Towards a theory of semantic communication

7 authors, including:

Prithwish Basu
Boston University
79 PUBLICATIONS   **1,843** CITATIONS

SEE PROFILE

Mike Dean
Raytheon BBN Technologies
47 PUBLICATIONS   **5,161** CITATIONS

SEE PROFILE

Craig Partridge
Raytheon BBN Technologies
131 PUBLICATIONS   **9,109** CITATIONS

SEE PROFILE

Ananthram Swami
Institute of Electrical and Electronics Engineers
455 PUBLICATIONS   **14,634** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project   Artificial Intelligence, machine learning, semantic query in Pathology and Laboratory Medicine View project

Project   CAMPFIRE View project

Towards a Theory of Semantic Communication

Jie Bao

Tetherless World Constellation
Rensselaer Polytechnic Institute
Troy, NY 03060
baojie@cs.rpi.edu

Prithwish Basu,Mike Dean,Craig Partridge

Raytheon BBN Technologies
Cambridge, MA, 02138
{pbasu,mdean}@bbn.com,craig@aland.bbn.com

Ananthram Swami

US Army Research Lab
Adelphi, MD 20783
ananthram.swami@us.army.mil

Will Leland

Raytheon BBN Technologies
Cambridge, MA, 02138
wel@bbn.com

James A. Hendler

Tetherless World Constellation
Rensselaer Polytechnic Institute
Troy, NY 03060
hendler@cs.rpi.edu

Abstract—**This paper studies methods of quantitatively measuring semantic information in communication. We review existing work on quantifying semantic information, then investigate a model-theoretical approach for semantic data compression and reliable semantic communication. We relate our approach to the statistical measurement of information by Shannon, and show that Shannon's source and channel coding theorems have semantic counterparts.**

I. Background

It has long been recognized that the broad subject of communication goes beyond what Shannon's theory [18] and many of its extensions cover. Weaver [21], just one year after Shannon introduced his information theory, proposed that communication involves problems at three levels as follows:

> "LEVEL A. How accurately can the symbols of communication be transmitted? (The technical problem.)
> LEVEL B. How precisely do the transmitted symbols convey the desired meaning? (The semantic problem.)
> LEVEL C. How effectively does the received meaning affect conduct in the desired way? (The effectiveness problem.)"

Shannon's Classical Information Theory (CIT) is deliberately focused on only Level A (technical level), thus, "semantic aspects of communication are irrelevant to the engineering problem" [18]. As a metaphor, Weaver said that "an engineering communication theory is just like a very proper and discreet girl accepting your telegram. She pays no attention to the meaning, whether it be sad, or joyous, or embarrassing". On the other hand, Weaver argued that Shannon's information theory is general enough to be extended

to consider communication on levels B and C, for instance, by adding "semantic transmitter", "semantic receiver" and "semantic noise" to Shannon's communication model. This vision is illustrated in Figure 1[1].
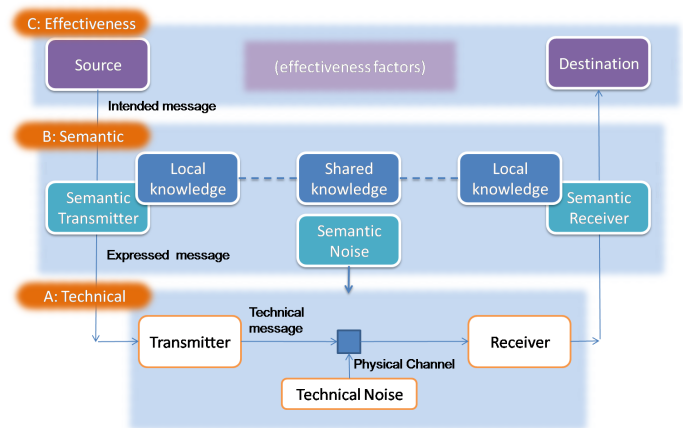

Fig. 1.   A 3-Level Communication Model

The assumption that "semantics is not relevant" is no longer true in many forms of modern communications, such as in database queries, distributed systems, human-computer interactions, and the Web (particularly the Semantic Web [3]). There is now a strong need for an extension of the classical communication model to characterize not only *sequences of bits*, but also the *meanings* behinds these bits. For this goal, various researchers have studied theories of "semantic information" (details discussed in Section II). Notable examples include the pioneering work of Carnap and Bar-Hillel [4], Floridi [8, 9], Barwise and Seligman [2, 17], among others.

However, a generic model of semantic communication, as suggested by Weaver, has still largely remained unexplored after six decades. Existing works on semantic information are limited in addressing some fundamental questions in

[1]Local knowledge and shared knowledge in the diagram are not mentioned by Weaver.

communication when the semantics of exchanged contents is no longer negligible. Some of these problems include: How can semantics help in data compression and reliable communication? How are semantic coding/decoding related to the engineering coding/decoding problems? What is semantic noise? Are there achievable bounds in semantic coding, analogues to the bounds established by Shannon in engineering communication? What factors should we consider to improve efficiency and reliability in semantic communication?

This paper summarizes some of our initial work in realizing Weaver's vision, by extending Shannon's theory of (technical) communication to a theory of Level B (semantic) communication. Our work is influenced by Carnap and Bar-Hillel [4], with new contributions in the following areas:

- We show that the work of Carnap and Bar-Hillel is a special case of a model-theoretical characterization of semantic information sources, and present a generic model of semantic communication;
- We discuss the role of semantics in reducing source redundancy, and establish theoretical bounds in lossless semantic data compression;
- We define the notions of semantic noise and semantic channel. By extending the Shannon's channel coding theorem, we obtain the semantic capacity of a channel.

The model developed in the paper is crude, and many non-trivial simplifications are made. Most importantly, the modeling of Level C (utility or effectiveness) communication is beyond the scope of this paper. We also note that the logic-based approaches we adopted may not be adequate to capture semantics in human communications. However, we believe that these simplifications are necessary for us to focus on the "core" issues of semantic communication, and that even this crude model readily yields some interesting results. We believe this model, after some of the suggested extensions, may form a foundation for a general theory of semantic communication.

This paper has an accompanying technical report which contains additional discussions and more detailed proofs: http://www.cs.rpi.edu/~baojie/pub/2011-03-28_nsw_tr.pdf

II. RELATED WORK

Efforts to extend CIT to capture semantic aspects of communication started shortly after Shannon published his paper. Carnap and Bar-Hillel (1952) [4] were among the first to introduce a "semantic information theory" (SIT), which is henceforth referred to as Classical Semantic Information Theory (CSIT).

They distinguish the concepts of information and the *amount of information*, and measure the amount of information in a sentence in a given language based on *logical probabilities* (as opposed to the statistical probabilities used in CIT) ranging over the contents. Intuitively, "A and B" has more information than "A" because it is less likely to be true: whenever "A and B" is true, "A" is true, but not vice versa. Similarly, "A" has more information than "A or B", and a tautology (which is trivially true) provides no information.

The logical probability of a sentence, therefore, is measured by the likelihood that the sentence is true in all possible situations. For instance, suppose "A" and "B" are independent of each other, and both are true or false as a result of the flip of a fair coin. There are 4 possible situations with equal possibilities (i.e., 0.25): A is false/true, B is false/true.

Therefore, "A and B" is true in one situation and its logical probability is 0.25. Similarly, the logical probability of "A or B" is 0.75. These can be denoted using a function m as:

$$m(A \wedge B) = 0.25, m(A \vee B) = 0.75$$

The amount of semantic information in a sentence A is defined as the negative logarithmic value of $m(A)$, i.e.,[2]

$$H_s(A) = -\log_2(m(A))$$

Thus, $H_s(A \wedge B) = 2$ and $H_s(A \vee B) = 0.415$, while $H_s(A) = H_s(B) = 1$, matching the intuitions given above.

It has been shown that logical inference does not provide additional semantic information, that is:

$$A \vdash B \Rightarrow H_s(A) \geq H_s(B)$$

where $\vdash$ is the logical entailment relation.

Essentially, CSIT can be regarded as a *model-theoretical approach* to assign probabilistic values to logical sentences. Since paper [4] is limited to propositional logic, Carnap and Bar-Hillel use truth tables (with each row called a "state description"), which can be seen as the universe of all possible models of a propositional sentence, to find the chance that a sentence is true. In CSIT, there is a close relationship between the quantity of information in a sentence and the set of its models. If a consistent sentence has fewer models, it is more "surprising" and contains more information. This is similar to the probabilistic logics of Nilsson [16] and Bacchus [1], which can be extended to first-order languages.

In [8, 9], Floridi developed a *Theory of Strongly Semantic Information* (TSSI). One of his major motivations is to solve the so-called Bar-Hillel-Carnap Paradox (BCP) in CSIT, which states that contradictions have an infinite amount of information, i.e., $m(\bot) = 0$, thus $H_s(\bot) = \infty$, where $\bot$ is shorthand for $A \wedge \neg A$ for arbitrary A. The basic idea is that the informativeness of a statement is measured by the positive or negative degree of semantic distance or deviation from "truth". This is quite different from CSIT, which defines informativeness as a function over all situations, not over a particular situation that is chosen to be true.

However, it has been noted that TSSI is incomplete with regard to quantifying all possible statements [6]. There exist propositional sentences that cannot be evaluated using the approach described in [8]. For these reasons, D'Alfonso ([6] section 4) proposed the "value aggregate" method that captures both inaccuracy and vacuity, based on formal models of truthlikeness. This method aggregates the differences of all models of a sentence to those of the "true" state.

---

[2]Carnap and Bar-Hillel used inf instead of H_s.

Both Floridi and D'Alfonso's approaches measure the *relative* information or misinformation of a statement against another "true" reference statement. Thus, the information value is always a value between 0 and 1. This approach is rooted in the semantic information framework using information flow and situation theory by Seligman and Barwise [2, 17] and that of Devlin [7]. However, Floridi and D'Alfonso's approaches cannot determine the objective amount of information when there is no reference statement. Essentially, their work offered a semantic *similarity* (or divergence) measurement between two sentences, not a measurement of *uncertainty* as Shannon, Carnap and Bar-Hillel proposed.

Several authors have investigated other approaches of modeling semantic information, e.g., algebraic information theories [12, 13], universal semantic communication [10, 11] and semantic coding [22]. Some recent work has been collected in two proceedings [15, 20]. However, these works do not offer a quantitative measure of semantic information in inference-capable sources, nor the study of the role of semantics in coding, which are our main foci.

III. Semantic Communication: a General Model

Before we can investigate the measurement of semantic information, we need to clearly define *semantic information* and *semantic communication*. The concept of semantic information is certainly not new. Here we will restrict ourselves to the engineering description of this notion. For more information about the philosophical account of semantic information, see the excellent survey in [9].

A. Goal of Semantic Communication

Note that there is a fundamental difference between the goal of engineering communication and that of semantic communication. Shannon stated in his paper [18] that

> The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point.

Weaver [21] stated that

> The semantic problems are concerned with the interpretation of meaning by the receiver, as compared with the intended meaning of the sender.

Comparing the two statements, we can state that the goal of semantic communication is *not* to reproduce, exactly or approximately, the *messages* transmitted, but their *interpretations*. For example, consider the conversation:

> Alice: "Are you free this weekend?"
> Bob: "No, I'm busy on both Saturday and Sunday."

Alice is a semantic source (sender) and Bob is a semantic destination (receiver). Bob is able to interpret the meanings of the received message and relates it to the meanings of the vocabulary he already knows. He knows that "free" is an antonym of "busy" and that "weekend" means "Saturday" or "Sunday". He is able to infer that "free this weekend" is the same as "not busy on both Saturday and Sunday", even if the two statements are syntactically different.

For a classical information source, a message is a *sequence of symbols*. In a semantic information source, a message, which may still be syntactically viewed as a sequence of symbols, is in fact an *expression* composed using the symbols in the language of the source. What we want to achieve is the faithful transmission of meanings of these expressions, not their syntactic representations, which is the concern of engineering communication.

Now consider a conversation between three persons:

> Alice: "Bob, is Charlie free this weekend?"
> Bob: "Charlie, if you are available this weekend?"
> Charlie: "No, I'm not available on both Saturday and Sunday."

Here Bob serves as a *semantic channel* between Alice and Charlie. Bob does not faithfully convey the original message from Alice, however, he is still able to preserve the original *meaning* of the message of the sender. There may be an engineering failure if we measure the success of communication "literally", but there is no semantic failure.

Even if there is no engineering communication failure, there may still be semantic communication failure. Considering a conversation about a "Lecturer" in universities, a US person who is not familiar with UK academic ranks may interpret it to be similar to a non-tenure-track position in US, whereas "Lecturer" in the UK is roughly equivalent to "Assistant Professor" in the US system.

B. Semantic Sources

A real world semantic source may be a complicated system which can make statements with subtle semantic distinctions. In this paper, we will not try to model every form of semantic source, but a very basic type that can make factual statements in propositional logic. This simplification will help us focus on the key modeling problem, and we will discuss its extensions later.

For a hypothetical example, suppose a child asks her father what is "Tweety". The father, as an information source, may do the following:

- (Observing World) He searches the Web and finds a webpage about Tweety. There are many such pages. Most of them are about Tweety the bird, but a few are about a Twitter client, or a basketball player.
- (Inferring) Depending on which page the father visits and trusts, the father may use his knowledge to come up with an appropriate answer for his child. For instance, the webpage may tell him that "Tweety is a canary", but since the child may not understand "canary" yet, and the father knows that canaries are birds, he may infer that "Tweety is a bird".
- (Transmitting) The father most likely answers his child in English that "Tweety is a bird", but there is some positive probability that he instead answers "Tweety is software" or "Tweety is a man".

For the message "Tweety is a bird", the unit of symbols is English words. Thus, from a non-semantic (syntactic) point of
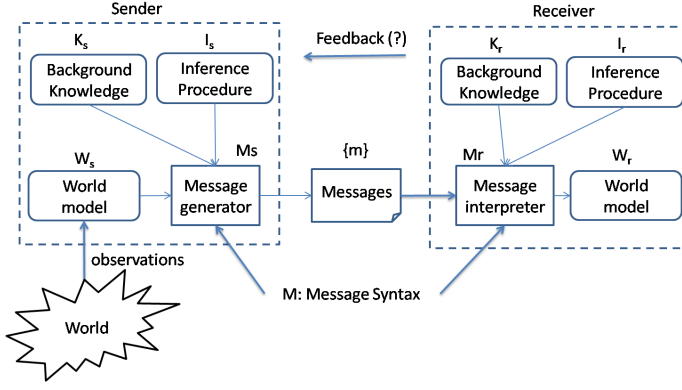
Fig. 2. Semantic Information Source and Destination

view, the message is a sequence of 4 symbols. Its classic information can be approximately determined by the frequencies of English words. Now, we regard this message as a semantic message, e.g., a human friendly coding of the proposition $bird_{Tweety}$. The source states it because the source believes that it is "true" w.r.t. its observations about the world[3]. On the other hand, whether a message is true or not is irrelevant in classical information theory.

Informally, we say a **semantic source** is an entity that can emit messages using a given syntax, such that these messages are "true" in the source, according to its state and inference capabilities.

C. A Semantic Communication Model

What, then, is *semantic communication*? When a semantic information source (e.g., the father in the example above) sends a message, the source expects the destination (e.g., the child) to "understand" the message to some degree. The destination, thus, rather than mechanically decoding the syntax of the message, will be able to draw conclusions from the received message, as well as from its current local knowledge. In the above example, the child, after learning that "Tweety is a bird", may infer that "Tweety is an animal", if her knowledge base tells her that "birds are animals".

Figure 2 characterizes a model of semantic communication we will use in this paper. Formally, a semantic information source is a tuple (W_s, K_s, I_s, M_s), where

- W_s is the model of worlds potentially observable by the source;
- K_s is the background knowledge base of the source;
- I_s is the inference procedure used by the source;
- M_s is the message generator used by the source to encode a message.

In this model, the source builds its own world model by observing the outside world. In the "Tweety" example, the world is observable using a search engine. In this generic model, we do not specify *how* the world is represented, and

[3]It's possible that a source intentionally sends out wrong messages to deceive the destination. However, we believe that such situations should be studied as Level C communication, not as Level B (semantic).

the kind of semantic relations between the world model and the messages. There are several different ways this may be done, e.g., by using model-theoretic semantics, operational semantics, lexical semantics, or by many forms of cognitive models of semantics [5].

The message generator (or semantic encoder) generates messages according to defined strategies. Since usually there are many different but semantically valid ways to describe one situation, the message generator has great freedom in picking a "good" code. For instance, the generator may send messages that are most accurate, or that are easy to generate (according to some cost function), or that the destination is most interested in. Also, similar to the engineering transmitter, the message generator may deal with both how to reduce redundancy in messages (source coding), and how to improve the reliability of the transmission (channel coding).

Possible outputs of the message generator can be seen as an *interface language* for the source. For instance, regarding a graph, one interface language may be the reachability between nodes; another may be minimal distances between nodes.

The generated message will be transmitted over a conventional (i.e., non-semantic) channel, in which a conventional transmitter and a conventional receiver will take care of the engineering coding/decoding tasks.

Analogous to the source, a semantic information destination (receiver) is a tuple (W_r, K_r, I_r, M_r), where

- W_r is the world model of the receiver;
- K_r is the background knowledge base of the receiver;
- I_r is the inference procedure used by the receiver;
- M_r is the message interpreter (semantic decoder).

A *semantic communication error* occurs if the message to be sent is "true" at the source (w.r.t. W_s, K_s and I_s), but the received message is "false" at the destination (w.r.t. W_r, K_r and I_r). The error may be due to losses in source coding, noise in the channel, losses in decoding, or their combinations.

Note that background knowledge and inference procedures may be fully or partially shared by the source and the destination in semantic communication. It is possible for them to use different background knowledge or inference rules, which may lead to different truth evaluations and, hence, semantic mismatches. There may also be feedback channels from the destination to the source. The source, channel and destination all may have memories (e.g., a Markov source), or may be continuous. To simplify discussion, we leave these extensions for future work.

IV. MEASURING SEMANTIC INFORMATION AND SEMANTIC DATA COMPRESSION

Now we discuss the general principles of measuring the amount of semantic information of sources, and the role of semantics in data compression (source coding). A model-theoretic semantics is studied in this and the next section, but we note that this is not the only possible approach in realizing our generic semantic communication model.

A. Entropy of Semantic Messages

In CIT, the entropy of a message is determined by the *statistical* probability of the symbols appearing it. In CSIT, the entropy of a statement is determined by its *logical* probability, i.e., the likelihood of observing a possible world (model) in which this statement is true. To see the difference, for instance, the message "Rex is not a tyrannosaurus" (M1) is less "surprising" than "Rex is not a dog" (M2), not because the word "tyrannosaurus" is more common than "dog", but because the individuals represented by "tyrannosaurus" (now considered extinct) are less common than the individuals represented by "dog". Thus, M1 has less semantic information than M2, even if it may have more Shannon information based on the statistical distribution of English words.

Below, we define *semantic entropy*, following and extending the CSIT approach. For simplicity, as in [4], we restrict our discussion to propositional logic.

We assume that the source has the following properties:

- The world model W_s is a set of *interpretations* with a probability distribution μ. For propositional logic, an interpretation is a set of positive propositions.
- The inference procedure I_s is a satisfiability reasoner for propositional logic.
- The message generator M_s generates messages by some fixed coding strategy, such that if the observed value of the world model is w and it generates a message x, it must be the case that $w \vDash x$ (verified by I_s), where $\vDash$ is the usual propositional satisfaction relation.

We will omit the subscript $_s$ when there is no confusion. Let $H(W)$ be the Shannon entropy of W, i.e.,

$$H(W) = - \sum_{w \in W} \mu(w) \log_2 \mu(w)$$

If the source is a classical source with W as the symbol set, $H(W)$ will be precisely the entropy of the source. We call $H(W)$ the *model entropy* of the semantic source.

For a message (sentence) x, let W_x be the set of its models, i.e., worlds in which x is "true", $W_x = \{w \in W | w \vDash x\}$. Note that, unlike CSIT, which relies on counting models of a sentence, when interpretations have different probabilities, what matters is the total probability of models of the sentence, not the cardinality of the set of models. Then, the logical probability of a message (sentence) x is

$$m(x) = \frac{\mu(W_x)}{\mu(W)} = \frac{\sum\limits_{w \in W, w \vDash x} \mu(w)}{\sum\limits_{w \in W} \mu(w)}$$

Since μ is a probability measure, when W is not constrained by the background knowledge, $\sum\limits_{w \in W} \mu(w) = 1$.

As in CSIT, we define the semantic entropy of x as

$$H_s(x) = -\log_2(m(x))$$

Carnap and Bar-Hillel [4] gave some justifications for using logarithm in their definition. The measurement satisfies some common-sense requirements for measuring semantics. For propositional logic, we observe:

- $H_s(A \wedge B) \geq H_s(A)$
- $H_s(A \vee B) \leq H_s(A)$
- $H_s(A \vdash B) \Rightarrow H_s(A) \geq H_s(B)$
- $H_s(A \vee \neg A) = 0$

B. Conditional Entropy and Background KB

CSIT is concerned with inferring logical probability (thus, semantic information) of a propositional expression when

- There is no background knowledge
- These propositions are independent of each other

In this subsection, we relax these two assumptions. When there is a background knowledge base K, the set of possible worlds will be restricted to the set compatible with K. The semantic entropy of a sentence is represented as a conditional logical probability:

$$m(x|K) = \frac{\sum\limits_{w \in W, w \vDash K, x} \mu(w)}{\sum\limits_{w \in W, w \vDash K} \mu(w)}$$

$$H_s(x|K) = \log_2 m(x|K)$$

For a simple example, suppose[4] $p(A) = p(B) = 0.5$, A, B independent and we have the background knowledge $K = \{A \to B\}$. The truth table is

#	A	B	$A \to B$	probability
1	0	0	1	0.25
2	0	1	1	0.25
3	1	0	0	0.25
4	1	1	1	0.25

Then the universe of possible worlds "shrinks" to the set of truth assignments in which $A \to B$ is true, i.e., cases 1, 2 and 4. Therefore, we now have conditional logical probabilities

$$m(A|K) = 1/3$$

$$m(B|K) = 2/3$$

$$m(A \wedge B|K) = 1/3$$

Logical probabilities are different from a priori statistical probabilities due to the presence of background knowledge. In the new distribution, A and B are no longer logically independent (as $m(A|K)m(B|K) \neq m(A \wedge B|K)$).

Let μ' be the new distribution of the set of models when K is present, that is,

$$\mu'(w) = \frac{\mu(w)}{\sum_{v \in W, v \vDash K} \mu(v)}$$

$$H(W|K) = - \sum_{w \in W, w \vDash K} \mu'(w) \log_2(\mu'(w))$$

---

[4] We always use p to represent statistical probabilities, and m for logical probabilities.

The model entropies of the source in the example without and with the background knowledge are

$$H(W) = -4 * 0.25 \log_2(0.25) = 2$$

$$H(W|K) = -3 * 1/3 \log_2(1/3) = 1.585$$

It seems that the presence of background knowledge *reduces* the informativeness of the source. This is true when the source does not share background knowledge with the destination. However, if the background knowledge is shared, the reduction in semantic entropy means that we can *compress* the source without losing information. In general, with the help of shared background knowledge, we will be able to communicate with shorter messages to achieve the maximal informativeness of the source. In the example above, this means that state descriptions (the most informative messages) need only 1.585 rather than 2 bits to describe. The 21% saving is the contribution of the shared background knowledge in compressing the source.

C. Semantic Source Coding

For a propositional logic with finite n propositions, the size of all possible interpretations (worlds) is finite (2^n). The number of all possible messages (syntactically valid propositional expressions), however, may be infinite if the length of messages is not restricted. Since an interpretation in general cannot uniquely determine messages, a semantic coding strategy is necessary.

For an information source of engineering interest, the number of all possible messages is in general only finite, or is restricted in other ways. The interface language of the source thus only allows a subset of all possible messages. For example, a Twitter post is limited to 140 characters, and a G-rated movie cannot contain scenes unsuitable for children. For a given interface language, a semantic coding strategy needs to achieve two potentially conflicting goals:

- Maximizing expected faithfulness in representing observed worlds;
- Minimizing expected coding length.

Let X be a finite set of allowed messages. A semantic coding strategy is a conditional probabilistic distribution $P(X|W)$. A deterministic coding is a special case of coding, where each $w \in W$ has at most one possible coded message. Given $\mu(W)$ and $P(X|W)$, the distribution of expressed messages $P(X)$ can be determined using

$$P(x) = \sum_w \mu(w) P(x|w)$$

Let us define $H(X)$ as the Shannon entropy of messages X with the distribution $P(X)$, i.e.,

$$H(X) = - \sum_{x \in X} P(x) \log_2 P(x)$$

The following theorem establishes the relation between the model (semantic) entropy and the message (syntactic) entropy of a source:

Theorem 1: $H(X) = H(W) + H(X|W) - H(W|X)$.

Proof sketch: By definitions of entropy and conditional entropy.

Intuitively, $H(X|W)$ measures semantic redundancy of the coding, and $H(W|X)$ measures semantic ambiguity of the coding. The theorem states that message entropy can be larger or smaller than model entropy, depending on whether redundancy or ambiguity is larger.

When $H(X) < H(W)$, there is an information loss ($H(W) - H(X)$). Sometimes, the loss in coding is an intentional and desired compression of the source. For instance, textual description of an image gives only a semantic abstract of the image. A temperature report about a city usually gives only an average value, hiding detailed reports from participating temperature monitoring stations.

We can also view the model entropy of a semantic information source as the maximal expected (message) entropy *per message* without redundancy. Such maximality is reached when the messages are descriptions of the models themselves. In the case of CSIT, this means that a most informative coding will always give the full state description.

D. Use Semantics for Data Compression

Some extensions of CIT exploit *side information*, i.e., receiver's prior knowledge about the sender, to reduce the length of the code. Classical results in this area [19] describe how to achieve optimal coding with respect to the joint entropy of the source and the side information. In semantic communication, shared knowledge and inference procedures may act as a special kind of side information to improve coding efficiency (i.e., compression). On the other hand, unlike in CIT, semantic side information is not represented as distributions, but as logical statements and inference procedures.

With the presence of semantics, some messages may be semantically equivalent to other messages, and if the equivalency is captured by shared knowledge, this can be used to compress the source. For example $a \to (a \wedge b) \vee (b \wedge c)$ can be reformulated as $a \to b$. If a message has many equivalent forms, we can pick a subset of the forms, hence reducing the entropy of the source without a "real" (semantic) loss.

To what extent is semantic compression possible? For a source with a message interface language X and message distribution $P(X)$, let $\overline{X}$ be the smallest subset of X such that

$$\forall x \in X, \exists \overline{x} \in \overline{X} \text{ s.t. } x \leftrightarrow \overline{x}$$

and

$$P'(\overline{x}) = \sum_{x \text{ s.t. } x \leftrightarrow \overline{x}} P(x)$$

For a message x in X, $\overline{x}$ is its unique semantic normal form in $\overline{X}$. The next theorem states the bound for lossless semantic compression.

Theorem 2: For a semantic source with interface language X, there exists a coding strategy to generate a semantically equivalent interface language X' with message entropy $H(X') \geq H(\overline{X})$. No such X' exists with message entropy $H(X') < H(\overline{X})$.

Proof sketch: The existence part is trivial. The non-existence part is shown by the uniqueness of semantic equivalent normal forms.

The difference $H(X) - H(\overline{X})$ can be a large reduction if the redundancy in semantically equivalent messages is large. For example, a formula in full disjunctive normal form with j different clauses and k different propositions has at least $k!j!$ semantically equivalent forms. For propositional logic with n propositions, 2^{2^n} semantic equivalence classes of messages exist. If $2^{2^n} < |X|$ ($|X|$ is the cardinality of the set of messages), the reduction can be significant. For example, suppose our vocabulary allows only connectives $\vee, \wedge$ and n proposition names. A propositional message can be represented with a grammar tree with internal nodes labeled with connectives and leaves being propositions. A grammar tree of depth d (thus, with message length $O(2^{d^2})$) may have $2^{d(d+1)/2}n^{2^d}$ possible variations. Thus, for

$$d \geq \sqrt{(1/2 + 2\log n)^2 + 2^{n+1}}$$

$2^{2^n} < |X|$ is true. This translates into a message length limit of $O(2^{2^{n+1}})$ or larger.

Other semantic data compression strategies may be explored. One possible approach is to reduce the model entropy of a source, e.g., instead of measuring all models, measure only minimal models [14]. When some semantic infidelity is allowed, *lossy* semantic coding strategies may be used based on semantic similarity between messages (e.g., "black"→"dark").

V. SEMANTIC NOISE AND CHANNEL CODING

A. Semantic Noise

For communication over a noisy channel, the received message may contain errors. The noise may be added either at the engineering level or at the semantic level. Below are some examples of semantic infidelity in communication:

- The meaning of a message is changed due to transmission errors, e.g., from "copy machine" to "coffee machine".
- Translation of one natural language into another language where some concepts in the two languages have no precise match;
- The source uses English units, while the receiver understands it using metric units (e.g., during the loss of the Mars Climate Orbiter[5]);

A key difference between engineering communication and semantic communication is how infidelity is handled. Let X be the input of the channel and Y be the output of the channel. In engineering communication, the goal is to minimize the expected difference between X and Y, and a particular mapping $x \rightarrow y$ (x is a value of X and y is a value of Y) is either a match or not. In semantic communication, we are concerned with, instead of syntactic preservation of the message, the semantic similarity between the input and output messages. Also note that not all syntactic errors will lead to

semantic errors. Suppose that the input message is $x_1 \rightarrow x_2$ and the received message is $x_2 \vee \neg x_1$, there is *no* semantic loss. Thus, the semantic effect of noise may be lower than its impact on syntax transmission due to the presence of *semantic redundancy*.

For a source state (interpretation) w, an input message x and an output message y, there are two kinds of semantic errors[6]:

- Unsoundness: the sent message is true but the received message is false, i.e., $w \vDash x$ but $w \nvDash y$
- Incompleteness: the sent message is false but the received message is true, i.e., $w \vDash y$ but $w \nvDash x$

Some communication tasks may tolerate one kind of error (e.g., incompleteness) more than the other. In this paper, since we do not consider lossy source coding, i.e., $w \vDash x$ is always true, our goal is to reduce unsoundness, formally stated as:

$$\max \sum_{w \vDash y} p(w, x, y)$$

where $p(w, x, y)$ is the joint distribution of w, x, y. For a semantic source, $p(w, x, y) = p(y|w, x)p(w, x)$ where

$$p(y|w, x) = p(y|x)$$

since transmission of the message is independent of source coding. Note that $p(y|x)$ is the semantic channel transition distribution.

$$p(w, x) = p(x|w)\mu(w)$$

where $p(x|w)$ is determined by the semantic encoder (message generator), and $\mu(w)$ is the logical distribution of interpretations. Thus, our goal is

$$\max \sum_{w \vDash y} p(y|x)p(x|w)\mu(w)$$

Since $p(y|x)$ is determined by the semantic channel, and $\mu(w)$ is determined by the source, the goal of semantic channel coding thus is to optimize the coding scheme $p(x|w)$, i.e., given an observed world, choose the strategy that can best tolerate noise. For instance, if a voice channel has a high possibility of confusing "p" and "ff", "copy machine" may be received as "coffee machine". Alternatively, assuming that both sides use "Xerox" as a synonym of "copy machine", "Xerox" may reduce the chance of misunderstanding.

Another way to overcome noise is to introduce semantic redundancy into a message. For example, in HTML, an 'img' object (image) may have an 'alt' attribute which gives a textual description of the image and will be shown instead if the image itself is not transmitted. Note that semantic redundancy may not necessarily lead to syntactical redundancy. For example, suppose the topic of communication is weekdays, then the message "Mon∨Tue∨Wed∨Thu∨Fri" can be reformulated as a shorter message "¬Sat∧¬Sun". The two parts of the reformulated message contain semantic redundancy such that if one part is lost in transmission, the received message is still sound (although not semantically equivalent to the original message).

B. Semantic Channel Capacity

Analogous to CIT, a noisy semantic channel has a capacity limit such that a transmission rate can be achieved with arbitrarily small semantic errors within the limit. First, we explain some notations to be used in the theorem.

- $I(X;Y) = H(X) - H(X|Y)$ is the mutual information between X and Y. It represents *syntactical* channel equivocation, which may be a result of technical noise or non-literal semantic transmission.
- $H_{K_s, I_s}(W|X)$ is the equivocation of the semantic encoder, given the sender's local knowledge K_s and inference procedure I_s. Intuitively, a higher $H_{K_s, I_s}(W|X)$ means higher semantic ambiguity in semantic coding.
- $\overline{H_{s;K_r,I_r}(Y)} = -\Sigma_y p(y) H_s(y)$ is the average logical information of received messages, given the receiver's local knowledge K_s and inference procedure I_s. A higher $\overline{H_{s;K_r,I_r}(Y)}$ means stronger ability of the receiver to interpret received messages.

For a simplified model, we assume $K_s = K_r$ and $I_s = I_r$ and omit the subscript. The limit is given in the theorem below:

Theorem 3 (Semantic Channel Coding Theorem): For every discrete memoryless channel, the channel capacity

$$C_s = \sup_{P(X|W)} \{I(X;Y) - H(W|X) + \overline{H_s(Y)}\}$$

has the following property: For any $\epsilon > 0$ and $R < C_s$, there is a block coding strategy such that the maximal probability of semantic error is $< \epsilon$.

The argument of $\sup$ is the semantic coding strategy. A proof sketch is given in the accompanying technical report. The proof uses a strategy similar to that used by Shannon [18] in deriving engineering channel capacity, using the Asymptotic Equipartition Property (AEP).

Semantic channel capacity may be higher or lower than the engineering channel capacity ($\sup\{I(X;Y)\}$), depending on whether $\overline{H_s(Y)}$ or $H(W|X)$ is larger. This implies that using a semantic encoder with low semantic ambiguity and a semantic decoder with strong inference ability and/or a large shared knowledge base, we may achieve high-rate semantic communication using a low-rate engineering channel.

VI. Conclusion

In this paper, we presented some initial results of our investigation into measuring semantic information and semantic coding. We proposed a model-theoretical framework for measuring semantic information in information sources and communication channels.

An interesting result is that the fundamental theorems of classical information theory have semantic counterparts. These theorems reveal the *existence* of some semantic coding algorithms for data compression and reliable communication. However, as in Shannon's paper [18], these theorems do not tell us how to develop optimal coding algorithms. We note that for both source coding and channel coding, bound-achieving algorithms could be computationally difficult. Efficient semantic coding algorithms deserve further investigation.

This paper is intentionally focused on an abstract basic model of semantic communication so that we can focus on the "core" issues. We will extend the framework in future work, e.g., for supporting more expressive languages, lossy data compression, handling semantic mismatches, typical semantic noise patterns, and measuring semantic misinformation.

References

[1] BACCHUS, F. On probability distributions over possible worlds. In *UAI* (1988), pp. 217–226.

[2] BARWISE, J., AND SELIGMAN, J. *Information Flow : The Logic of Distributed Systems*. Cambridge University Press, 1997.

[3] BERNERS-LEE, T., HENDLER, J., AND LASSILA, O. The semantic web. *Scientific American 284*, 5 (May 2001), 34–43.

[4] CARNAP, R., AND BAR-HILLEL, Y. 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] CROFT, W., AND CRUSE, D. A. *Cognitive Linguistics (Cambridge Textbooks in Linguistics)*. Cambridge University Press, Feb. 2004.

[6] D'ALFONSO, S. On quantifying semantic information. *Information 2*, 1 (2011), 61–101.

[7] DEVLIN, K. *Logic and information*. Cambridge University Press, New York, NY, USA, 1991.

[8] FLORIDI, L. Outline of a theory of strongly semantic information. *Minds Mach. 14*, 2 (2004), 197–221.

[9] FLORIDI, L. Philosophical conceptions of information. In Sommaruga [20], pp. 13–53.

[10] JUBA, B., AND SUDAN, M. Universal semantic communication i. In *STOC* (2008), pp. 123–132.

[11] JUBA, B., AND SUDAN, M. Universal semantic communication ii: A theory of goal-oriented communication. *Electronic Colloquium on Computational Complexity (ECCC) 15*, 095 (2008).

[12] KOHLAS, J., AND SCHNEUWLY, C. Information algebra. In Sommaruga [20], pp. 95–127.

[13] LANGEL, J. *Logic and Information, A Unifying Approach to Semantic Information Theory*. Ph.d. dissertation, Universitat Freiburg in der Schweiz, 2009.

[14] LIFSCHITZ, V. Computing circumscription. In *IJCAI* (1985), pp. 121–127.

[15] NAFRIA, J. M. D., AND ALEMANY, F. S., Eds. *What is really information? An interdisciplinary approach* (2009), vol. 7, tripleC.

[16] NILSSON, N. J. Probabilistic logic. *Artif. Intell. 28*, 1 (1986), 71–87.

[17] SELIGMAN, J. Channels: From logic to probability. In Sommaruga [20], pp. 193–233.

[18] SHANNON, C. E. A mathematical theory of communication. *Bell System Technical Journal 27* (1948), 379–423, 625–56.

[19] SLEPIAN, D., AND WOLF, J. Noiseless coding of correlated information sources. *IEEE Transactions on Information Theory* (July 1973), 471–480.

[20] SOMMARUGA, G., Ed. *Formal Theories of Information: From Shannon to Semantic Information Theory and General Concepts of Information [Muenchenwiler Seminar (Switzerland), May 2009]* (2009), vol. 5363 of *Lecture Notes in Computer Science*, Springer.

[21] WEAVER, W. *The Mathematical Theory of Communication*. 1949, ch. Recent Contributions to the Mathematical Theory of Communication.

[22] WILLEMS, F. M., AND KALKER, T. Semantic compaction, transmission, and compression codes. In *Proceedings of International Symposium on Information Theory (ISIT)* (2005), pp. 214–218.