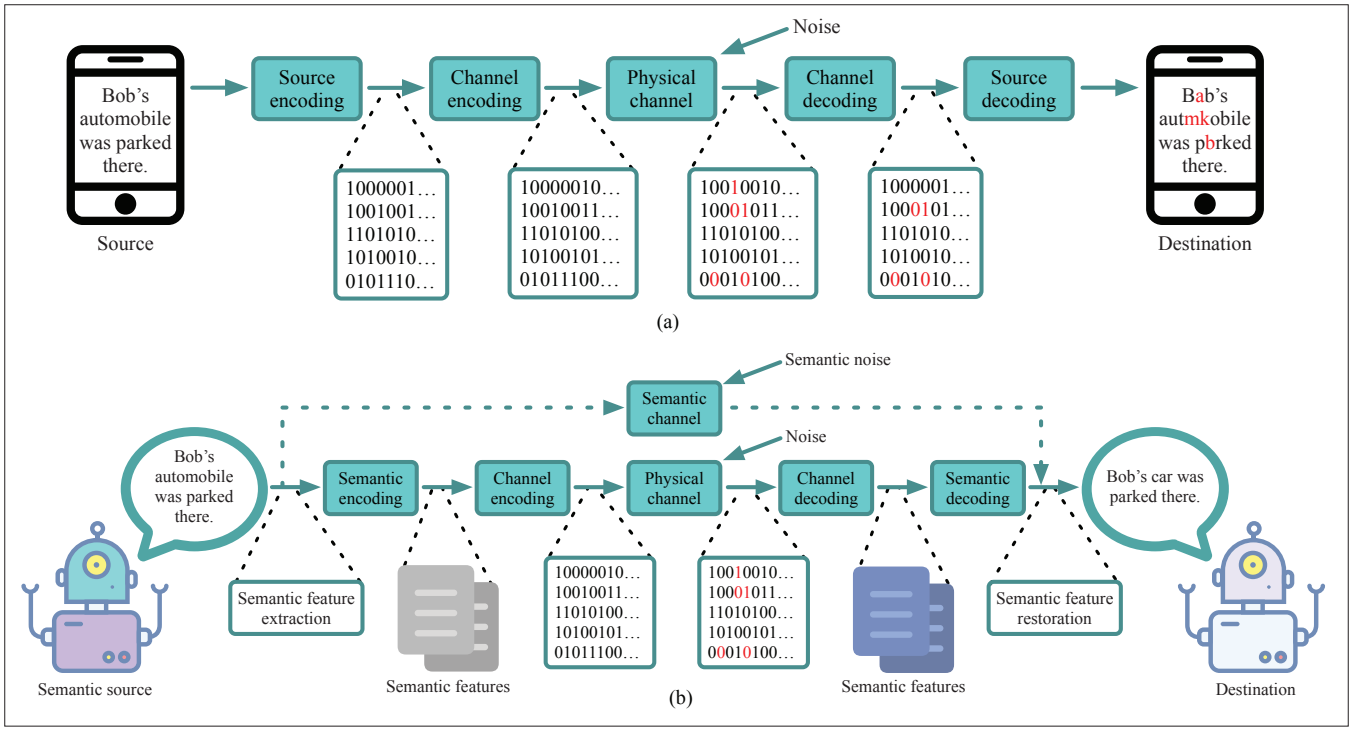

FIGURE 2. A comparison between traditional and semantic communication systems: a) Workflow of a traditional communication system; b) Workflow of a semantic communication system.

source is able to extract the semantic features of raw messages and encode these features into symbols (bits) for transmission. The destination should be able to "understand" and infer the messages sent by the semantic source. For example, in a natural language system, there is a syntax for the semantic source and destination to understand the meanings of words in sentences. Moreover, the agents should distinguish even very subtle differences in words, which behave just like a human reading polysemy and synonym. An example of polysemy is the word "cherry," which becomes a person's name when "C" is an uppercase letter; otherwise it means a type of fruit. Car and automobile are synonyms despite the fact that they are different in syntax, and they are the same in a semantic sense.

Semantic Channels with Errors

As shown in Figs. 2a and 2b, in addition to a physical channel, a "virtual" channel, that is, a semantic channel exists, through which the semantic information is transmitted from a source and interpreted at a destination. Unlike traditional communications, there are two different types of noises in semantic communication systems. The first type is physical channel noise, which exists ubiquitously in wireless communications and is caused by physical channel impairments, such as additive white Gaussian noise (AWGN), channel fading, multiple path propagation, and so on. It is noted that the errors caused by channel propagation usually occur before channel decoding and can be corrected by channel decoding. In addition, the co-channel interferences from different users cannot be ignored. The second type of noise is semantic noise, which appears in message interpretation processes due to the ambiguity existing in words, sentences or symbols used in the sent messages [3, 7].

Semantic noise can cause semantic errors at a receiver and induce misunderstanding of the received messages. On the semantic level, semantic errors can be caused by the mismatch between the background knowledge bases (KBs) used by the semantic source and destination. For instance, the source is an English language system, but in the destination only the Chinese language is used. In addition, polysemy and synonym may also induce semantic errors. On the technical level (as shown in Fig. 1), semantic errors may occur from symbol or bit errors during transmission due to the noise or interference in the physical channels, and it is hard to distinguish these errors caused by semantic noise and channel propagation. Therefore, in order to interpret the meanings successfully at a semantic destination, we need to overcome not only physical channel noise, but also semantic noise in a semantic communication system.

Source/Channel Encoding and Decoding

In traditional wireless communications, data should be compressed by source encoding first and then by channel encoding to combat channel impairments, aiming to achieve an optimal transmission performance in each processing block. In semantic communications, semantic encoding not only compresses data at the source as much as possible, but also extracts the meanings and their semantic features of the data. The goals of semantic encoding are twofold [5], that is, maximizing expected faithfulness in representing observed worlds and minimizing the amount of data to be transmitted. The semantic features should be transmitted over a physical channel, and channel encoding should be added to improve the robustness. At a receiver, the received signal is decoded via channel decoding to extract the semantic features before the original messages are recovered.

Empowered by DL-based E2E communications and natural language processing (NLP) technologies, semantic encoding (decoding) and channel encoding (decoding) can be implemented by deep neural networks (DNNs). In such a system, semantic source and destination are auto-encoder and auto-decoder to perform semantic encoding (decoding) and channel encoding (decoding) jointly, which can achieve a global optimality if compared to the block-based structure in a traditional communication system.

Background Knowledge Bases

Different from traditional communications, another important characteristic feature in semantic communications is that a semantic communication system is a knowledge-based system [1]. This means that the semantic source and destination can establish their own background knowledge bases (KBs) by self-learning, just like human brains, which form the core of a semantic communication system. The KBs are the world models that the source and destination observed previously. The semantic source extracts the semantic information of the messages based on its KB. After receiving the messages, the receiver is able to interpret and infer the meanings of sent messages based on the destination KB. There are different types of KBs based on text, image, speech or video, and most works in the literature focused only on text or image based semantic communications due to mature DL-based NLP and image processing technologies. The establishment of KBs is a complex and time-consuming process, which basically is a learning process, just like the learning process through which humans learn knowledge of the world from a child to an adult. The KBs can learn from the perceived environment and can continuously expand and update their knowledge through training and sharing via communications. In addition, the KBs at the semantic source and destination may be different because the worlds and environments they observed are different and their abilities to understand things are also different, which may cause a semantic mismatch. However, as shown in Fig. 1, the semantic source and destination can share their KBs with each other in order to minimize the semantic mismatches.

Performance Metrics

Traditional communications need to minimize bit-error rate (BER) or symbol-error rate (SER) and transmit more bits utilizing as little communication resource as possible. In semantic communications, a receiver is supposed to extract the semantic information with the least ambiguity of the sent messages, and several performance metrics are used to ensure that the semantic information is transmitted and retrieved correctly. Different from the performance metrics used in traditional communication systems, the performance metrics used to measure semantic communication systems are diverse due to different KB types. We will discuss this issue in the sequel.

Text: For text messages in semantic communications, the performance can be measured by the similarity between the sent words or sentences and the interpreted words or sentences. The semantic dissimilarity between the two is measured by the semantic distance, which can be used to evaluate the distortion between the words on the semantic level [12, 13]. The average semantic distortion or error is defined as the average semantic distance in probability, which is statistically expressed by the probability of words and the conditional probability of receiving wrong meanings under the condition of the sent messages. The word error rate is an edit distance normalized by the length of a sentence [6]. Another common measurement is the bilingual evaluation understudy (BLEU) score, which measures the similarity between decoded text and raw text. However, because the BLUE score compares only the difference between the two text messages, it cannot distinguish more subtle difference in words, such as polysemy and synonym [3]. Thus, the sentence similarity is proposed to calculate the semantic similarity between the originally sent sentence and recovered sentence [7].

Image: Two performance metrics for image messages were proposed. First, the performance of an image semantic communication system can be measured by peak signal-to-noise ratio (PSNR), which is the ratio between the maximum signal and noise powers [9]. If the mean squared error (MSE) between the transmitted image and the reconstructed image is smaller, PSNR is larger and the reconstruction quality of the image is better. PSNR can also be used in video transmission because video files consist of many image frames. Furthermore, for image recognition, recognition accuracy is a measurement for a joint transmission-recognition scheme in an image semantic communication system [10].

Speech: In the literature, very few works focus on speech based semantic communications. In a very limited number of existing works, signal to distortion ration (SDR) was utilized to measure the errors between raw speech vector and restructured speech sequence, where a higher SDR indicates the fact that the reconstructed speech signal is easy to understand [11]. Another good metric for speech based semantic communications is perceptual evaluation of speech distortion (PESQ), as proposed in [11], which evaluates various speech signal conditions, such as background noise, analog filtering, and so on.

Finally, it should be noted that Shannon information theory does provide a design guidance for semantic communications. The semantic information of text, image or speech should eventually be encoded into bit streams and then transformed into physical signals for their transmission via communication channels. Thus, modulation, demodulation and other signal processing schemes are also required in semantic communications, and advanced wireless communication technologies can improve the efficiency of semantic communication systems.

An Overview of Semantic Communications

In the previous section, we compared semantic communications with traditional communications from several aspects, and illustrated how a semantic communication system works. In this section, we will introduce detailed semantic communication system models, including E2E semantic communication systems and multi-user semantic communication systems.

Different from traditional communications, another important characteristic feature in semantic communications is that a semantic communication system is a knowledge-based system. This means that the semantic source and destination can establish their own background knowledge bases by self-learning, just like human brains, which form the core of a semantic communication system.

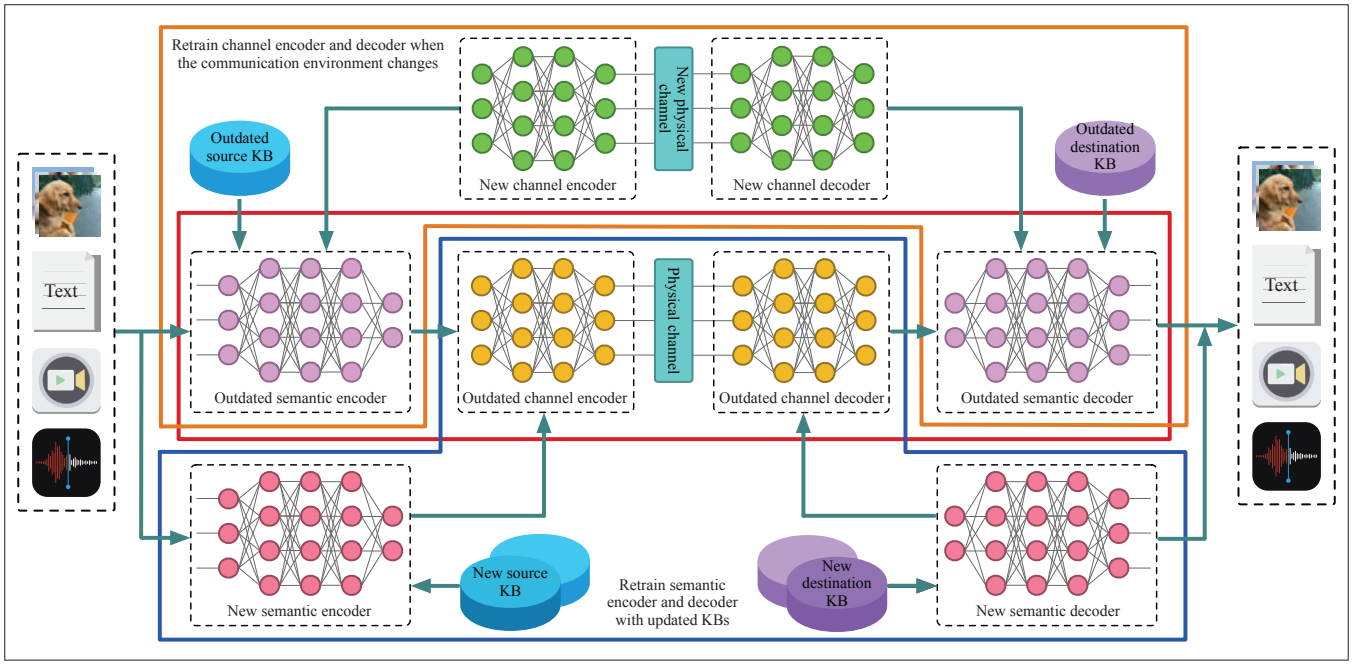

FIGURE 3. A DL-based semantic communication model.

End-to-End Semantic Communications

A generic DL-based E2E semantic communication model is shown in Fig. 3, where the semantic encoder (decoder) and channel encoder (decoder) are implemented by DNNs. With a given static source and destination KBs and a communication environment, the semantic encoder (decoder) and channel encoder (decoder) are trained jointly by a stochastic gradient descent (SGD) algorithm, as shown in the blocks highlighted by the red lines in Fig. 3. Due to the generalization ability of DNN, it is possible to send a new message which may not be included in the source and destination KBs without degrading interpretation performance. However, if a sent message is from a different type of KB, for example, an image in a linguistic system, it must take a lot of time to retrain the semantic encoder (decoder). Owing to dynamic communication environment and KBs' expansion/update, transfer learning is an effective and efficient approach to train the encoder (decoder) due to the flexible structures of DNNs [3]. The encoder and decoder training and KB updating processes in a dynamic communication environment are denoted by yellow and blue blocks in Fig. 3, respectively. If the communication environment is changing, the channel encoder and decoder should be retrained using a new channel model, while the parameters of the semantic encoder and decoder remain invariant. Otherwise, if the semantic source and destination update their KBs through learning and sharing, the semantic encoder and decoder should be retrained based on the new KBs with the given parameters of channel encoder and decoder.

Next, we will give an overview of semantic communications according to different types of transmitted semantic information, including text, image, and speech.

Text: An intuitive approach to preserve the semantic similarity between two words is to assign them similar indexes. A semantic index assignment problem is formulated when assigning a binary codeword for each word, where semantic similar words are coded with a short Hamming distance, and semantic independent words (i.e., most different codewords) are coded with the longest Hamming distance [12, 13]. For instance, "car" and "automobile" are coded by "0010" and "0000", and the semantic index of "magician" is "1011". In this way, the words reconstructed at a receiver via inverse index assignment have their semantic similarity very close to the words transmitted by a semantic source, in spite of the presence of channel noise and interference. Semantic index assignment is a good way to distinguish semantic similar and semantic independent words when the number of words is limited. However, the length of a codeword is exponentially proportional to the number of words, which makes the assignment process extremely time-consuming and complicated.

Recently, DL was proposed to be used in joint source-channel coding (JSCC), benefiting from the powerful representation capability of DNNs. Inspired by the success of DNNs in NLP, Farsad *et al.* [6] proposed a JSCC scheme for text-based semantic communications, where the encoder and decoder were implemented by two recurrent neural networks (RNNs), and the channel was represented by a dropout layer. Compared to a separate source-channel coding (SSCC) scheme, the DL-based JSCC scheme offers better performance. Although there are some insignificant errors, such as punctuation errors and so on, the semantic information could be conveyed accurately.

Enabled by intelligent E2E communications, a novel framework of semantic communication systems was proposed in [7], which aimed to design a joint semantic source and channel coding scheme while maximizing system capacity. This work considered technical level and semantic level jointly. In addition, based on a transformer and self-attention mechanism, a destination can easily understand long sentences. Considering a dynamic communication environment with different background

Source type	Metrics	NN	Loss function	Research content	KB	Dynamic environment	References
Text	Average semantic distortion	–	–	Binary codeword design for words	Limited word set	–	[12]
	Average semantic error	–	–	Semantic communication in the presence of an external entity, friend or foe	Limited word set	–	[13]
	Word error rate	RNN	–	Joint source-channel coding for text transmission	Proceedings of the European Parliament	–	[6]
	BLEU score and sentence similarity	Transformer	Cross-entropy and mutual information	Design of DL-based semantic communication systems	Proceedings of the European Parliament	Transfer learning	[3, 7]
	BLEU score	Transformer	Cross-entropy	A lite semantic communication system design for IoT networks	Proceedings of the European Parliament	–	[8]
Image	PSNR	CNN	Average MSE	Joint source-channel coding for image transmission	CIFAR-10 image dataset	–	[9]
	Recognition accurary	ResNet, CNN	Cross-entropy	Joint image transmission-recognition scheme for the IoT devices	CIFAR-10 image dataset	–	[10]
Speech signal	SDR and PESQ	Attention mechanism SE network	MSE	Semantic communication system design for speech signals	Edinburgh DataShare	Transfer learning	[11]

Notation
" – " indicates that the information is not available in the literatures.
NN: Neural network.
KB: Knowledge base.
RNN: Recurrent neural network.
BLEU: Bilingual evaluation understudy.
DL: Deep learning.
IoT: Internet of things.
PSNR: Peak signal-to-noise ratio.
CNN: Convolutional neural network.
MSE: Mean squared error.
SDR: Signal to distortion ration.
PESQ: Perceptual evaluation of speech distortion.
SE: Squeeze-and-excitation.

TABLE 1. A summary of semantic communication systems.

KBs, the authors in [3] utilized transfer learning to train the semantic encoder (decoder) and channel encoder (decoder) DNNs jointly.

Image: The structure of an image-based semantic communication system is almost the same as that of a text-based semantic communication system. The biggest difference between the two lies on the structure of DNNs. The transformer, dense layer [3, 7] and long short-term memory (LSTM) [6] have been widely utilized to extract semantic information from text messages. However, convolutional neural network (CNN) has its advantages in image processing, and thus it is more efficient for image feature extraction. A CNN-based E2E JSCC scheme was proposed for the first time in [9] to transmit high-resolution images in both AWGN and Rayleigh channels. Compared to most conventional compression algorithms, such as JPEG and JPEG2000, the proposed scheme showed a graceful performance and did not suffer a "cliff effect". Moreover, a joint transmission and recognition scheme was proposed in [10] to improve recognition accuracy in the Internet of Things (IoT) networks. Dense layer, convolutional layer and ResNet were used in image feature extraction and recognition, because video files are comprised of different image frames, such that the image semantic communication systems are basically able to transmit videos by extracting the feature of each frame.

Speech: Based on advanced NLP technologies, speech can be translated into text effectively, and then the text can be transmitted into semantic communications. However, unlike text that consists of characters only, speech signal is more complex and more difficult to deal with, because its quality involves not only speech signal fidelity and loudness, but also its frequency and tone. Speech signals can convey emotions, such as happiness, sadness, doubt, and so on. The same text may express different emotions in speech-based semantic communications. For instance, on one hand, "What is wrong with you?" can express a kind of concern for a person's health condition in a gentle and low intonation. On the other hand, it is a complaint if a person speaks in an anxious and high intonation. In this case, text cannot express the exact emotion of the transmitter. In addition, it is difficult to recognize a dialect in speech signals. Considering the nature of speech, Weng *et al.* [11] utilized attention mechanism squeeze-and-excitation (SE) networks to capture imperfections and non-linearities of speech signals. It was shown that this system performs better than a traditional communication system, and is more robust even in a low SNR region. In Table 1, a summary of recently reported works on semantic communications is given.

So far we have introduced single-modal semantic communication systems. However, multi-modal semantic communications should not be ignored, where semantic destination may receive different types of messages from a source. To the best of our knowledge, there is no specific research focusing on this aspect in the literature. It is possible for a semantic source to send text (or speech) and the destination to receive speech (or text). A method for this multi-modal semantic communication system can be realized based on advanced NLP technologies, where the sent text (or speech) can

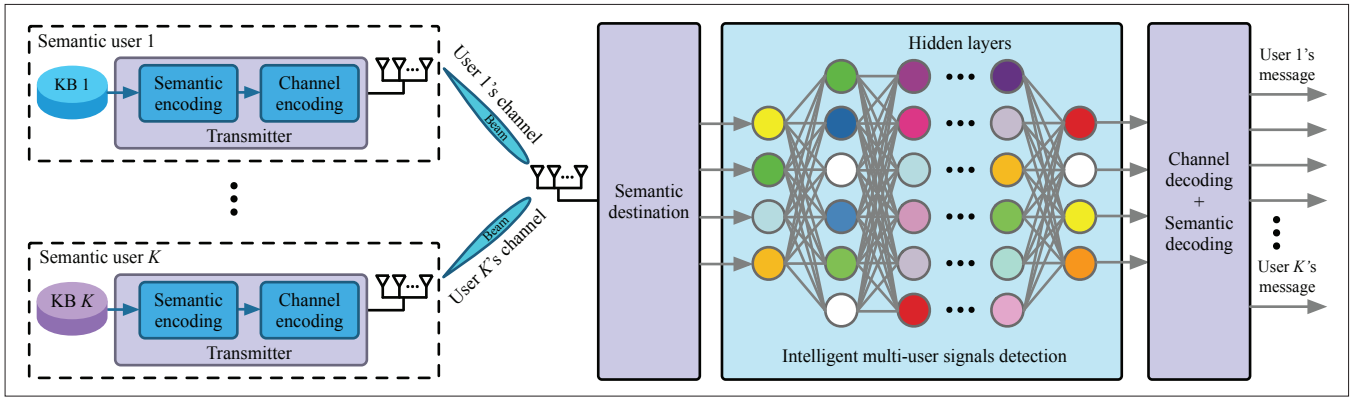

FIGURE 4. A structural architecture of multi-user semantic communication systems.

be translated into speech (or text).

Multi-User Semantic Communications

In the previous subsection, several semantic communication systems were introduced, where DL plays an important role in semantic information extraction and communication processes. However, all aforementioned systems do not involve multi-user transmissions. In general, connectivity density in 5G is 10^6 devices per km^2, whereas the connectivity density in 6G networks will grow up to 10 times of 5G, and area traffic density should be one hundred times of 5G, which requires a significant improvement in spectrum efficiency. In this context, NOMA is an efficient way to improve spectrum efficiency in semantic communication systems. As we discussed in the earlier sections, the KBs in a semantic communication system may be very much different from each other, which makes it necessary to design multi-user semantic communication systems, where users can learn from each other to communicate as human beings. For instance, Alice and Eve are native English speakers who cannot speak Chinese, and Bob can speak Chinese only. If Alice and Bob talk to Eve at the same time, Eve can only understand Alice and the Chinese from Bob is ignored. In this way, even though the Chinese seems to behave like an interference from Bob, Eve can still get the English messages from Alice without semantic errors. Similarly, in a semantic communication system, due to the diversity in KBs, multi-user signals can be transmitted using the same channel resources, such as frequency or time-slot. In this way, the bandwidth can be saved for an improved spectrum efficiency. However, multi-user signal detection and the complexity of interpretation process at a receiver are critical issues.

Here, we propose a structural architecture of multi-user semantic communication systems based on intelligent radio (IR), as shown in Fig. 4. A receiver in IR can estimate the channel state information (CSI) of each user and separate multi-user signals by training an intelligent multi-user signal detection DNN [14]. Then, the separated signals can be decoded by channel decoding and semantic decoding. Moreover, equipped with multiple antennas at a transmitter and a receiver, beamforming and precoding techniques can be used to enhance multi-user signal transmissions in a semantic communication system due to the spatial diversity gain of MIMO. It is an interesting research topic to study beamforming and precoding based multi-user semantic communication systems for multi-user signal detection. With IR and multiple antenna technologies, although the KBs of two users are not the same, it is possible to separate their messages at the receiver.

Use Cases of Semantic Communications

The use cases of semantic communications are extremely important for the implementation of semantic communication systems in the future. Three possible use cases of semantic communications are illustrated in Fig. 5.

IoT Networks

In various data monitoring applications from desert, ocean, cities, and home, IoT devices play significant roles in 5G and B5G networks. The wide proliferation of various intelligent devices, such as VR/AR glasses, unmanned aerial vehicle (UAV), sensors, and so on, has pushed IoT networks to provide more advanced functions, which in turn consumes more radio resources. Furthermore, in status update systems or age of information (AoI) aware systems, the IoT devices should perceive their working environments and upload the real-time status of environment information to cloud centers or mobile edge computing (MEC) servers for signal analysis and processing. Thus, the IoT devices have to support many sophisticated functions such as intelligent monitoring, data process, communications, and so on. Obviously, the data transmitted from IoT devices are mostly time-sensitive, which may not require a very high data rate but need a low latency and high accuracy.

Semantic communication is a promising technology for accurate and real-time data transmissions in IoT networks [8] as it consumes little radio resource and is relatively insensitive to channel noise. However, due to limited computation and storage capabilities, the IoT devices cannot use complex DNNs on board, and how to train the semantic encoder (decoder) and channel encoder (decoder) is a crucial issue in IoT networks. In order to simplify the structure of DNNs, model compression can be achieved with the help of network sparsification and quantization [8]. Furthermore, federated learning (FL) and distributed learning can be an alternative option to train a DL-based semantic communication system efficiently. DNN models in FL can be trained for a large number of IoT devices jointly, and the training process can be coordinated by a cloud/edge server. The parame-

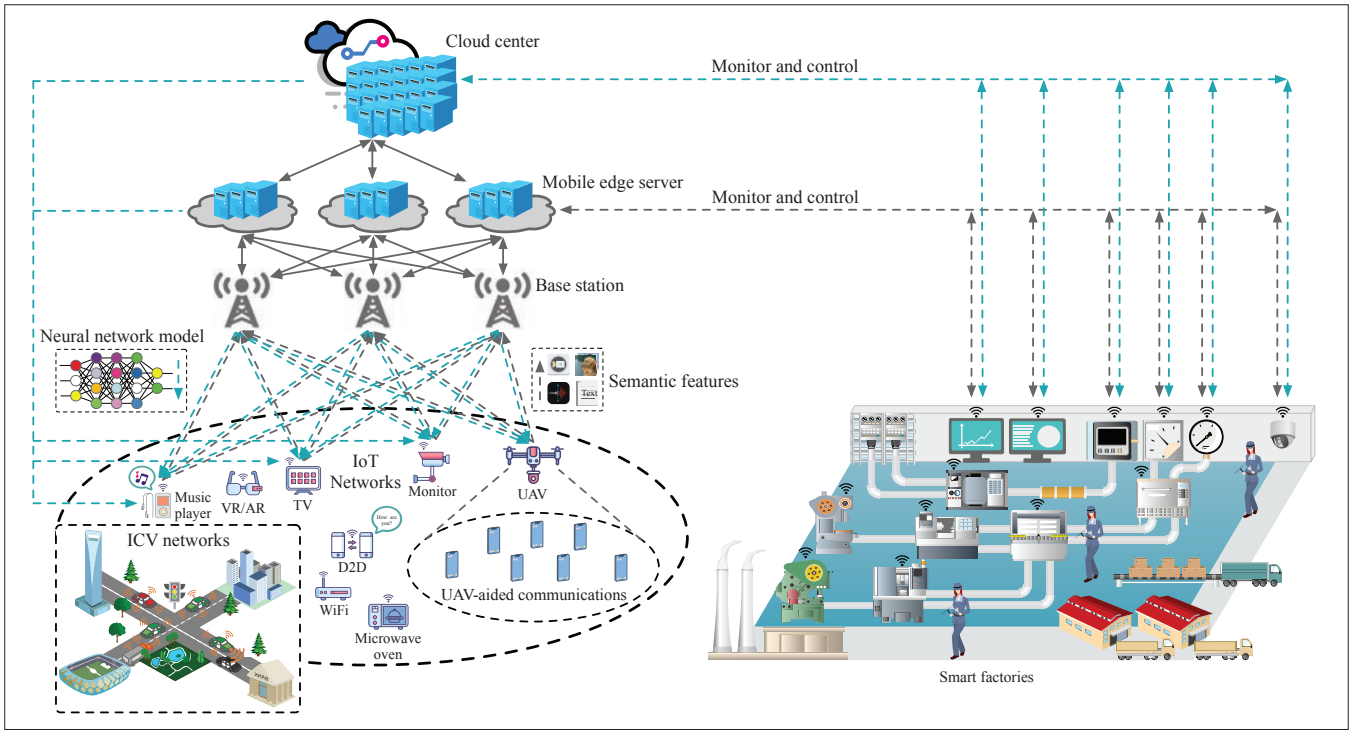

FIGURE 5. Three use cases of semantic communications, including IoT networks, ICV networks, and smart factories.

ters of DNNs at the IoT devices are uploaded periodically to cloud/edge servers for updating, and then the updated parameters are fed back to the IoT devices, such that each IoT device does not need to consume a lot of computing power for training. Another advantage is that FL may utilize the diversity in locally sensed data at IoT devices to speed up the training process.

Intelligently Connected Vehicle Networks

Another application of semantic communications is intelligently connected vehicle (ICV) networks, where a vehicle can perceive information in the environment and predict its driving trajectory, traffic flow, network congestion, CSI, and so on. In ICV networks, a vehicle with various sensors may generate about tens or even thousands of gigabytes per day, including videos and images of traffic information. Although most of these data are processed at vehicles or discarded, the amount of remaining data uploaded to roadside units (RSUs) or cloud/edge servers can be huge, and the uploading latency should be as short as possible. Comparisons of average run time between semantic communication systems and traditional schemes in [3, 9] suggest that DL-based semantic communications can help to compress and extract semantic information to reduce latency, so that semantic communications are applicable to a large amount data transmission in ICV networks with a low latency. In addition, semantic information is more robust against channel noise and interferences than bit streams in traditional communications, which enhances the reliability of data transmission and improves driving and road safety in ICV networks. Moreover, in device-to-device (D2D) based vehicular communications, vehicles usually share radio resources with cellular users in an underlay fashion, which may cause severe co-channel interferences. In semantic communications, due to the diversity of KBs, this interference can be minimized as long as a receiver can understand the meanings of transmitted messages.

Smart Factories

Semantic communications can also be applied to smart factories. Smart factories rely on communications between machines and the interactions between human and machines. Furthermore, advanced communication technologies, such as 5G and DL-based communication technologies, make factories more intelligent, efficient, energy saving, and environmentally friendly. In futuristic smart factories, unmanned management, real-time control and monitoring are important features to run machines and equipment. The semantic features of the monitoring information, such as the status of machines, temperature, humidity, and so on, can be extracted and uploaded to a central controller or a cloud/edge server to analyze the status of materials and the quality of products. Another important issue in smart factories is to control the operation of machines to perform a specific action. In this sense, semantic control is an efficient way to achieve goal-centric communications, where semantic information of the control signals is conveyed to the machines [15]. The operational efficiency of smart factories is improved and the communication cost can be reduced as only the intentions (i.e., the semantic information) of control signals are transmitted and fewer errors will occur.

Both ICV networks and industrial Internet attach great importance to reliability, which can be ensured by channel encoding and decoding in semantic communications, where structured redundancy is introduced to improve the robustness against interference/noise in the channels and increase communication reliability.

The KBs are always being expended and updated frequently, just like the learning process of human beings, which makes the sharing process much longer and more difficult. Thus, how to communicate, share and infer semantic information with inconsistent KBs is a wide open issue in semantic communications.

Open Issues

Despite the fact that semantic communication is not a completely new research topic, which can be traced back to the pioneering work of Weaver [2], there are a variety of challenges to be tackled before it can be used in real applications. In this section, we would like to list several major open issues for future investigations.

Insufficient Theoretical Research on Semantic Communications

Classic semantic information theory (CSIT) was introduced first by Carnap and Bar-Hillel in 1952 [4] based on logical probability. Inspired by this seminal work, some theoretical research works have been done in the literature in the past two decades, such as [5, 13] and the references therein, but they are not sufficient, especially on semantic communications in a framework based on DL. Some important challenges in theoretical research include the lack of theoretical guidance for joint semantic-channel coding designs. The technical level and semantic level should be considered jointly (such as the impact of data transmission rate on semantic communications, how much semantic information can be transmitted in a wireless channel). More investigations on SIT under interference channels are also needed, plus a specific definition of a semantic channel and its capacity, and so on. A general framework for a DL-based semantic communication system should be explored, including its proper performance metric, suitable DNN architecture, and so on.

Inconsistent KBs at Semantic Source and Destination

In semantic communications, the KBs of the semantic source and destination are normally inconsistent. Although these KBs can be made more homogeneous through KB sharing, it is an extremely time-consuming and resource-consuming process because the sharing between KBs needs effective communications between the semantic source and destination. The KBs are always being expended and updated frequently, just like the learning process of human beings, which makes the sharing process much longer and more difficult. Thus, how to communicate, share and infer semantic information with inconsistent KBs is a wide open issue in semantic communications.

Multi-User Interpretation Algorithm Design

As we discussed in the previous sections, multiple users can utilize the same frequency or timeslot to transmit semantic information due to the diversity in their KBs. However, the complexity of semantic information interpretation at a receiver in a multi-user environment is very high because it must consider multi-user detection, channel decoding and semantic decoding jointly. In addition, the KB at a receiver should include different types of data in order to separate multiple users' messages. More effective and yet efficient interpretation algorithms for joint semantic-channel decoding of an intended user should be designed. Although an IR based approach was proposed in the literature, it is just the first step in multi-user semantic communication systems. Thus, in our future work, more research efforts should be made on the design of low complexity multi-user interpretation algorithms.

Effectiveness Level in Semantic Communications

The design of any communication systems aims to minimize the consumption of channel resources, such as bandwidth and power. The effectiveness level in semantic communications is responsible for the management of radio resources. Despite the fact that many works have been done on radio resource management (RRM) in traditional communication systems, their effectiveness in semantic communications remains to be verified, and more comparisons with traditional communication systems are important for practical applications of semantic communications. Moreover, in resource limited IoT networks, resource consumption for download and parameter upload in a DNN model, and gradient back propagation (BP) of the parameters from a receiver to a transmitter, should also be considered carefully.

Implementation of Semantic Communications

Today, the research on semantic communications is only in its infant stage. It is widely believed that more theoretical research can definitely help to promote real implementation of semantic communication systems, as theoretical incompleteness in semantic communications may restrict its implementation. In addition, it is still a question in both industry and academia whether we really need semantic communications as the existing communication technologies are very mature. The investigations of DL-enabled semantic communication systems are limited not only by semantic theories, but also by AI hardware, which is required to run DNNs efficiently at a relatively low cost. At present, a mobile system-on-chip (SoC) with a dedicated AI core, for example, a neural network processing unit (NPU), is able to run AI models on embedded AI accelerators. The most advanced mobile SoCs, such as Snapdragon 888 and HiSilicon Kirin 9000 manufactured by a 5 nm silicon process, can execute DL models in even a few milliseconds for image classification, face recognition, and so on.[1] However, these SoCs still cannot meet ultra-low latency requirements in wireless communications, for example, 1 ms in 5G ICV networks and 0.1 ms in B5G networks. Therefore, it is a big challenge to develop a DL-based E2E semantic communication system based on these SoCs, and more advanced microelectronic and chip technologies are needed to address this problem.

Conclusions

This article gives an overview of the most recently reported works on feature extraction based semantic communications, which are relevant to future intelligent communications. Semantic communication works very much differently from traditional communication in many aspects, such as communication channels, source and channel encoding (decoding) schemes, performance metrics, and so on. Moreover, the design of an E2E semantic communication system is related to the types of messages transmitted, and thus the DNN structures of source encoder (decoder) and channel encoder (decoder) can be very much different. In particular, the use cases in IoT networks, ICV networks and smart factories were discussed for possible implementation of seman-

[1] A performance ranking of mobile SoCs can be found in AI-Benchmark (https://ai-benchmark.com/ranking_processors.html), which provides 46 AI and computer vision test results obtained by neural networks running on smartphones and measured in terms of more than 100 different AI performance metrics, such as speed, accuracy, initialization time, and so on.

tic communications. In addition, the open issues were summarized to highlight the challenges in theoretical research and practical implementation of semantic communications. In summary, semantic communications will definitely play an important role in the development of futuristic AI-based communication technologies beyond 5G.

Acknowledgment

The work presented in this article was supported in part by the Natural Science Foundation of China (No. U1764263) and Taiwan Ministry of Science and Technology (Nos. 109-2221-E-006-175-MY3 and 109-2221-E-006-182-MY3).

References

[1] E. C. Strinati and S. Barbarossa, "6G Networks: Beyond Shannon Towards Semantic and Goal-Oriented Communications," *Computer Networks*, vol. 190, no. 8, May 2021, pp. 1–17.
[2] W. Weaver, "Recent Contributions to the Mathematical Theory of Communication," *The Mathematical Theory of Communication*, 1949.
[3] H. Xie *et al.*, "Deep Learning Enabled Semantic Communication Systems," *IEEE Trans. Signal Process.*, 2021, doi: 10.1109/TSP.2021.3071210.
[4] R. Carnap and Y. Bar-Hillel, "An Outline of a Theory of Semantic Information," RLE Technical Reports 247, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge MA, Oct. 1952.
[5] J. Bao *et al.*, "Towards a Theory of Semantic Communication," *Proc. 2011 IEEE Network Science Workshop*, West Point, NY, USA, 2011, pp. 110–17.
[6] N. Farsad, M. Rao, and A. Goldsmith, "Deep Learning for Joint Source-Channel Coding of Text," *Proc. 2018 IEEE Int'l. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 2326–30.
[7] H. Xie *et al.*, "Deep Learning based Semantic Communications: An Initial Investigation," *Proc. 2020 IEEE Global Commun. Conf.*, Taipei, Taiwan, Dec. 2020, pp. 1–6.
[8] H. Xie and Z. Qin, "A Lite Distributed Semantic Communication System for Internet of Things," *IEEE JSAC*, vol. 39, no. 1, Jan. 2021, pp. 142–53.
[9] E. Bourtsoulatze, D. Burth Kurka, and D. Gündüz, "Deep Joint Source-Channel Coding for Wireless Image Transmission," *IEEE Trans. Cogn. Commun. Netw.*, vol. 5, no. 3, Sept. 2019, pp. 567–79.
[10] C. Lee *et al.*, "Deep Learning-Constructed Joint Transmission-Recognition for Internet of Things," *IEEE Access*, vol. 7, 2019, pp. 76 547–61.
[11] Z. Weng, Z. Qin, and G. Y. Li, "Semantic Communications for Speech Signals," 2020; available: https://arxiv.org/abs/2012.05369.
[12] B. Guler and A. Yener, "Semantic Index Assignment," *Proc. 2014 IEEE Int'l. Conf. Pervasive Computing and Communication Workshops (PERCOM WORKSHOPS)*, Budapest, Hungary, Mar. 2014, pp. 431–36.
[13] B. Gler, A. Yener, and A. Swami, "The Semantic Communication Game," *IEEE Trans. Cogn. Commun. Netw.*, vol. 4, no. 4, Dec. 2018, pp. 787–802.
[14] Q. Yu *et al.*, "Intelligent Radio for Next Generation Wireless Communications: An Overview," *IEEE Wireless Commun.*, vol. 26, no. 4, Aug. 2019, pp. 94–101.
[15] M. Kountouris and N. Pappas, "Semantics-Empowered Communication for Networked Intelligent Systems," 2021; available: https://arxiv.org/abs/2007.11579v3.

Biographies

Xuewen Luo received his B.E. degree in communications engineering from Jilin University, Changchun, China, in 2017. He is currently working toward his Ph.D. degree with the School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin, China. His research interests include physical layer security, intelligently connected vehicle networks, mobile edge computing, and semantic communications.

Hsiao-Hwa Chen [S'89, M'91, SM'00, F'10] is currently a Distinguished Professor in the Department of Engineering Science, National Cheng Kung University, Taiwan. He obtained his B.Sc. and M.Sc. degrees from Zhejiang University, China, and a Ph.D. degree from the University of Oulu, Finland, in 1982, 1985, and 1991, respectively. He has authored or co-authored over 400 technical papers in major international journals and conferences, six books, and more than ten book chapters in the areas of communications. He has served as the general chair, TPC chair, and symposium chair for many international conferences. He has served or is serving as an editor or guest editor for numerous technical journals. He is the founding Editor-in-Chief of Wiley's *Security and Communication Networks Journal*. He was the recipient of the best paper award at IEEE WCNC 2008 and the recipient of the IEEE 2016 Jack Neubauer Memorial Award. He served as the Editor-in-Chief for *IEEE Wireless Communications* from 2012 to 2015. He was an elected Member-at-Large of IEEE ComSoc from 2015 to 2016. He is serving as TPC Chair for IEEE Globecom 2019. He is a Fellow of IEEE, and a Fellow of IET.

Guo Qing received the B.S. degree in radio engineering from Beijing Institute of Posts and Telecommunications in 1985, and M.S. and Ph.D. degrees in information and communication engineering from Harbin Institute of Technology in 1990 and 1998, respectively. He is a professor in the School of Electronics and Information Engineering at Harbin Institute of Technology, director of the Key Laboratory of Wideband Wireless Communications and Networks, Heilongjiang Province, China. His research interests include satellite communications, space information networks and wireless communication networks, etc. He has published one authored book and more than 200 papers in journals and international conferences.

It is widely believed that more theoretical research can definitely help to promote real implementation of semantic communication systems, as theoretical incompleteness in semantic communications may restrict its implementation. In addition, it is still a question in both industry and academia whether we really need semantic communications as the existing communication technologies are very mature.