

Challenges and Directions for Semantic Communication

Volkan Rodoplu, Member, IEEE, and Snehal S. Vadvalkar, Student Member, IEEE
Department of Electrical and Computer Engineering
University of California, Santa Barbara

Abstract—This paper aims to present ideas to bridge the gap between two conflicting views of information: Shannon information, which is used to model physical layer information, and semantic information, which is used in databases, distributed systems, human-computer interaction, and in the application layer of the OSI communication stack. We introduce the notion of a semantic domain, and define the amount of semantic information relative to a semantic domain. We show that translations between semantic domains achieve a consistent picture of semantic information. We indicate where and how the conceptual divide occurs between semantic information and Shannon information. We discuss mobile devices and users and a central server, as a possible setting for the application of the concepts developed in this paper. The discussion is aimed at pointing out the need for a unified theory, and at directions for how such a theory can be pursued.

I. INTRODUCTION

This paper aims to describe the need and directions for building a framework for semantic communication. The recent developments in communications and computing have led to the convergence of these technologies, as well as the vertical integration of the layers of the OSI stack that have been traditionally separate. The design of the physical layer of the stack has been driven by Shannon’s information theory [1], which models the source and the channel probabilistically, and the semantics of the data are ignored at this layer. However, the upper layers must continue to view information as semantic information. This results in two conflicting views of information, and no satisfactory solution to guide cross-layer designs have been developed so far.

The main difference in the two views lies in where and how the information is measured. The following simple example illustrates the difference: Examine two sources S_1 and S_2 , where S_1 produces two letters a with probability $1/3$, and b with probability $2/3$, whereas S_2 produces two words *apple* with probability $1/3$ and *orange* with probability $2/3$. Shannon’s theory views the information rate of these two sources as the same, namely $H(1/3)$, where $H(p)$ is the binary entropy function. Even though the same optimal encoding is used at the channel, when the words are detected, and stored at the receiver, the semantic information that is stored in these two cases is very different. Each time a bit is detected in R_2 , namely the receiver of S_2 , it indexes into an entire word, rather than a single letter.

In general, we can show that by indexing into a word space, rather than a letter space, we can achieve a much faster communication in the semantic sense, i.e. in the units

of semantic knowledge transmitted across per second. The reason that Shannon’s framework does not capture this is that Shannon entropy does not model how the different source alphabets, i.e. that of S_1 and S_2 are related to each other, and how the alphabet of one source may subsume that of the other. In order to be able to measure the difference, we must construct a representation of semantic knowledge at the receiver, so that these different pieces of information can be compared. Hence, a semantic information model must be receiver-centric, and must be able to relate the information across different streams that are operating with different source alphabets, even though their optimal encoding at the channel might be the same.

Continuing with the above simple example, a semantic model of each receiver’s knowledge space can be built as follows: The semantic atoms of R_2 may be restricted to the set of valid words in the English language, including many of the proper nouns, as is done in online extended dictionaries. The semantic atoms of R_1 are letters. Now, the key is that a “translation” exists between these two “semantic domains.” Namely, the mapping that maps a word such as “hello” into its constituent letters ‘h’, ‘e’, ‘l’, ‘l’, ‘o’, performs the translation of every word in the semantic domain of R_2 into the semantic domain of R_1 . The inverse mapping does not exist for many strings that are recognizable by R_1 but not by R_2 because they are not valid English words. However, this phenomenon is one of the key properties of semantic information hierarchies. Namely, the receivers such as R_2 that have compactly encoded the world knowledge, operate at a higher level of abstraction using a sparse subset of the total number of strings that can be represented as random combinations of all of the constituents at the lower layer. What has been gained at the higher level is compaction in the amount of memory space that must be devoted to storing the different combinations, as well as a decrease in the access time. For example, in a speech recognition system, R_2 would aim to distinguish between only the sounds that correspond to valid words, rather than the sounds that correspond to all possible combinations of the letters (or all phonemes in the case of speech recognition).

An important goal of a theory of semantic information is the specification of such translations between different semantic domains. A related important goal is to specify a measure on the information tokens to characterize the “amount of information” that has been transmitted. As discussed in the example above, this amount of semantic information is not the Shannon information capacity of the channel since

that measure does not distinguish between different source alphabets.

In the past, there have been a few papers that proposed a theory of semantic information. The first of these was by Bar-Hillel and Carnap [2]. The major difference between that theory and the directions outlined in this paper is that we are motivated by emerging applications of a plethora of devices with different capabilities, and we aim to drive towards a theory that can achieve a unified view of information at the upper layers and the physical layer. Reference [3] outlines a quantitative theory of strongly semantic information based on truth-values rather than probability distributions. Again, the setting there is a very small subset of the range of applications for which we aim to model semantic information. A survey of the semantic conceptions of information appears in [4].

The rest of the paper is organized as follows: In Section II, we introduce the notion of a “semantic domain” and discuss how to measure semantic information. In Section III, we describe a model for semantic communication between a transmitter and a receiver, where these are construed as structures with memories that have a world knowledge representation. In Section IV, we present ideas to investigate the ultimate limits of semantic communication. In Section V, we give an example application setting to which this framework can be applied. In Section VI, we present our conclusions.

II. SEMANTIC DOMAINS AND MEASURING SEMANTIC INFORMATION

So far, we have used the term “semantic domain” loosely. We now make this notion precise. We define the semantic domain (J_R, P_R) of a receiver R as a set of objects J_R and a set of operations P_R that operate on the set of objects and on the products produced. (Hence, the definition is recursive but well-defined.) For the above example, the semantic domain of R_1 is $J_{R_1} = \{a, b, c, \dots, y, z\}$ where the set of operations on the objects P_{R_1} is chosen as follows: Each time we use a letter, we assume that we invoke an atomic operation *construct* for that letter. In addition, we have operations to concatenate letters as many times as we wish. The set of operations P_{R_1} is chosen to be the union of these construct operations, and the concatenation operation. The semantic domain of R_2 is J_{R_2} is the set of all valid words in the English dictionary, and set of operations on the objects P_{R_2} is the union of the set of construct operations, one for each word, and the concatenation operation that concatenates words with a single space in between.¹ Now, the semantic information theory specifies the set of translations between these two semantic domains, as we discussed before. The question is now how the amount of information should be measured.

¹Neither of these receivers’ semantic domains can produce grammatically correct sentences; however, such receivers with a grammar can be defined. Our use of the term “semantic domain”, which is very general, should not be confused with the syntax-semantic separation in linguistics. In our case, semantic domains are capable of specifying a grammar as a set of rules by which the objects in their domain can be combined. More sophisticated models can incorporate some semantic rules in linguistics to form such sentences with meaning. However, even with the state of the art in linguistics research today, this is possible to only a very limited extent.

It can be seen that the *same* information, represented in different domains, will have different measures, in terms of the number of objects required in that domain, for the construction of this information. For example, the word “apple” is a single object in J_{R_2} ; it requires that the “apple” object be constructed in the semantic domain of R_2 , but does not need to invoke the concatenation operation at all. We may define the amount of semantic information of an information token, as the *minimum* number of operations that need to be performed in a given semantic domain, in order to construct that token using the operations in that semantic domain. In the case of the word “apple”, only a construct operation on the word “apple” with no concatenations are needed; hence, the semantic information measure of this in (J_{R_2}, P_{R_2}) is 1. However, the same information token would require a total of 5 constructs, and 4 concatenations in (J_{R_1}, P_{R_1}) , resulting in a semantic information measure of 9, in that domain.

This example shows a more general fact, namely that a single measure of information cannot be assigned uniquely to an information token. The “amount of information” in an information token is different, depending on in which semantic domain that information token is being constructed and represented. However, the translations between the semantic domain assure that, when referenced to a single semantic domain, the same information token results uniquely in a single number to represent the amount of information in that token. Hence, we can say that the amount of semantic information in a piece of information is relative to the particular semantic domain in which the information is represented, but translations retain the equivalence of the same information token across different semantic domains.

III. SEMANTIC COMMUNICATION BETWEEN A TRANSMITTER AND A RECEIVER

In our discussion so far, we have focused on a very simple couple of semantic domains. To generalize from these slightly further, we may think of each semantic domain as a database where particular connections between the objects have been stored. In this model, each semantic domain is a set of objects, and the main operation is *Connect*(a, b) which draws an arc between a and b to indicate that these are related. Such a model can be used as a first order model for classical semantic networks in the literature, that place the most closely related words together. This model also allows us to begin to model two cognitive entities and the communication between them. Assume that there is now a transmitter T and a receiver R such that the transmitter has a semantic domain (J_T, P_T) and the receiver has a semantic domain (J_R, P_R) , where the semantic domains contain objects of knowledge (such as words) in the object sets, and the set of operations is given by these *Connect*(a, b) functions. This model may be used, for example, for acquisition of relationships between different objects at the transmitter, by the receiver. A direct application is the synchronization of data between a desktop and a Palm Pilot, for a bidirectional version of this communication. This may also serve as a model for the transfer of semantic world relationships acquired by two robots, who are now

synchronizing their knowledge with each other. We shall focus on the one-directional case where T would like to transfer all the knowledge that it has to R . Now, the part that needs to be transferred is the part that R does not know, namely $G_T \setminus G_R$, where G_T and G_R are the knowledge graphs of T and R that show the acquired connections, respectively. We shall assume here that T 's knowledge is more accurate or recent, hence, it will override the formed connections in G_R if no such connection exists between those two objects on G_T .

There are many algorithms in the literature [5][6][7] that aim at synchronizing G_R with G_T with the minimum number of bits sent from T to R . The key point for our purposes is to note that such minimal representations *depend* on the existing knowledge set of R , namely the graph G_R . If $G_T = G_R$, then there is nothing new that R can learn from T . In Shannon theory, a similar notion in terms of $H(X|Y)$ is formulated, where X and Y are (possibly correlated) random variables. In order to create a unified framework between the upper layers that formulate similar notions for semantic information, and the physical layer that formulates a similar notion in terms of conditional entropies of random variables, one may attempt to proceed as follows: One may aim to model the knowledge base at T and at R as matrix random variables where the set of objects $J_T = J_R$ are fixed; however, the connections are random. One may then formulate a model for how we expect the probabilities of these connections to be correlated between G_T and G_R . Based on this, one may derive $H(G_T|G_R)$, namely how much extra information T may impart to R . However, imposing this model on this application would be a very forced solution because the entropy framework presumes a continuous transmission of information between two entities whose realizations of the underlying processes must be ergodic. No such continuous communication exists in the original model, where two robots may be synchronizing only once, before they part their ways. Further, the amount of information transmitted in that case is finite, and likely to be small; hence, there is no assurance that the processes be ergodic, or that long enough data from T are transmitted such that Shannon's channel coding arguments (which presume long channel codes) may apply.

This brings us to another important divide between semantic information at the upper layers, and Shannon information at the physical layer. At the upper layers, we would like to model a finite, small amount of information that needs to be transmitted only once. Shannon information models assume a long, continuous stream of data coming from an ergodic underlying process. The way that the two connect is as follows: At the upper, semantic layers, the number of objects in the semantic domain is small, and this semantic domain uses a very sparse subset of the entire space available, when the objects are decomposed into lower-layer objects. We saw this in the example of the two receivers $R1$ and $R2$ that operated on letters versus words, in Section I. Hence, what appears as a finite, small amount of information in the semantic domain in the upper layers, expands, when mapped to its constituents at the lowest layers. This can be visualized as a pyramid, as shown in Fig. 1. Shannon information theory

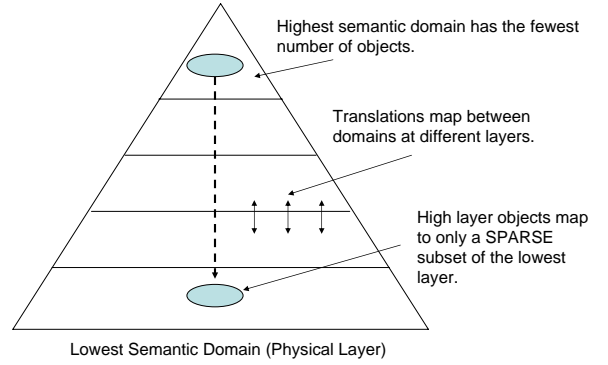


Fig. 1. The relationship of the objects at different semantic layers or layers of abstraction. The lowest layer has the most number of objects, and thus can be viewed as a stream and modelled probabilistically. The upper-most layer has the fewest objects and results in the view of information where a few objects are transmitted only once. The challenge lies in unifying these views of information.

takes this expanded version at the lower layers, treats it as an ergodic source, and models the probability distribution of these long streams. However, when “translations” are performed, in the sense that we defined them in this paper, to upper-layer semantic domains, there is much semantic structure in those long streams. The Shannon theory implicitly assumes that the receiver is not intelligent enough to see this semantic structure. This also simplifies the communication model by providing a universal binary interface for all physical communication.

However, with today's technology, the situation is very different. We can index directly into the upper semantic layers of the communication stack. This requires the development of new technologies that build a more sophisticated “codebook” that has objects, not simply as long streams of bits, but rather as objects and combinations of objects in the upper-layer semantic domains of the transmitters and receivers, where we no longer conceive of the transmitter and the receiver as physical layer entities, but rather as cognitive entities, and represent the entire stack reaching their semantic layers. For example, if T knows that R is able to perform predicate logic, or basic forms of reasoning with a given piece of data, then, this might change the encoding in the sense that the channel might need to be much fewer times (in the Shannon sense) in order to send a fixed amount of semantic information as measured in a common reference semantic domain. Here, we see that the aim is to communicate semantic information, as it should be, and Shannon-Kolmogorov information emerges as the minimum description length of the semantic information, where the minimum description length is computed, not traditionally by treating the data as strings of 0's and 1's and forming an adaptive dictionary, but in terms of the semantic domain objects and operations that are common to the transmitter and the receiver. Hence, we see that the minimum description lengths, in this setting, will be *relative* to the common set

of semantic domain objects and operations of the transmitter and receiver; that is, the minimum description length will not be a *universal* measure of information. We believe that this is the right direction for the theory to proceed, given the proliferation of devices with widely different capabilities in today's communication and computing market.

IV. INVESTIGATING LIMITS OF SEMANTIC COMMUNICATION

In this new setting, we may ask the question that Shannon asked originally, namely, "What is the fundamental limit to the communication of information?", where we now take information to mean semantic information. It is illustrative to examine media in which the channel itself is not the main bottleneck. For example, when two humans communicate using language, at near distance, the channel can be assumed to be very clear and is not the main bottleneck to the communication of information. In this setting, the bottleneck to how much information can be communicated between two humans is whether a common set of abstractions exists at both ends, into which the transmitter and the receiver can index. So far in the paper, we have modeled these semantic domains (the objects and the operations) as fixed; however, the brains on which they are implemented can be plastic. Hence, when this plasticity, or the ability to learn new objects and operations is incorporated into the model, the limits of communication may be substantially increased. Our second model, where we assumed an operation $Connect(a, b)$ existed at both nodes, can now be expanded, using plasticity, to create new operations, such as $ConnectAsSquare(a, b, c, d)$, which can connect these four nodes into a square. These new operations can be learned and can add to the set of common operations of two communicating systems. Then, the communication can more efficiently index into these complex instructions, to communicate the desired semantic information token. Hence, in this setting, we may say that a fundamental limit to the communication of semantic information is the plasticity of the brains or substrates on which the semantic domains are constructed. The higher the plasticity, the more both the range and the efficiency of the communication are expected to be.

V. EXAMPLE APPLICATION

The application of communicating semantic information to a mobile user that utilizes a PDA device may be taken as a setting to apply the above concepts. In this setting, the semantic domain may be modeled by treating the user's brain and the device itself as a single entity. The upper layer semantic operations, such as reasoning, are performed by the user, and lower layer operations such as physical layer transmission are performed by the device. The device and the user also interact with each other through a well-defined interface, such as that of a Blackberry or Palm. In this setting, we would be interested in questions such as "What is the maximum rate at which semantic information can be communicated to this user plus the device, by the server?" Note that, here, the semantic

information will be represented as residing, eventually, in the brain of the user.

Two tasks can be distinguished in this setting: Learning tasks, in which the user is learning some new application, such as using a GPS functionality, or a device interface, and second, information retrieval tasks where the procedural knowledge of the user is assumed to be fixed, and we focus on the amount of semantic information that can be retrieved from the server. In the former task, a key measure by which we may infer the amount of semantic knowledge gained is to measure the minimum amount of time taken to carry out a particular learning task, where the minimum is taken over all learning strategies. The larger this minimum time, the more semantic connections that can be assumed to have taken place in the user's brain. In the latter task, the amount of semantic information can be measured as a set difference between what the user already knows, in that particular semantic domain, and what the server can provide it as new knowledge. A good application is the use of maps in which the server can adapt the data that it transmits based on its current knowledge of the knowledge set of the particular user at hand. Such adaptive technologies are currently not used; e.g. Blackberry services are not customized to the knowledge base of particular users.

VI. CONCLUSIONS

The aim of this paper has been to highlight the need for a unified theory of semantic and physical layer communication, and to present some ideas for how such a unified theory can be pursued. We presented the idea of a semantic domain as a set of objects and a set of operations on the objects. A theory of semantic communication provides the translations between the different semantic domains, at different layers on the same machine, as well as across different machines. We argued that a single number cannot be assigned universally as a measure of the amount of semantic information; instead, it must be measured relative to a semantic domain. Translations between the semantic domains must find that the same information token is measured consistently across all of the domains.

We presented a semantic domain model in which the transmitter and the receiver have knowledge graphs and we wish to transfer the missing knowledge from the transmitter to the receiver. We argued that the Shannon model is not applicable to this scenario due to the fact that relatively small, finite amounts of information are to be transmitted in this case. We presented a connection between the semantic information and Shannon information models by way of a pyramid that shows how relatively sparse sets are mapped and expanded to long streams of physical layer information. We argued that new physical layer technologies must be developed that index directly into the semantic domains of the receivers, exploiting not only existing relationships between semantic objects but also the abilities to reason (such as via predicate logic) at the receiver. We argued that for a fixed amount of semantic information, dramatic reductions in the number of channel uses may be achieved by indexing into such cognitive abilities at the receiver.

We discussed a model where the semantic domains are constructed on plastic substrates, and hence new operations can be

learned, and then utilized to cut down on the communication cost in the communication of semantic information from a transmitter to a receiver.

Finally, we gave an example application setting of mobile devices and users, where the framework can be applied. The theory must be developed in such an application setting, first for a single server (transmitter) and user plus device (receiver). In the future, semantic information networks can be considered where the problems of peer-to-peer semantic communication may be posed for an ad hoc network of users. These indicate new directions for research in this area.

REFERENCES

- [1] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379-423 and 623-656, Jul. and Oct. 1948.
- [2] Y. Bar-Hillel and R. Carnap, "Semantic Information," *British Journal for the Philosophy of Science*, vol. 4, no. 14, pp. 147-157, 1953.
- [3] L. Floridi, "Outline of a Theory of Strongly Semantic Information", *Minds and Machines*, 14(2), 197-222.
- [4] "A survey of the semantic conceptions of information", <http://plato.stanford.edu/entries/information-semantic/>, 2005.
- [5] D. Barbara, Rj. J. Lipton, "A class of randomized strategies for low-cost comparison of file copies," *IEEE Trans. Parallel Distributed Systems*, pp. 160-170, Apr. 1991.
- [6] Y. Minsky, A. Trachtenberg, R. Zippel, "Set reconciliation with nearly optimal communication complexity," in *International Symposium on Information Theory*, Jun. 01, p. 232.
- [7] A. Trachtenberg, D. Starobinski, S. Agarwal, "Fast PDA synchronization using characteristic polynomial interpolation," in *Proc. IEEE INFOCOM 2002*, vol. 3, pp. 1510-1519, Jun. 2002.