Semantic Communications: A Paradigm Whose Time Has Come

Emrecan Kutay and Aylin Yener
INSPIRE@OhioState Research Center
Dept. of Electrical and Computer Engineering
The Ohio State University
kutay.5@osu.edu, yener@ece.osu.edu

***Abstract*—Communication system design to date is predicated on principles that abstract information as digital sequences irrespective of their meanings. Such a semantic-agnostic approach leads to fundamental limits that are application and technology independent, and offers efficient engineering of communication systems. Emerging applications that involve communications however, call for going beyond the engineering problem of reliably reconstructing a digital sequence. Most current communication devices are computing devices that execute tasks. Increasingly communication is needed for a purpose and is integrated with decision making, machine learning and sensing. Further, human networks operate over machine networks, which necessitate more sophisticated human-machine communication. The goal of this vision paper is to advocate for Semantic Communications, i.e., communication system design that, at the outset, pays attention to the content, its meaning, context, and purpose. We will argue that taking into account the meanings and context of information can lead to better communication designs. In particular, we argue in favor of semantic distortion, a novel metric introduced nearly a decade ago, based upon which communications systems design aiming to convey the meaning and purpose can be designed. We review the current efforts of Semantic Communications which has recently become a popular area of 6G, and potential directions of Semantic Communications, which can explore various different directions with novel metrics including using Knowledge Graphs (KL), information theoretic approaches, and machine/Deep Learning (DL).**

***Index Terms*—Semantic Communications, Task Oriented Communication, Context Awareness, Deep Learning, Information Theory.**

I. Introduction

Shannon established the fundamental limits of representation and communication of information in 1948 in his landmark paper [1]. A fundamental starting assumption in Shannon's approach is to strip the meanings from the communication messages: "Frequently the messages have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering." [1]. This allows for representing messages as digital sequences and the problem of reliable reconstruction becomes one of identifying one of these sequences. The resulting framework which quantifies information and connects the limits of compression and reliable communication to the information measures have been the cornerstone of communications system design ever since. In particular, insights quantifying information capacity and separation of source and channel coding have arguably led to today's compression and communication as we know it.

Shortly after Shannon's work, this fundamental assumption has been brought into question. Shannon and Weaver identified three levels where a problem is likely to occur during communication [2]: i) technical level; ii) semantic level; iii) effectiveness level. The first level deals with the accurate transmission of symbols while the second level focuses on how the meanings are conveyed over the link. The third level investigates the impact of the transmitted message on the receiver and how effectively it is processed. These levels of classification clearly highlight the semantic aspect, and despite having focused on the first level, shows the broader thinking early on.

Over the years, practical metrics such as throughput, Bit Error Rate (BER), and Symbol Error Rate (SER) have been used to quantify the performance of communication links. Additionally, decades of research brought us closer to Shannon's fundamental limits. The digital revolution transformed society at large, to the point that we rely on devices that are designed by these principles for all aspects of life.

This brings us to present day where a new revolution is brewing: 6G. The envisioned applications set forth for the next decade in 6G clearly points to a more natural and integrated human-machine interaction, including fundamentally better ways for machines to analyze and respond to human behavior. Additionally, emerging applications like remote surgeries, digital twins, fully autonomous vehicles, smart environments [3], [4], [5], [6] are all demanding with respect to wireless resources (bandwidth), and call for better ways of communicating *needed information* with reliability and ultra-low latency, rather than all information. For these applications to be pervasive, we need to consider the content, the meaning and the goal of messages. In other words, to enable future applications, we need to add the second level (and even third) of communication of [2].

This paper reviews the past decade on semantic communications research and aims to emphasize the potential benefits of this approach. Specifically, we will argue that unlike classical communications, semantic communications

978-1-6654-7300-2/22/$31.00 ©2022 IEEE
DOI 10.1109/CIC56439.2022.00020

focus on only conveying the information that is pertinent to the transmission task and/or its meaning. This mind-set naturally leads to resource savings which can be significant with careful design and tailored to the application. This approach requires tools to extract the semantic information, and metrics to quantify the semantic knowledge for sources with different modalities. There are different approaches utilizing knowledge graphs (KL), information theory (IT), and deep learning (DL) models in the literature. These will be summarized in the next sections starting with the early efforts defining the semantic communication metric.

II. Semantic Communications: The beginnings

Although, including meanings of messages has been considered in various works following Shannon, there have not been significant advances in this area until recently. [7] is among the earliest references to recognize that semantic communications can lead to better utilization of resources. In particular, a transmitter design that minimizes semantic distortion of the receiver has been considered, effectively replacing the age-old syntactic Word Error Rate (equivalently BER or SER) with semantic error as the performance metric.

Conventional transmitter design maps source symbols to codewords (index assignment) to minimize the symbol errors over the noisy communication link [8], [9]. The inclusion of semantics in this design, naturally calls for a mapping that optimizes semantic error, called *semantic distortion* in [7]. This differentiation between the errors is captured via the utilization of a semantic measure. In [7], a taxonomy-based semantic measure was used. The similarity between two concepts is defined as:

$$sim(c_i, c_j) = max_{c \in S(c_i, c_j)}[-log(p(c))] \tag{1}$$

where $p(c)$ is the probability of the concept c, S is the set of general concepts in taxonomy including both c_i and c_j. The similarity of words is then:

$$sim(w_i, w_j) = max_{(c_i, c_j)}[sim(c_i, c_j)] \tag{2}$$

Expressions c_i, c_j denote the different senses (meanings) of the words w_i, w_j respectively. The average semantic distortion is then given by:

$$D(\pi) = \sum_{i=1}^{|W|} p(w_i)(\sum_{j=1}^{|W|} p(\pi(w_j)|\pi(w_i)) * d(w_i, w_j)) \tag{3}$$

where $\pi(.)$ is the index assignment, $p(w_i)$ denotes the probability of word w_i, and $p(\pi(w_j)|\pi(w_i))$ denotes the probability of receiving the codeword assigned to word w_j when codeword of w_i was transmitted [7].

Minimizing semantic distortion leads to mappings where semantically similar codewords are placed to points that are close in terms of Hamming Distance [7]. The provides a semantically similar codeword that is likely to be decoded in case of error brought on by channel noise, minimizing the semantic loss of the link. The inclusion of meanings in the metric removes the limitation in conventional systems where the errors are treated equally independent of their similarity in semantic space. Hence, errors in transmission become *interpretable* from the human perspective. As an example, consider in a noisy communication link, the transmitted message is "A car is approaching", and receivers A and B received messages "A table is approaching", and "A vehicle is approaching" respectively. Although the errors of A and B are similar from the traditional perspective (with BER and SER), the semantic error of B is much lower compared to A. This example also connects us to *context-aware* communications, one can also see that the message received by A is received in error immediately. Utilizing context or knowledge can lead to better designs.

The concept of Knowledge Graph (KG) [10] is utilized in [7] to achieve the minimum number of codewords. The setup in [10], [11] includes two users, each with their own messages (facts), x_i and $\hat{x}_i$ for users 1 and 2 respectively. Users are interested in reaching conclusions by exchanging facts in an interactive communication scenario. Knowledge graphs for both users are constructed with individual facts as the vertex set. An edge between two facts from the user's knowledge, (x_i, x_j) is present if they reach a conclusion with the same fact, $\hat{x}_k$, from the other user. The problem then becomes graph coloring with vertex set, *W*, and edge set, *E*, where the chromatic numbers of graphs are taken into consideration. To ensure valid coloring, a function was introduced [7],

$$f(c) = \sum_{\substack{(w_i, w_j) \in E \\ w_i, w_j \in W}} I_{\{c(w_i) = c(w_j)\}} \tag{4}$$

Where,

$$I_{\{c(w_i)=c(w_j)\}} = \begin{cases} 1, & \text{if } c(w_i) = c(w_j) \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

$c(w_i)$ is the proposed color for the word w_i and c is the proposed coloring. For coloring, c, to be valid, $f(c) = 0$ in (4). Adopting this constraint to index policies, the mapping that minimizes the average semantic distortion $D(\pi)$ over all valid colorings is then sought in [7]:

$$\begin{aligned} \boldsymbol{\pi_{opt}} &= argmin_{\pi} D(\pi) \\ &\text{s.t.} f(\pi) = 0 \end{aligned} \tag{6}$$

This optimization problem is NP-hard and thus simulated annealing was employed to find near-optimal mappings for large scale problem. It is worthwhile to note that with the advent of data-driven machine learning methods, semantic distortion optimization is much more within reach, see, for example, the next section for extensions utilizing deep learning.

Bringing the importance of context and side information to the forefront, [12] further explored a game theoretic approach to obtain optimal transmission policies for minimum end-to-end semantic error. In addition to the transmitter decoder pair, the system model considered in [12] includes an agent who can influence the communication process by supplying side information. One can imagine such an agent being helpful by providing context to the decoder as to the incoming message.

It can also be detrimental by providing adversarial side information. Modeling a probabilistic agent, a Bayesian game is played where Player 1 tries to optimize the encoding and decoding (the system), and Player 2 is the agent (influencer whose nature is uncertain). Corresponding payoff functions were set for the players and Nash Equilibrium is investigated. Furthermore, the case when the agent is known to be helpful for sure is investigated for achieving minimum semantic error where the behavior can be considered as side information to the decoder [12].

These early studies show the potential benefits of including semantics in communication system design, and introduce pathways to explicitly taking semantics of information into account in future designs. In recent years, especially with the advent of deep learning, there has been as resurgence of interest from the research community in various directions of content-aware communications design, i.e., bringing semantics. Today, semantic communications is emerging as a major direction in the design of 6G systems.

III. Recent Literature

A. Rate-Distortion Approaches

There is recent work that frames semantic communications as rate-distortion problems. Of specific interest is indirect rate-distortion or remote source coding. Indirect rate-distortion problem, introduced in [13], adds a noisy channel between a source *S* and the encoder, meaning the encoder observes a noisy source. This is a classical problem in information theory. The output of the encoder propagates through a noisy communication channel and is obtained at the receiver. Witsenhausen studied this problem further in [14] by considering side information at the decoder and providing a reformulation for the expected distortion function. This formulation is widely used in recent studies [15], [16], [17]. These works consider the input channel as a semantic channel that introduces distortion to the source parallel to the notion started by [7].

In [15], data at the encoder is represented with a tuple *(S, X)*. In this notation, *S* is the semantic information of the source and is not directly observable. *X* is defined as appearance of *S* to the encoder. The receiver obtains $\hat{X}$ and $\hat{S}$ through a communication channel. Distortion constraints D_s and D_a are defined separately for (S, $\hat{S}$) and (X, $\hat{X}$) respectively. The problem was considered as a combination of rate-distortion with multiple distortion constraints and indirect rate-distortion problems [15]. These recent works can be seen as theory approaches to semantics, although they mainly remain in the Shannon theory domain.

B. Semantic Communications Utilizing Deep Learning Models

In [18], a communication system is modeled as an autoencoder. Different from traditional models, the transmitter, and receiver are jointly optimized for end-to-end data reconstruction. Although, not specific to semantic communications, this work certainly has motivated researchers in communications to build deep learning-enabled models, where the source and channel coding blocks are jointly optimized [19].

In this approach, source and channel encoder/decoder blocks are replaced with neural networks and the channel is represented with a non-trainable noise layer. The problem is considered as an end-to-end reconstruction task and training is performed with the selected loss function. For this structure, there are recent approaches to replace the source coding neural network with models that are successful at extracting features from the source in DL tasks. This structure thus will perform semantic coding *to preserve the meaning* through communication.

In [20], joint source-channel coding (JSCC) for text transmission is done utilizing long-short-term memory (LSTM) networks. LSTM models are widely used in natural language tasks and are utilized in this model to capture the meanings [21]. Authors in [22] design a DL-enabled JSCC-based system for text transmission. Transformers are used to capture the meanings thanks to their attention mechanism. The use of transfer learning in cases of knowledge and channel alterations is investigated. A sentence similarity metric was defined to evaluate the similarity level of the two sentences that extends the word similarity in [12] to the whole sentence. In addition to text, the transmission of speech signals is investigated in [23], where Squeeze and Excitation (SE) Networks are used to extract essential information from speech signals. The extension in [24] utilizes CNN and RNN models for a speech recognition task where speech is captured at the transmitter and text is generated at the receiver. Instead of sending speech signals, extracted semantic features are transmitted decreasing the traffic. To generalize the concept to multi-user multi-modal cases, a study in [25] was conducted. Machine Translation (MT), and Image Retrieval (IR) tasks are selected for multi-user case and Visual Question Answering (VQA) task is selected for multi-user multi-modal case. A unified transformer structure was proposed to process the data for multiple tasks and multi-modals. Results show that this framework increases robustness by providing enhanced communication in low SNR regimes and achieving higher semantic similarity scores compared to the traditional separation-based methods like Huffman, Brotli (for text), Adaptive Multi-Rate Wideband (AMR-WB) (for speech) JPEG (for image) for source coding and Turbo, Reed-Solomon (RS), Polar Codes for channel coding [22], [24], [25].

For image transmission, DL enabled JSCC model was proposed in [19] involving CNN to process the image source. It was seen that model outperformed the state of art techniques, JPEG source coding with capacity achieving channel coding by offering a higher PSNR score. Similar work was done in [26] for video transmission. Video signals are first considered as a sequence of images and selected key frames from this sequence are sent by using a similar technique in [19]. Motion information was extracted from the remaining frames and encoded by an interpolation encoder. Results showed that the proposed model outperformed the state of art H.264 video encoder with LDPC channel coding. [26].

All of these recent works showcase the utility of deep learning in designing non-traditional end-to-end communication systems, essentially affirming the merits of joint source-

channel coding in practical systems of the future. Extracting and conveying the *needed information* and explicitly taking the content semantics into account integrates well with end-to-end approaches employing deep learning.

IV. Conclusion and Forward Look

In this vision paper, we have discussed the emerging field of Semantic Communications. We have talked about the motivation for and origins of semantic communications and summarized recent developments that nicely dovetail with the recent advents in machine learning for communications. For a more comprehensive survey on semantic communications and the adjacent area of goal oriented communications, the reader is referred to [27].

In this paper, we have argued the case for replacing the traditional symbol/word error metrics that have driven practical communication network design for decades, with *semantic* error. Such a leap brings in the possibility of end-to-end designs conveying *useful* information to the destination. This mindset departs from measuring the performance of a network by throughput, which separates communications from the actual goal of the network. As 6G designs on the horizon point to convergence of task execution, computing, learning and communications with distributed networked intelligence, semantic communications paradigm, we posit, is needed to execute this vision.

Semantic communications is still a nascent field. There are a number of exciting directions and formulations taking semantics into account in network design. There is also significant attention from researchers towards defining the transformative directions. There are as many recent opinions on what semantics in communications is as is not. While the power of deep learning is evident for content and semantics aware multi-modal communications design of the future, it is also important to note that exciting directions remain on the modeling side as well, in particular towards quantifying semantic information. Further, better integration with natural language processing is needed in order to utilize context and structure in content better. We conclude by advocating for hybrid approaches: a mix of theoretical foundations and practical designs, for example building on rate-distortion/JSCC framework, and considering semantic error as an application dependent metric; and a mix of learning and model based designs harnessing the power of deep learning and (wireless) network models.

References

[1] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 625–656, July, Oct. 1948.

[2] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*. Urbana, IL: University of Illinois Press, 1949.

[3] F. Jameel, Z. Chang, J. Huang, and T. Ristaniemi, "Internet of autonomous vehicles: Architecture, features, and socio-technological challenges," *IEEE Wireless Communications*, vol. 26, no. 4, pp. 21–29, 2019.

[4] D. Gündüz, D. B. Kurka, M. Jankowski, M. M. Amiri, E. Ozfatura, and S. Sreekumar, "Communicate to learn at the edge," *IEEE Communications Magazine*, vol. 58, no. 12, pp. 14–19, 2020.

[5] L. Atzori, A. Iera, and G. Morabito, "The internet of things: A survey," *Comput. Netw.*, vol. 54, no. 15, p. 2787–2805, oct 2010.

[6] S. Mihai, M. Yaqoob, D. V. Hung, W. Davis, P. Towakel, M. Raza, M. Karamanoglu, B. Barn, D. Shetve, R. V. Prasad, H. Venkataraman, R. Trestian, and H. X. Nguyen, "Digital twins: A survey on enabling technologies, challenges, trends and future prospects," *IEEE Communications Surveys & Tutorials*, pp. 1–1, 2022.

[7] B. Guler and A. Yener, "Semantic index assignment," in *IEEE Int. Conf. on Pervasive Comp. and Comm. Workshops (PerCom)*, 2014, pp. 431–436.

[8] X. Wang, X. Wu, and S. Dumitrescu, "On optimal index assignment for map decoding of markov sequences," in *2006 IEEE International Symposium on Information Theory*, 2006, pp. 2314–2318.

[9] X. Wu, H. D. Mittelmann, X. Wang, and J. Wang, "On computation of performance bounds of optimal index assignment," in *2010 Data Compression Conference*, 2010, pp. 189–198.

[10] B. Guler, A. Yener, and P. Basu, "A study of semantic data compression," in *2013 IEEE Global Conference on Signal and Information Processing*, 2013, pp. 887–890.

[11] B. Guler, A. Yener, P. Basu, and A. Swami, "Two-party zero-error function computation with asymmetric priors," *Entropy*, vol. 19, no. 12, 2017.

[12] B. Güler, A. Yener, and A. Swami, "The semantic communication game," *IEEE Transactions on Cognitive Communications and Networking*, vol. 4, no. 4, pp. 787–802, 2018.

[13] R. Dobrushin and B. Tsybakov, "Information transmission with additional noise," *IRE Transactions on Information Theory*, vol. 8, no. 5, pp. 293–304, 1962.

[14] H. Witsenhausen, "Indirect rate distortion problems," *IEEE Transactions on Information Theory*, vol. 26, no. 5, pp. 518–521, 1980.

[15] J. Liu, W. Zhang, and H. V. Poor, "A rate-distortion framework for characterizing semantic information," in *2021 IEEE International Symposium on Information Theory (ISIT)*, 2021, pp. 2894–2899.

[16] J. Liu, S. Shao, W. Zhang, and H. V. Poor, "An indirect rate-distortion characterization for semantic sources: General model and the case of gaussian observation," *IEEE Transactions on Communications*, vol. 70, no. 9, pp. 5946–5959, 2022.

[17] T. Guo, Y. Wang, J. Han, H. Wu, B. Bai, and W. Han, "Semantic compression with side information: A rate-distortion perspective," *arXiv preprint arXiv:2208.06094v1*, 2022.

[18] T. O'Shea and J. Hoydis, "An introduction to deep learning for the physical layer," *IEEE Transactions on Cognitive Communications and Networking*, vol. 3, no. 4, pp. 563–575, 2017.

[19] E. Bourtsoulatze, D. Burth Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 3, pp. 567–579, 2019.

[20] N. Farsad, M. Rao, and A. Goldsmith, "Deep learning for joint source-channel coding of text," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 2326–2330.

[21] X. Liu, D. Cao, and K. Yu, "Binarized LSTM language model," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 2113–2121.

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

[23] Z. Weng and Z. Qin, "Semantic communication systems for speech transmission," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 8, pp. 2434–2444, 2021.

[24] Z. Weng, Z. Qin, and G. Y. Li, "Semantic communications for speech recognition," in *2021 IEEE Global Communications Conference (GLOBECOM)*, 2021, pp. 1–6.

[25] H. Xie, Z. Qin, X. Tao, and K. B. Letaief, "Task-oriented multi-user semantic communications," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 9, pp. 2584–2597, 2022.

[26] T.-Y. Tung and D. Gündüz, "Deepwive: Deep-learning-aided wireless video transmission," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 9, pp. 2570–2583, 2022.

[27] D. Gunduz, Z. Qin, I. E. Aguerri, H. S. Dhillon, Z. Yang, A. Yener, K. K. Wong, and C.-B. Chae, "Beyond transmitting bits: Context, semantics, and task-oriented communications," *To appear in IEEE Journal on Selected Areas in Communications, arXiv preprint arXiv:2207.09353v2*, 2022.