

# A Semantic Information Formula

## Compatible with Shannon and Popper's Theories

Chenguang Lu ( Independent Researcher )  
Email: [survival99@gmail.com](mailto:survival99@gmail.com)

**Abstract:** Semantic Information conveyed by daily language has been researched for many years; yet, we still need a practical formula to measure information of a simple sentence or prediction, such as “There will be heavy rain tomorrow”. For practical purpose, this paper introduces a new formula, Semantic Information Formula (SIF), which is based on L. A. Zadeh’s fuzzy set theory and P. Z. Wang’s random set falling shadow theory. It carries forward C. E. Shannon and K. Popper’s thought. The fuzzy set’s probability defined by Zadeh is treated as the logical probability sought by Popper, and the membership grade is treated as the truth-value of a proposition and also as the posterior logical probability. The classical relative information formula ( $Information = \log(Posterior\ probability / Prior\ probability)$ ) is revised into SIF by replacing the posterior probability with the membership grade and the prior probability with the fuzzy set’s probability. The SIF can be explained as “ $Information = Testing\ severity - Relative\ square\ deviation$ ” and hence can be used as Popper’s information criterion to test scientific theories or propositions. The information measure defined by the SIF also means the spared codeword length as the classical information measure. This paper introduces the set-Bayes’ formula which establishes the relationship between statistical probability and logical probability, derives Fuzzy Information Criterion (FIC) for the optimization of semantic channel, and discusses applications of SIF and FIC in areas such as linguistic communication, prediction, estimation, test, GPS, translation, and fuzzy reasoning. Particularly, through a detailed example of reasoning, it is proved that we can improve semantic channel with proper fuzziness to increase average semantic information to reach its upper limit: Shannon mutual information.

### **Highlights:**

Fuzzy set membership grade is used to express semantic meaning and to extend Bayes’ formula. Revising classical information formula ( $information = \log(posterior\_probability / prior\_probability)$ ) to obtain the semantic information formula.

Generalized Kullback-Leibler formula and generalized mutual information formula have some interesting properties.

Semantic channel (such as that for weather forecast) can be improved to match Shannon’s channel to convey most information.

Semantic information formula and fuzzy information criterion can be used to optimize predictions, estimates, GPS, translation, and fuzzy reasoning.

**Key words:** Shannon, Popper, semantic information, logical probability, factual test, semantic channel

## **1 Introduction**

After Shannon [29] published his famous article titled *The Mathematical Theory of Communication* in 1948, Weaver proposed to research Semantic Information (IS) [34]. Compared to Shannon’s information, semantic information has two features: one is that IS

involves meaning of messages which can be right or wrong or deviant; another is that uncertainty is not only from probability but also from fuzziness of conceptual extension.

To measure semantic information, Bar-Hillel and Carnap [4] proposed using logical probability (LP) to measure the information carried by a proposition. The amount of information is  $Inf(i) = \log m_p(i)$ , where  $i$  is a proposition and  $m_p$  is a logical probability. The paper also provides the relative information formula:  $Inf(j/i) = Inf(i, j) - Inf(i)$  where  $j$  is another proposition. These formulas are very meaningful. The later one is also used in the author's information formula for reasoning. Yet, there are too many restrictions in their methods. For example, 1) the sentences selected must be either true or false without fuzziness; 2) the sentences are always correctly used without wrong statements, lies or deviations; 3) the probability of objective event is not considered and hence the logical probability is only the relative volume of logical space (for example, three predicates can divide the logical space into  $2^3=8$  lattices and hence the minimum of the logical probability of a proposition is  $1/8$ ). In addition, with their formula  $Inf(i) = \log m_p(i)$ , an inconsistent proposition can imply more information than a correct one, which is called Bar-hillel and Carnap Paradox (BCP) by Floridi [10]. Because of these restrictions and the paradox, it is hard to apply their formulas.

Popper as early as 1935 in his book *"The logic of Scientific Discovery"* [24] proposed to use testability, falsifiability, or information as criterion to demarcate and evaluate scientific theories, and affirmed that the testability or information of a proposition was "in inverse ratio to its logical probability" ([24], 96, 269). Later, in his book *"Conjectures and Refutations"* [25], he made more clear statement: *"It characterizes as preferable the theory which tell us more; that is to say, the theory which contains the greater amount of experimental information or content; which is logically stronger; which has greater explanatory and predictive power; and which can therefore be more severely tested by comparing predicted facts with observations. In short, we prefer an interesting, daring, and highly informative theory to a trivial one."* ([25], 294)

Different from Barhillel and Carnap, Popper stressed factual tests for semantic information ([25], 309). Why does Popper use the information criterion? Let's consider an example of weather forecast. One person always predicts tomorrow's weather as the same as today or predicts in very fuzzy wording; while another person often makes daring or unexpected predictions. If judged by ordinary error criterion, the first person has a higher chance of success while the second person has a lower chance. Such type of criterion promotes conservative and fuzzy predictions. like the saying "Mistake-free is a virtue", a criterion used to evaluate human performance. However, this criterion is detrimental to emerging talents. From the perspective of information value, predictions that the ordinary error criterion promotes also have less value. Therefore, Popper proposes that if propositions are easy to be falsified logically, they are in fact able to withstand tests, the more information there is and the more valuable the proposition is; a tautology such as "The sun will or will not rise" contains no information, therefore is not scientifically meaningful. Popper in his book also made a meaningful attempt to establish a function to denote the severity of tests to test propositions or hypotheses ([25], 526). Yet, his formula is still not practical.

In recent decades, many scholars have been researching on IS measure, as summarized by [9], and on the philosophy of information, as summarized by [8, 21], following Barhillel, Carnap, and Popper. The most influential one is Floridi [8,9,10,11]. For IS measure, they took into account the true-values of propositions and the deviations [9,10] between facts and propositions as the author did. Yet, the difference is that the author uses fuzzy propositions with the truth-values  $\in [0, 1]$  which has implied the deviation already. Floridi and the others still use clear-cut propositions with truth-values  $\in \{0, 1\}$  [9, 10, 11] so that it is not easy to define deviation between facts and statements. Floridi's categorization of IS is meaningful, but the formula he

proposed [10, 11] is still not practical as discussed by others [1, 9, 21]. There are also other studies on generalized information or complexity related to IS more or less [12, 26, 1].

The author's research on information theory began from trying to prove that human mechanism of color vision with higher discrimination [14, 15] can receive more information from colors. Later the research is extended to semantic information and general information, such as information from sensations, daily language, weather forecasts, predictions of various indexes, and various meters, not only for measuring information, but also for optimizing communications [16-20]. Different from other researchers of classical information theory, the author uses both statistical probability and logical probability; different from other researchers of semantic information, the author uses both prior and posterior logical probabilities, and considers the fuzziness of language.

This paper intends to, first, introduce a method to revise the classical relative information formula to get the Semantic Information Formula (SIF); second, demonstrate how the SIF is compatible with Popper's theory [25] and Shannon's theory [29]; and then, introduce the SIF's applications including new developments on the optimization of semantic channel, and the information measure for reasoning.

## 2 Research background and methods

### 2.1 Classical relative information formula

Let  $A = \{x_1, x_2, \dots\}$  and  $B = \{y_1, y_2, \dots\}$  denote two sets,  $X$  denote an element of  $A$  and  $Y$  denote an element of  $B$ . According to the classical information theory, the relative information formula [27] is:

$$I(x_i; y_j) = \log \frac{P(x_i | y_j)}{P(x_i)} = \log \frac{1}{P(x_i)} - \log \frac{1}{P(x_i | y_j)} \quad (1)$$

where  $P(x_i)$  is the prior probability of  $x_i$  and  $P(x_i | y_j)$  is the condition probability or posterior probability of  $x_i$  after  $y_j$  occurs. Based on Shannon's coding theory,  $\log(1/P(x_i))$  and  $\log(1/P(x_i | y_j))$  indicate the prior and posterior averagely optimal codeword lengths of  $x_i$  respectively, so information  $I(x_i; y_j)$  means spared average codeword length.

Because of Bayes' formula  $P(y_j | x_i) = P(x_i | y_j)P(y_j)/P(x_i)$ , there is

$$I(x_i; y_j) = \log \frac{P(x_i | y_j)}{P(x_i)} = \log \frac{P(y_j | x_i)}{P(y_j)} \quad (2)$$

This formula is the core of Shannon's mutual information formula. Averaging  $I(x_i; y_j)$ , we can get Shannon's mutual information  $I(X; Y)$ . Yet, Shannon never used this formula. The reason may be that the use of this formula may bring negative information. However, Shannon's formulas of entropy and mutual information only measure mean information, which is always positive. The SIF comes from Eq. (1) because the author thinks that negative information is possible and meaningful to semantic information. This is because if we believe liars, the information will be negative; coding based on nonsense or lies will increase the average codeword length.

### 2.2 Distinguishing three types of LP to avoid BCP

To avoid Barhillel-Carnap Paradox (BCP), we need to distinguish three types of logical probability. Let us use  $X \in A = \{x_1, x_2, \dots\}$  to denote an age. There are three types of LP :

1) Logical probability related to prior factual statistics

For example, the prior logical probability for “A thief ( $X$ ) is about 80 years old” refers to the probability in which this statement is judged true for all possible thieves. This should be a very small number, say, 0.05. We had better say that it is the LP of a predicate or the prior LP of a proposition. It is this LP that can express Popper’s severity of tests.

2) Logical probability related to posterior fact

It refers to the truth-value of a proposition, determined by language usage. Take the proposition “The thief ( $x_i$ ) is about 80 years old” for example. If  $x_i=80$ , then the truth-value of the proposition is 1; If  $x_i=70$ , then the truth-value is about 0.5-0.7; and as the deviation increases, the truth-value decreases. Later we also call it conditional LP or posterior LP.

Note that “logical probability of a proposition” is an unclear expression which has affected the research of semantic information for a long time. Strictly speaking, it means the truth-value of a proposition; but in many cases, such as in Barhillel and Carnap’s studies [4] and Popper’s studies [25], it actually indicates the LP of Predicate  $y_j=y_j(X)$ , or the prior LP of a proposition.

3) Logical probabilities of tautological and contradictory propositions

Each of them has the same prior and posterior logical probability, 1 or 0, regardless of changes in facts, and hence carries no information.

It is clear that as long as we distinguish three types of logical probability, it is easy to avoid BCP. In addition, we also need to distinguish statistical probability and logical probability of a sentence. For example, the probability in which a sentence is selected is statistical probability. The statistical probability is normalized which means that the sum of probabilities of all selected sentences is 1, while logical probabilities of all selected sentences or the truth function of a predicate are not normalized and have the maximum value 1.

### 2.3 Fuzzy sets, proposition’s truth-value and logical probability

In order to explain LP and conditional LP used in the SIF, we need the notions of fuzzy sets and random sets. Zadeh in 1965 [36] proposed the notion of fuzzy sets, using  $m_{A_j}(x_i)$  to denote the membership grade of  $x_i$  in fuzzy subset  $A_j$  of  $A$ . Now we use  $y_j$  to denote a sentence or predicate and let  $A_j$  include all  $x_i$  that makes proposition  $y_j(x_i)$  true, and use  $B = \{y_1, y_2, \dots\}$  to denote the set of selected sentences. Note the difference between propositions and facts:

Proposition:  $y_j(x_i)$  which is equivalent to “ $x_i \in A_j$ ” For example, “*it will rain tomorrow*” is a proposition, and probably wrong since actually it might be clear tomorrow.

Fact:  $x_i \in A_j$  which is equivalent to the judgment: “ $x_i \in A_j$ ” is true.

According to the above definitions, the truth-value of Proposition  $y_j(x_i)$  is membership grade  $m_{A_j}(x_i)$ . We also understand it as the conditional LP of Predicate  $y_j$ , which is the LP of “ $X \in A_j$ ” for a given condition  $X=x_i$ . So, we can write  $T(A_j | x_i) = T(X \in A_j | X = x_i) = m_{A_j}(x_i)$ .

The LP of Predicate  $y_j$  happens to be the fuzzy set probability  $T(A_j)$  defined by Zadeh [37] and is obtained from averaging conditional LP:

$$T(A_j) = \sum_i P(x_i)T(A_j | x_i) \tag{3}$$

We can also say that  $T(A_j)$  is the prior LP of  $y_j$ , and  $T(A_j|x_i)$  is the posterior LP of  $y_j$ . Assuming  $A_j = \{A_1, A_2, \dots\}$ , then  $A_j \in A_j, j=1,2,\dots$ , and the different subsets in  $A_j$  may be joint.

How do we obtain the membership grade function of a fuzzy set? Wang [31] defines it as the falling shadow of a random set (see Fig.1). By this definition we can obtain the membership grade function of a fuzzy set through the statistics of a random set.

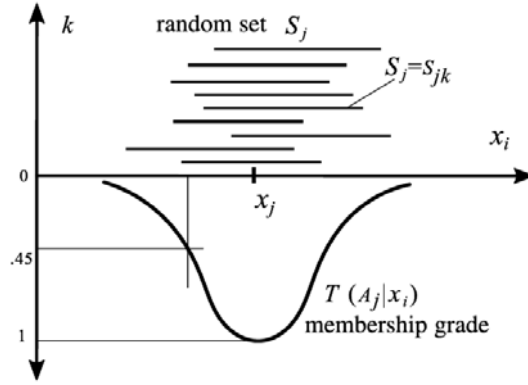


Fig. 1 A membership grade function from a random set falling shadow

For a given predicate  $y_j$ , let many people divide a subset  $s_{jk}, k=1,2,\dots,n$ , of  $A$ , which make  $y_j$  true. When  $n \rightarrow \infty$ , the probability of  $x_i$  in the random set  $S_j$  is

$$T(A_j | x_i) = \frac{1}{n} \sum_k T(s_{jk} | x_i), \quad (4)$$

where  $T(s_{jk} | x_i)$  is the feature function of set  $s_{jk} \in \{0, 1\}$ . We can also treat  $x_j$  as Idea (proposed by Ancient Greek philosopher Plato) of  $A_j$ , and hence membership grade  $m_{A_j}(x_i)$  is similarity degree or confusion probability of  $x_i \in A_j$  with the Idea  $x_j$  of  $A_j$ . The SIF uses not only objective messages' occurring probability as the classical information formulas but also subjective confusion probability.

#### 2.4 The set-Bayes' formula for probability prediction on semantic meaning

In order to obtain the SIF, we bring the feature function of a set or membership grade function of a fuzzy set into Bayes' formula, which is a crucial step. Given the probability distribution of  $X$ ,  $P(x_i), i=1, 2, \dots$ , and knowledge  $X \in A'$ , which is a subset of  $A$ , we can use the feature function  $T(A' | x_i) \in \{0, 1\}$  to replace conditional probability  $P(y_j | x_i)$  to obtain backward conditional probability:

$$P(x_i | A') = P(X = x_i | X \in A') = \frac{T(A' | x_i)P(x_i)}{T(A')} \quad (5)$$

We call this formula as the set-Bayes' formula, illustrated in Fig. 2.

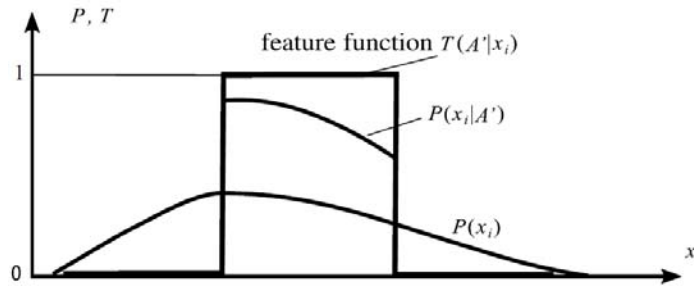


Fig. 2 The illustration of set-Bayes' formula for a set  $A'$

If  $A'$  becomes a fuzzy set  $A_j$ , we have backward conditional probability function ([16], Sec. 3.1) as shown in Fig. 3 and Eq. (6):

$$P(x_i | A_j) = \frac{T(A_j | x_i)P(x_i)}{T(A_j)} \quad (6)$$

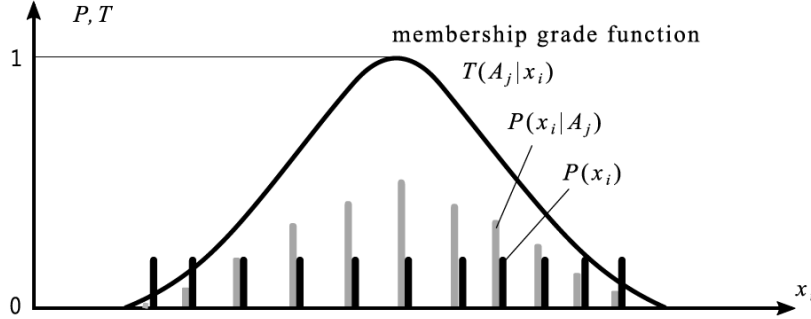


Fig. 3 The illustration of set-Bayes' formula for a fuzzy set  $A_j$

This formula is meaningful in that it establishes the relationship between statistical probability  $P$  and logical probability  $T$ , specifying that probability predictions depend not only on the semantic meaning  $T(A_j|X)$ , but also on the context  $P(X)$ .

### 3. The SIF and its support for Popper's theory

#### 3.1 Revising the classical relative information formula

We replace  $y_j$  in Eq. (1) with  $y_j$  is true, then Eq. (1) becomes:

$$I(x_i; y_j) = \log \frac{P(x_i | A_j)}{P(x_i)} \quad (7)$$

According to Eq. (6) and (7), we obtain the semantic information formula (8) illustrated in Fig. 4:

$$I(x_i; y_j) = \log \frac{T(A_j | x_i)}{T(A_j)} \quad (8)$$

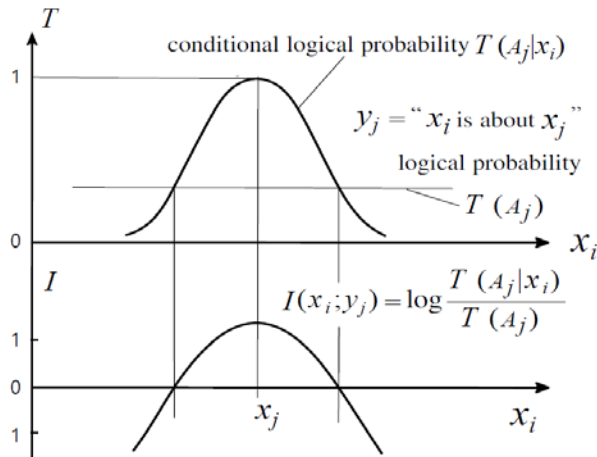


Fig. 4 The illustration of the semantic information formula

Floridi [11] classifies semantic information into factual information and instructional information; factual information is further classified into the true or the untrue, and the untrue into the unintentional (misinformation) and the intentional (disinformation). Apparently, any factual information can be measured by the above formula, because they are the same kind of

information which comes from comparing propositions with facts. The consistent prediction provides positive information while the inconsistent provides negative information. The instructional information proposed by Floridi can also be measured by the same information formula. For example,  $y_j$  = "Set the next electrical pole 50 meters ( $x_j$ ) away". Without the instruction, the distance has a probability distribution  $P(x_i)$ ,  $i=1, 2, \dots$ . With the instruction, the actual distance is  $X=x_j$  which provides the information of  $y_j$  as shown in Eq. (7) and (8). If there is no deviation, that is  $X=x_j$ , the amount of the information is the maximum,  $I=-\log T(A_j)$ . When the deviation reaches to a certain level, the information becomes negative.

### 3.2 How the semantic information measure serves to Popper's information criterion

Let us consider information regarding stock market index predictions, "The index ( $x_i$ ) will be about  $x_j$  at the end of this year". The truth function is probably bell-like distribution surrounding  $x_j$ :

$$T(A_j | x_i) = \exp[-(x_i - x_j)^2 / (2d^2)] \quad (9)$$

where  $d$  is standard deviation. Hence the SIF can be written as

$$I(x_i; y_j) = \log[1 / T(A_j)] - (x_i - x_j)^2 / (2d^2), \quad (10)$$

The above formula means *Information = Testing severity - Relative square deviation*.

If  $A_j$  is not fuzzy, and  $x_i$  is must in  $A_j$ , or  $A_j$  is fuzzy and  $x_i$  has no deviation from the Idea  $x_j$ , then above information measure becomes Barhillel-Carnap information measure [4].

The above semantic information measure has three characteristics:

1. To determine the amount of information, we need to test the prediction by the fact. When the fact is exactly consistent with the prediction, that is  $x_i=x_j$ , the information reaches the maximum. The information decreases as deviation increases. When deviation reaches a certain level, information is negative. This relationship just right manifests ordinary error criterion.

2. The smaller the  $T(A_j)$  is, or the lower the horizontal line in Fig. 3 is, the more information there is if the prediction is correct. This exactly manifests Popper's notion: the smaller the logical probability of a predicate is, the more information there is if it can withstand the test. Under the following two conditions the logical probability becomes small: 1) The coverage of the truth function of the predicate is small, indicating an unusual event. 2) The probability of  $X$  covered by the truth function of a predicate is small, indicating an occasional event. Therefore, smaller logical probability of predicates means the specificity and occasionality of events described by predicates. It is specificity and occasionality that Popper [25] used to explain the severity of tests and his information criterion.

3. Popper affirms that a tautology contains no information, because it is not verifiable or logically non-falsifiable and hence has no scientific value. The above formula reaches the same conclusion: if the truth function of a propositions is always 1 or 0 or a constant, then its average is also the same and information  $I$  is always 0. In this way, it is easy to avoid BCP.

### 3.3 Generalized Kullback-Leibler formula and Popper's falsification

Averaging  $I(x_i; y_j)$  in Eq. (8), we can obtain the mean information of Predicate  $y_j$

$$I(X; y_j) = \sum_i P(x_i | y_j) \log \frac{T(A_j | x_i)}{T(A_j)} \quad (11)$$

where  $P(x_i | y_j)$  is conditional probability from statistics as evidence. Note that it is improper to

put LP on the left of the log because LP is not normalized and hence improper for averaging.

It can be proved that if there is a counter-example  $x_i$  for which  $P(x_i|y_j)>0$ ,  $T(A_j|x_i)=0$ , yet  $T(A_j)>0$ , then the average information is  $-\infty$ . This coincides with Popper's view: one exception is enough to falsify a universal proposition. However, this view is about clear-cut propositions. How do we test probabilistic propositions (such as "The stock market will go up most likely") and fuzzy propositions (such as "There will be a lot of rain this year")? Popper did not offer a proper method. In daily life and social science, most predictions and predicates are probabilistic or fuzzy. The above formula allows a reasonable evaluation of these predicates and avoids negative infinite information.

The above formula can be also written as

$$I(X; y_j) = \sum_i P(x_i | y_j) \log \frac{P(x_i | A_j)}{P(x_i)} \quad (12)$$

where  $P(x_i | A_j)$ ,  $i=1, 2, \dots$ , can be understood as theoretical prediction; the  $P(x_i | y_j)$ ,  $i=1, 2, \dots$ , can be understood as factual test; the  $P(x_i)$ ,  $i=1, 2, \dots$ , can be understood as background knowledge or context. The Eq. (12) is similar to a formula proposed by Theil [32]. It is illustrated in Fig. 5.

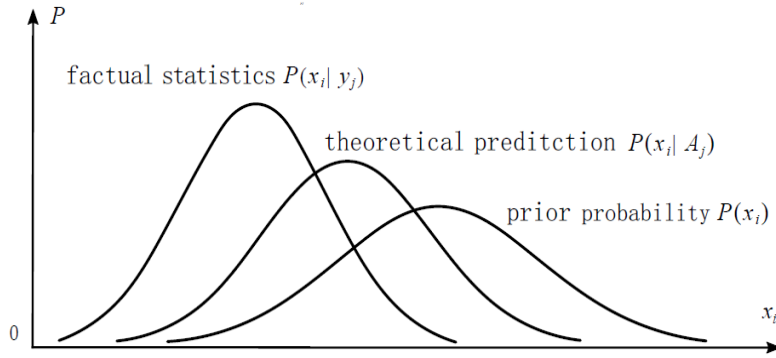


Figure 5 The illustration of the generalized Kullback-Leibler formula

The above formula can be written as the difference of two Kullback-Leibler distances [13]:

$$I(X; y_j) = \sum_i P(x_i | y_j) \log \frac{P(x_i | y_j)}{P(x_i)} - \sum_i P(x_i | y_j) \log \frac{P(x_i | A_j)}{P(x_i | y_j)} \quad (13)$$

Since Kullback-Leibler distance is larger than or equal to 0, when the right part is 0, that is

$$P(x_i | A_j) = P(x_i | y_j), i=1, 2, \dots \quad (14)$$

the amount of information reaches the maximum, equivalent to the Kullback-Leibler information. Therefore, Eq. (11) or (12) can be called as the Generalized Kullback-Leibler Formula (GKLF). It conforms to ordinary error criterion: consistency is good. This formula also shows that the more different the factual statistics is from prior knowledge, the more information there is if the theoretical prediction is correct. This is the very manifestation of Popper's view: the more unexpected a proposition or predicate is and thus the more severe the test is, the more information there is if it can withstand the test.

The GKLF supports such an induction logic: We believe a predicate is due to that more positive examples and less negative examples can increase its average information instead of its logical probability. This induction logic can keep reasonable thought in Carnap and others' induction logic and compatible with Popper's theory.



### 3.4 Semantic mutual information and its significance for optimizing general communications

Averaging  $I(X_i; y_j)$ ,  $j=1, 2, \dots$ , we can get semantic mutual information or Generalized Mutual Information (GMI)  $I(X; Y)$  ([15], Sec. 4.5):

$$\begin{aligned} I(X; Y) &= \sum_j \sum_i P(y_j) P(x_i | y_j) \log \frac{T(A_j | x_i)}{T(A_j)} = H(Y) - H(Y | X) \\ H(Y) &= - \sum_j P(y_j) \log T(A_j) \\ H(Y | X) &= - \sum_j \sum_i P(x_i, y_j) \log T(A_j | x_i) \end{aligned} \tag{15}$$

where  $H(Y)$  is called generalized entropy and  $H(Y|X)$  is called generalized conditional entropy which can also be called fuzzy entropy. The formula of GMI is meaningful to general communications. The classical Rate Distortion Theory [7, 29] deals with the optimization of electronic communication: given a source and average distortion  $D$ , at least how much Shannon Mutual Information SMI, denoted by  $R(D)$ , is needed? In other words, given the source and SMI  $R$ , what is the minimum of average distortion  $D(R)$ ? However, the distortion in the classical information theory is defined subjectively without an objective standard. Now we can use the semantic information  $I(x_i; y_j)$  to replace the distortion  $d(x_i, y_j)$  to evaluate quality of general communications and to reform Rate Distortion Theory into Rate Fidelity Theory ([16], Sec. 5.6). According to Rate Fidelity Theory, we can get some meaningful conclusions. For example, if we believe one's talk randomly, we would be more ignorant about facts and hence would lose some information. If we want to deceive enemies better by lies, we also need a certain amount of SMI, which means that lies according to fact is more harmful than lies according to nothing ([16], Sec. 5.7). For image communication, there is an optimal match between visual discrimination and image resolution because the discrimination of colors or pixels by human eyes is limited, therefore an overly high resolution of images is actually not necessary ([16], Sec. 5.9).

Generalized entropy  $H(Y)$  and GMI  $I(X; Y)$  in Eq. (16) are also meaningful in coding. Suppose  $H(Y)$  is source instead of destination and any  $x_j \in A_j$  can be the destination of  $y_j$ , we change  $P(X|Y)$  to get the minimum  $I^*(X; Y)$  of  $I(X; Y)$ . We define  $R(A_j) = I^*(X; Y)$  which is called Rate Tolerance for given source  $P(Y)$  and  $A_j$  ([16], Sec. 5.5), [19]. It can be proved that if each  $A_j \in A_j$  is a classical set, then  $R(A_j)$  can be written in the form of  $H(Y)$ , and also called Generalized Hartley Measure by others [12, 26]. If each  $A_j \in A_j$  becomes the same size ball around the center  $x_j$  where  $A_j$  is in a multi-dimension space, then  $R(A_j)$  or  $H(Y)$  becomes Complexity Distortion [30]. Fuzzy entropy  $H(Y|X)$  indicates the fuzziness of language. It can be proved that fuzziness decreases information for correct propositions and also reduces information loss for wrong ones. If sets in  $A_j$  are fuzzy, then  $R(A_j)$  can be written in the form of GMI and has some kind of equivalence relation to  $R(D)$  ([16], Sec. 5.5).

## 4 The applications of SIF and Fuzzy Information Criterion

### 4.1 Optimizing semantic channel with Fuzzy Information Criterion (FIC)

Suppose  $A$  is the set of rainfall and sunshine means negative rainfall, then the available forecasts set  $B$  may includes "small rain" (short for "there will be small rain"), "moderate rain", "heavy rain", "torrential rain", "cloud", "sunshine", and also "small or moderate rain", "moderate or heavy rain" and son on. If the rainfall predicted is  $x_i$ , then we select different  $y_j$  to calculate  $I(x_i; y_j)$  by Eq. (8). The  $y_j$  that makes  $I(x_i; y_j)$  reach its maximum is most acceptable. If

the predicted rainfall is in an uneven distribution on time, or in a probability distribution, then we can use the Eq. (11) or (12) to measure average information  $I(X; y_j)$  carried by  $y_j$ . The membership function  $T(A_j|X)$  had better be obtained from statistics of random sets, or from definition as by the following Eq. (19) as a makeshift.

Now we consider another question. For given forecasting rules, i.e. Shannon channel  $P(Y|X)$ , and source  $P(X)$ , can we find a semantic channel  $T(A_j| x_i)$ ,  $i=1, 2, \dots; j=1, 2, \dots$ , denoted by  $T(A_j|X)$ , which can convey semantic information whose amount reaches its upper limit: Shannon mutual information? If audience understand the weather forecast by this channel, they can obtain more information, especially will not obtain  $-\infty$  information when a wrong forecast happens. If an observatory obtains this semantic channel  $T(A_j|X)$  from its past forecasts, it can compares this channel with those sentences' meaning in daily language, and then improves their selecting rule of sentences. For example, if it often makes wrong forecasts, it should select fuzzier sentences. Fortunately, we can solve semantic channel  $T(A_j|X)$  that matches Shannon's channel.

Suppose there are  $M$  kinds of weather in  $A$ ,  $i=1, 2, \dots, M$ , and  $N$  sentences in  $B$ ,  $j=1, 2, \dots, N$ .

$$T(A_j | x_i) = T(A_j)P(y_j | x_i) / P(y_j) \quad \text{Then the semantic channel is a } M*N \text{ matrix. From Eq.}$$

(14), we know that the average information  $I(X; y_j)$  reaches its maximum when

(16)

This is the inverse formula of the set-Bayes formula (6). Let the maximum of  $T(A_j| x_i)$ ,  $i=1, 2, \dots, M$  be 1, we have the optimal semantic channel:

$$T(A_j | x_i) = P(y_j | x_i) / P^*(y_j | x_i), \quad i=1, 2, \dots, M; j=1, 2, \dots, N$$

(17)

where  $P^*(y_j | x_i)$  is the maximum of  $P(y_j | X)$ . We call this criterion for selecting truth function according to Eq. (17) as Fuzzy Information Criterion (FIC).

Philosopher Wittgenstein has a famous remark: The meaning of language lies in how it is used in daily life. The above conclusion shows that understanding semantic meaning (truth function) according to conventional usage (conditional probabilistic distribution) can improve the average information, which supports Wittgenstein's view. When audience or listeners continue to improve their understanding, forecasters or speakers also continue to improve their selecting rules of sentences. Language is evolving in this way.

#### 4.2 To evaluate and optimize fuzzy predictions, estimates, and tests

According to Popper's theory, any scientific statement can be viewed as a hypothesis or prediction. Following predictions include scientific and daily statements, which can be consistent or inconsistent with facts. Hereafter, predictions, estimates, and tests discussed are all fuzzy even if "fuzzy" is not used.

A prediction, estimation, or test  $y_j$  can be defined by its truth function  $T(A_j|X)$ . We select  $y_j$  according to prior probability distribution  $P(X)$  and predicted probability distribution  $P(X|y_j)$ . Let's look at these general predictions "The stock index of this year will increase about 30%-50%", "There is moderate or heavy rain tomorrow", and "The thief is not a young man." The truth functions of these predictions have more than one peak and usually cover a certain range. If there is only one center with possible deviation for a prediction, then the prediction becomes an estimate. For example, GPS, thermometer, watch, and human color perception provide information by estimates. In mathematics an estimate is denoted by  $\hat{x}_j$ , which means  $y_j = \hat{x}_j$  = " $X \approx x_j$ ". In fuzzy estimate, we provide not only  $\hat{x}_j$  but also standard deviation. If the estimated value is one of two, then the estimate becomes test. An example of tests is to make positive or

negative sign in medical examinations. These types of information can be measured by Eq. (8) for a single event  $X=x_i$ , and by Eq. (11) or (12) for average information.

Theoretically, the truth function of a predicate can be obtained by the statistics of a random set; practically, in most cases, it can be defined by some parameters. A general truth function of predictions in one-dimension space can be determined by four parameters,  $d_1$ ,  $c_1$ ,  $c_2$ , and  $d_2(c_1 \leq c_2)$ , shown in Eq. (18) and Fig. 5.

$$T(A_j | x_i) = f(d_1, c_1, c_2, d_2, x_i) = \begin{cases} \exp[-(x_i - c_1)^2 / (2d_1^2)], & x_i < c_1; \\ 1, & c_1 \leq x_i \leq c_2; \\ \exp[-(x_i - c_2)^2 / (2d_2^2)], & x_i > c_2; \end{cases} \quad (18)$$

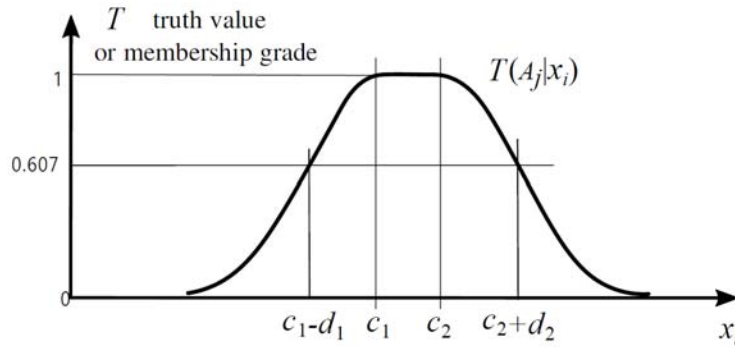


Fig. 5 Using four parameters to construct a truth function for semantic meaning

When  $c_1=0$ , the left part of the function is a horizontal line with height 1. When  $c_2=\infty$ , the right part of the function is a horizontal line with height 1. When  $c_1=c_2$  and  $d_1=d_2$ , the prediction becomes an estimate and the function is in bell-like distribution. Next we focus on the optimization of estimates for which  $c_1=c_2=\hat{x}_j$  and  $d_1=d_2=d$ . The optimization of predictions and tests can follow the same patterns.

If  $d$  is constant when  $y_j$  changes, then we can change only  $\hat{x}_j$  for the maximum  $I^*(X; y_j)$  of  $I(X; y_j)$ . The  $\hat{x}_j$  that makes  $I(X; y_j)=I^*(X; y_j)$  is the optimal estimate. If  $d=d_j$  that is variable with  $\hat{x}_j$ , then a pair of parameters  $\hat{x}_j$  and  $d_j$  that makes  $I(X; y_j)=I^*(X; y_j)$  is the optimal estimate.

According the last section for semantic channel optimization, we can make use of FIC to adopt an approximate method. First we select  $x_i^*$  that makes  $P(y_j | x_i^*)=P^*(y_j | x_j)$  as  $\hat{x}_j$ . Then we select some  $x_i$  to get its  $T(A_j|x_i)$  by Eq. (18). From this  $T(A_j|x_i)$  and Eq. (18), we can get  $d_j$ . Since the  $d_j$  is related “some  $x_i$ ”. It seems feasible to try different “some  $x_i$ ” to get different  $d_j$ , then to select their average. The FIC for estimates should be compatible with or similar to Akaike Information Criterion [2] and Bayesian Information Criterion [6], and also related to Maximum Likelihood Method [3, 22].

In Artificial Intelligence (AI) systems, the information granule is used to simplify statistics and extract information [23, 38]. To obtain more intension or specificity, the coverage of information granule for samples should be as small as possible; however, to reduce omissions and errors, the coverage should be sufficiently large and fuzzy. To balance specificity and coverage to optimize the information granules, we can also use the FIC.

### 4.3 To Optimize GPS and translation: for re-judgment

Theoretically, vehicles’ using GPS (Global Positioning System) to position (tell  $y_j$ ) is based on the distances from vehicles to three GPS satellites (calculated from time differences of

signals). GPS not only gives the most likely position of the vehicle but also error precision. The error precision of GPS is often represented by Root Mean Square Error, or RMS. “RMS=10 meters” means that in 68.2% of chances the vehicle is in the circle within 10-meter radius and with a reading  $x_j$  as the center. That is to say, GPS provides the truth function of  $y_j$  or  $\hat{x}_j$ :

$$T(A_j | x_i) = \exp[-(x_i - x_j)^2 / (2d^2)] \quad (19)$$

where  $d$  is RMS denoting error precision. We can also think that GPS provides similar conditional probability distribution  $P(y_j|x_i)$  with a coefficient for normalization, which does not affect the optimization of re-judgment. However,  $\hat{x}_j$  does not mean the actual position of the vehicle, because if  $y_j = \hat{x}_j$  refers to a position in a pond or on a roof, we can not judge that the vehicle is in the pond or on the roof. We must also consider the prior probability distribution of  $X$  or the larger possibility of the vehicle on roads. Therefore we need to use  $P(x_i)$  and  $T(A_j|x_i)$  or  $P(y_j|x_i)$ ,  $i=1, 2, \dots$ , to re-judge the position of the vehicle.

Assuming the re-estimate is  $\hat{x}_k$ , including only positions on roads, and the truth function of  $\hat{x}_k$  is  $T(A_k | x_i) = \exp[-(x_i - x_k)^2 / (2d'^2)]$  where  $d'$  ( $d' < d$ ) is standard deviation, we can use  $y^j_k$  to denote re-judgment  $y_k = \hat{x}_k$  when the calculated position is  $\hat{x}_j$ . Then, the average information is

$$I(X; y^j_k) = \sum_i P(x_i | A_j) \log \frac{P(x_i | A_k)}{P(x_i)} \quad (20)$$

Change  $y_k$  to maximize  $I(X; y^j_k)$ . The  $y_k$  that makes  $I(X; y^j_k)$  reach its maximum is the optimal re-judgment.

If  $\hat{x}_j$  has systematic deviations from  $x_i$ , we can use FIC to modify its  $T(A_j|X)$  and the positioning. For example, if  $T(A_j | x_i) = \exp[-(x_j - 100 - x_i)^2 / (2d^2)]$  where the systematic deviation is 100, then as  $x_i = x_i^* = x_j - 100$ ,  $P(y_j | x_i^*)$  is the maximum of  $P(y_j | X)$ . So,  $\hat{x}_j$  should be replaced by  $\hat{x}_j - 100$ .

To optimize GPS parameter  $d$ , we can also use the method for searching  $d$  in fuzzy estimate.

The same holds true for language translation. Assume that  $y_j$  is a sentence of source language set, and  $y_k$  is a sentence of target language set. Their truth functions are  $T(A_j|x_i)$  and  $T(A_k|x_i)$ ,  $i=1, 2, \dots$ . We can change  $y_k$  to maximize  $I(X; y^j_k)$  in Eq. (20), the  $y_k$  that makes  $I(X; y^j_k)$  reach its maximum is the most informative translation.

#### 4.4 Testing fuzzy reasoning and verifying FIC

Reasoning can be divided into logical reasoning and experiential reasoning (including scientific reasoning). In logical reasoning, we pay attention only to relation between concepts. Even if the reasoning is correct, for people knowing the reasoning rule, the reasoning is tautology and the information is 0. If it is wrong, the information is negative to people who believe. For example, “If John is Tom’s son’s son, then Tom is John’s grandfather” and “Because  $(a+b)(a-b) = a^2 - b^2$ ,  $(100+1)(100-1) = 100^2 - 1 = 9999$ ” are logical reasoning. Mathematical calculation is also logical reasoning. Certainly, for people who do not know the calculation rule or are inconvenient to calculate, the calculation can offer positive information.

Experiential reasoning or scientific reasoning changes the expected occurrence probability of fact  $X$ . If it is prone to error logically and cannot be falsified factually, then the amount of

information is large; if it is falsified, the information should be negative. We will focus on experiential reasoning and the conclusion should be compatible with logical reasoning.

There are many kinds of fuzzy reasoning which are experiential in daily language, such as old farmers' sayings, "Foggy spring brings rain; foggy summer brings heat; foggy fall brings cool breeze; foggy winter brings snow", "If a man's face looks yellowish, he might be ill". These examples of reasoning cannot be falsified by one counter-example, but can be falsified by statistical data or negative reasoning information.

Assume  $A$  is the set of people, weathers, or their various properties,  $y_1$  and  $y_2$  are fuzzy predicates with truth value functions  $T(A_1|x_i)$  and  $T(A_2|x_i)$ ,  $i=1,2,\dots$ . For the ease of illustration of the reasoning information, we assume  $A$  is one dimensional in Fig. 7.

Suppose two predicates  $y_1$  and  $y_2$  are not correlative, then the membership grade function of the intersection of two fuzzy sets  $T(A_1A_2|x_i)=T(A_1|x_i)T(A_2|x_i)$ . If  $y_1$  and  $y_2$  are strongly correlative, then  $T(A_1A_2|x_i)$  follows Zadeh's definition [36]:

$$T(A_1A_2|x_i)=\min(T(A_1|x_i), T(A_2|x_i)) \quad (21)$$

We also have the membership grade of difference set [14]  $T(A_1\bar{A}_2|x_i)=\max[T(A_1|x_i)-T(A_2|x_i),0]$ , which is used for color model [14] and different from Zadeh's definition [36]. The reason is that if  $A_1$  and  $A_2$  are positively correlative, then  $A_1$  and the [complement](#) of  $A_2$  must be negatively correlated. We calculate the information provided by strongly correlative  $y_1$  and  $y_2$  minus the information provided by single  $y_1$  to get the information provided by  $y_1 \rightarrow y_2$ . The formula is

$$\begin{aligned} I(x_i; y_1 \rightarrow y_2) &= I(x_i; y_1 y_2) - I(x_i; y_1) \\ &= \log \frac{T(A_1 A_2 | x_i)}{T(A_1 A_2)} - \log \frac{T(A_1 | x_i)}{T(A_1)} = \log \frac{T(A_1 A_2 | x_i) / T(A_1 | x_i)}{T(A_1 A_2) / T(A_1)} \end{aligned} \quad (22)$$

Compared with the Barhille and Carnap's formula  $Inf(j/i)=Inf(i,j)-Inf(i)$  [4], this formula needs both prior and posterior LPs of  $y_1 y_2$  (i. e.  $y_1$  and  $y_2$ ) and  $y_1$ ; two predicates  $y_1$  and  $y_2$  are fuzzy. The Eq. (22) is illustrated in Fig. 7.

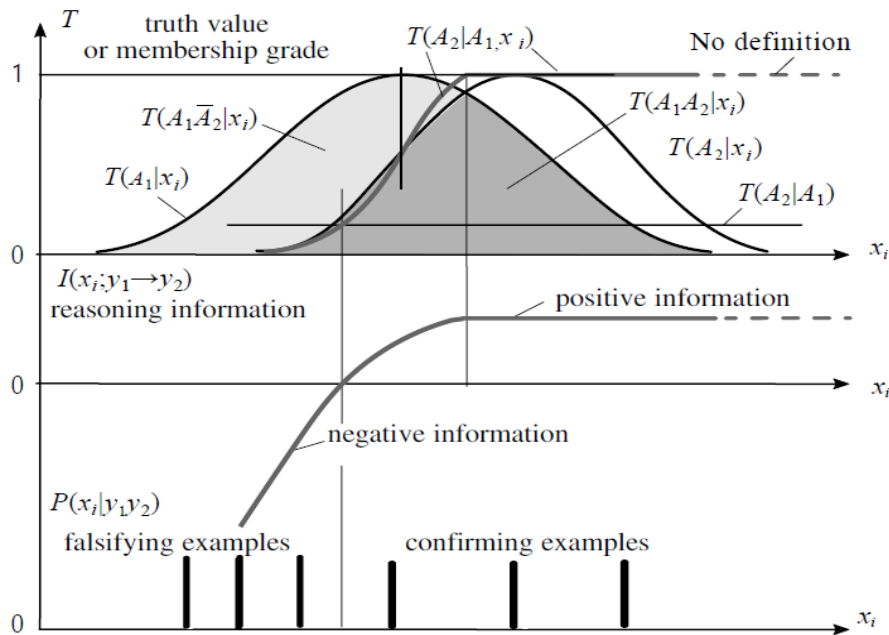


Fig. 7 The illustration of reasoning information

This formula indicates that, according to the prior LP distribution of  $X$ , the larger the  $T(A_1 \bar{A}_2)$  is and the less the  $T(A_2 A_1)$  is, the more unexpected the reasoning is and the easier the falsification is, and hence the more potential information the reasoning can provide. If there is always  $T(A_2 | x_i) \geq T(A_1 | x_i)$  factually, indicating that the reasoning is able to withstand the test, then the reasoning can provide more information. If there is  $T(A_2 | x_i) \geq T(A_1 | x_i)$  prior in any region, then the numerator and the denominator are always 1, and hence the reasoning is tautology and the information is 0.

Now we can use  $T(A_1 A_2 | x_i) / T(A_1 | x_i)$  as the truth-value of the reasoning  $y_1 \rightarrow y_2$ . We also write it as  $T(y_1 y_2 | x_i) / T(y_1 | x_i)$  or  $T(y_2 | y_1, x_i)$  since  $T(y_2 | y_1, x_i) = T(y_2 | (y_1, x_i)) = T(y_2, y_1, x_i) / T(y_1, x_i) = [T(y_1 y_2 | x_i) P(x_i)] / [T(y_1 | x_i) P(x_i)] = T(y_1 y_2 | x_i) / T(y_1 | x_i)$ . So there is

$$T(y_1 \rightarrow y_2 | x_i) = T(y_2 | y_1, x_i) = \min[T(A_1 | x_i), T(A_2 | x_i)] / T(A_1 | x_i) \quad (23)$$

Its prior logical probability is  $T(y_2 | y_1) = T(A_2 A_1) / T(A_1)$ . If  $y_1$  and  $y_2$  is not prior correlated, then  $T(y_2 | y_1) = T(y_2)$ , and Eq. (22) becomes

$$I(x_i; y_2 | y_1) = \log \frac{T(y_2 | y_1, x_i)}{T(y_2)} \quad (24)$$

In order to obtain the average information from the reasoning  $y_1 \rightarrow y_2$  for variable  $X$ , the statistical conditional probability  $P(x_i | y_1 \rightarrow y_2)$ ,  $i=1,2,\dots$  is used for test. There is

$$I(X; y_1 \rightarrow y_2) = \sum_i P(x_i | y_1 \rightarrow y_2) \log \frac{T(y_2 | y_1, x_i)}{T(y_2)} \quad (25)$$

If  $X$  can be divided into counter-examples and positive examples to the reasoning, the average information of  $y_1 \rightarrow Y_2$  is:

$$I(X; y_1 \rightarrow Y_2) = P(X \notin A_2 | y_1) \log \frac{T(y_2' | y_1, X \notin A_2)}{T(y_2')} + P(X \in A_2 | y_1) \log \frac{T(y_2 | y_1, X \in A_2)}{T(y_2)} \quad (26)$$

where the first item is for the counter-examples;  $y_2'$  is the negate of  $y_2$ ;  $Y_2$  is one in  $\{y_2, y_2'\}$ . Similarly, we use  $y_1'$  as the negate of  $y_1$ , and  $Y_1$  as one in  $\{y_1, y_1'\}$ . The average information of  $y_1' \rightarrow Y_2$  is:

$$I(X; y_1' \rightarrow Y_2) = P(X \notin A_2 | y_1') \log \frac{T(y_2' | y_1', X \notin A_2)}{T(y_2')} + P(X \in A_2 | y_1') \log \frac{T(y_2 | y_1', X \in A_2)}{T(y_2)} \quad (27)$$

where the second item is for the counter-examples. Then we have semantic mutual information:

$$I(X; Y_1 \rightarrow Y_2) = P(y_1) I(X; y_1 \rightarrow Y_2) + P(y_1') I(X; y_1' \rightarrow Y_2) \quad (28)$$

Now we calculate the information of the reasoning between triglyceride and fatty liver to illustrate the application of Eq. (26), (27) and (28). The reference [35] provides the data from health examines of 500 people as shown in table 1.

Table 1 The numbers of 4 types of people

numbers	no fatty liver	fatty liver	Total
low triglyceride	3223	427	3650
high triglyceride	647	703	1350
Total	3870	1130	5000

We use  $X$  to denote one of 5000 people,  $y_1$  to denote predicate “ $X$  has high triglyceride”,  $y_2$  to denote “ $X$  has fatty liver”. Now we calculate the average information of  $Y_1 \rightarrow Y_2$  under the framework of Shannon’s theory, shown in Table 2.

Table 2 The conditional probability  $P(Y_2|Y_1)$  and Shannon mutual information

$P(Y_2 Y_1)$	$y_2'$	$y_2$	$P(Y_1)$	$I$ (bit)
$P(Y_2 y_1')$	0.883	0.117	0.73	$I(y_1'; Y_2)=0.057$
$P(Y_2 y_1)$	0.479	0.521	0.27	$I(y_1; Y_2)=0.296$
$P(Y_2)$	0.774	0.226	1	$I(Y_1; Y_2)=0.121$

Now we come back to the discussion under the framework of semantic information. Assume that true-value of the reasoning for counter-examples is 0.2, then the Shannon’s channel becomes the semantic channel, shown in Table 3.

Table 3 The truth value  $T(Y_1 \rightarrow Y_2|X)$ , semantic channel, and average information

$T(Y_1 \rightarrow Y_2 X)$	$X \in A_2'$	$X \in A_2$	$P(Y_1)$	$I$ (bit)
$T(y_1' \rightarrow Y_2 X \in A_1')$	1	0.2	0.73	$I(X; y_1' \rightarrow Y_2)=0.186$
$T(y_1 \rightarrow Y_2 X \in A_1)$	0.2	1	0.27	$I(X; y_1 \rightarrow Y_2)=-0.286$
LP	0.784	0.416		$I(Y_1; Y_2)=0.059$

It is assumed that factual statistics equals to the conditional probability, that is,  $P(X \notin A_2 | y_1) = P(y_2' | y_1) = 0.479$  and  $P(X \in A_2 | y_1) = P(y_2 | y_1) = 0.521$ , According to Eq. (3),  $T(A_2) = P(y_1) + 0.2 * P(y_1') = 0.27 + 0.2 * 0.73 = 0.416$ ;  $T(A_2') = P(y_1') + 0.2 * P(y_1) = 0.73 + 0.2 * 0.27 = 0.784$ . According to Eq. (26), there is  $I(X; y_1 \rightarrow Y_2) = 0.479 \log(0.2/0.784) + 0.622 \log(1/0.416) = -0.286$  bit, According to Eq. (27),  $I(X; y_1' \rightarrow Y_2) = 0.186$  bit. According to Eq. (28),  $I(X; Y_1 \rightarrow Y_2) = 0.059$  bit, which is obviously smaller than the Shannon mutual information  $I(Y_1; Y_2) = 0.121$  bit in Table 2.

Can we adjust the true-values of the counter-examples to improve semantic information? Yes! According to FIC in Sec. 4.1, first we use Eq. (17) to get the optimized LPs:  $T(A_2) = 0.434$  and  $T(A_2') = 0.877$ . Then we use Eq. (18) to get the optimized true-values for counter-examples:  $T(y_2 | y_1', X \notin A_1) = 0.225$ ,  $T(y_2 | y_1, X \in A_1) = 0.543$ . According to Eq. (28), we get  $I(X; Y_1 \rightarrow Y_2) = 0.121$  bit which is equal to the Shannon mutual information  $I(Y_1; Y_2)$  in Table 2.

Table 4 Optimized  $T(Y_1 \rightarrow Y_2|X)$  so that  $I(X; Y_1 \rightarrow Y_2)$  is equal to the Shannon mutual information

$T(Y_1 \rightarrow Y_2 X)$	$X \in A_2'$	$X \in A_2$	$P(Y_1)$	$I$ (bit)
$T(y_1' \rightarrow Y_2 X \in A_1')$	1	0.225	0.73	$I(X; y_1' \rightarrow Y_2)=0.057$
$T(y_1 \rightarrow Y_2 X \in A_1)$	0.542	1	0.27	$I(X; y_1 \rightarrow Y_2)=0.298$
LP	0.877	0.434		$I(Y_1; Y_2)=0.121$

## 5 Discussion and Summary

There seem to be two kinds of information theories: one is classical, with Shannon’s theory at its core and data coding optimization as its purpose, without taking into account semantic meaning; another is the information theory related to semantic meaning. The later is for philosophy, artificial intelligence, and daily communication, and rarely uses Shannon’s formulas. Now we see that the SIF from the classical relative information formula with a slight change can greatly improve explanatory power and bridge two kinds of information theories.

This paper briefly reviews the research history of semantic information and introduces a method to obtain the SIF compatible both with Shannon and Popper’s theories. This paper also explains how the SIF and the FIC can be applied to the fields such as predictions, estimates, tests, AI, GPS, language translation, and fuzzy reasoning.

The semantic similarity measures have been discussed by many researchers [5, 27]. With the SIF, we can obtain the semantic mutual information measure between two predicates, such as the above  $y_1$  and  $y_2$ . This measure should be related to semantic similarity measure for fuzzy concepts, which needs further research. The relationship between the Fuzzy Information Criterion and the Maximum Likelihood Method [3, 31] is also worth to discuss in future. It is possible that SIF and FIC are widely used in AI.

#### **Acknowledgment**

I want to express my gratitude to Professor Pei-Zhuang Wang for his long time support for my research on semantic information, and to Dr. Ya Li and others for their help in preparing and revising this manuscript.

#### **References**

- [1] P. Adriaans, A critical analysis of Floridi's theory of semantic information, *Knowledge, Technology & Policy*, 23(1) and (2)(2010)41-56.
- [2] H. Akaike, A new look at the statistical model identification, *IEEE Transactions on Automatic Control* 19 (6)(1974)716–723
- [3] J. Aldrich, R. A. Fisher and the making of maximum likelihood 1912–1922, *Statistical Science*, 12(3)(1997)162–176.
- [4] Y. Bar-Hillel, R. Carnap. An outline of a theory of semantic information. Tech. Rep. No.247, Research Lab. of Electronics, MIT, 1952.
- [5] M. Batet, S. Harispe, S. Ranwez, D. Sánchez, V. Ranwez, An information theoretic approach to improve semantic similarity assessments across multiple ontologies, *Information Sciences*, 283(1)(2014)197-210.
- [6] K. P. Burnham, D. R. Anderson, ultimodel inference: understanding AIC and BIC in Model Selection, *Sociological Methods & Research*, 33(2004)261–304.
- [7] T. Berger, *Rate Distortion Theory*, Englewood Cliffs, N.J. Prentice-Hall, 1971.
- [8] G. D. Crnkovic, W. Hofkirchner, Floridi's "Open Problems in Philosophy of Information", Ten Years Later, *Information* 2011, 2(2)(2011)327-359
- [9] S. D' Alfonso, On Quantifying Semantic Information, *Information*, 2(1)(2011)61-101
- [10] L. Floridi, Outline of a theory of strongly semantic information. *Minds and Machines*, 14 (2004)197-221.
- [11] L. Floridi, Semantic conceptions of information, in *Stanford Encyclopedia of Philosophy*, <http://seop.illc.uva.nl/entries/information-semantic/>, substantive revision Jan 7, 2015.
- [12] G. Klir, *Uncertainty and Information: Foundations of Generalized Information Theory*, John Wiley, Hoboken, NJ, 499 pp., 2005.
- [13] S. Kullback, R.A. Leibler, On information and sufficiency, *Annals of Mathematical Statistics* , 22 (1)(1951)79–86.
- [14] C. Lu, Decoding model of color vision and verifications, *ACTA OPTIC SINICA*, 9(2)(1989), 158-163.
- [15] C. LU, B-fuzzy set algebra and a generalized cross-information equation, *Fuzzy Systems and Mathematics(in Chinese)* , 1(1991)76-80.
- [16] C. Lu, *A Generalized Information Theory( in Chinese)*, Hefei, China Science and Technology University Press, 1993.
- [17] C. Lu (1994), Meanings of generalized entropy and generalized mutual information for coding, *J. of China Institute of Communication(in Chinese)*, 15(6)(1994)37-44.
- [18] C. Lu, *Entropy Theory of Portfolio and Information Value (in Chinese)*, Hefei, Science and Technology University Press , 1997.
- [19] C. Lu, A generalization of Shannon's information theory, *Int. J. of General Systems*, 28 (6) 1999, 453-490.



- [20] C. Lu, GPS Information and Rate-Tolerance and its Relationships with Rate Distortion and Complexity Distortions(in Chinese), *Journal of Chengdu University Of Information Technology*, 6(2012)27-32.
- [21] J. Mingers, Prefiguring Floridi's theory of semantic information, *TripleC* 11(2)(2013)388-401.
- [22] A. Mohammad-Djafari, Entropy, Information Theory, Information Geometry and Bayesian Inference in Data, Signal and Image Processing and Inverse Problems, *Entropy*, 17(2015)3989-4027.
- [23] W. Pedrycz, New frontiers of computing and reasoning with qualitative information, *International Conference on Oriental Thinking and Fuzzy Logic* (held in Dalian, China, Aug. 17-20, 2015)
- [24] K. Popper, *Logik Der Forschung: Zur Erkenntnistheorie Der Modernen Naturwissenschaft*, Wien: J. Springer, 1935; English translation: *The Logic of Scientific Discovery*, London: Hutchinson, 1959.
- [25] K. Popper, *Conjectures and Refutations*, London and New York, Routledge, 2002.
- [26] M. Poulya, J. Kohlasb, P. Y A Ryanc, Generalized information theory for hints, *International Journal of Approximate Reasoning*, 54(1)(2013)228–251.
- [27] P. Resnik, Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language, *Intelligence Research* 11 (1999) 95-130.
- [28] A. M. Rosie, *Information and communication theory*, New w York, Gordon and Breach, 1966.
- [29] C.E. Shannon, A mathematical theory of communication. *Bell System Technical Journal*, 27 (1948) 379–429, 623–656.
- [30] D. M. Sow, Complexity Distortion Theory, *IEEE Trans on Information Theory*, 49(3)(2003) 604 - 608
- [31] Wang, P. Z., *Fuzzy Sets and Random Sets Shadow* (in Chinese), Beijing Normal University Press, 1985.
- [32] M. A. Tanner, *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*; Springer series in Statistics; Springer: New York, NY, USA, 1996.
- [33] H. Theil, *Economics and information theory*, North-Holland Pub. Co., Amsterdam and Rand McNally, Chicago, 1967.
- [34] W. Weaver, Recent contributions to the mathematical theory of communication. In: *The Mathematical Theory of Communication*, edited by C. E. Shannon and W. Weaver, University of Illinois Press, Urbana, 1949.
- [35] G. Yin, J. Ding, Q. Gong, L. Shi, J. Liu, The relationship of fatty liver with high blood lipids, high blood glucose and high UA, *Modern Medicine Journal of China*, 8(9)(2006)7-8.
- [36] L. A. Zadeh, Fuzzy sets, *Information and Control*, 8(3)(1965)338–353.
- [37] L. A. Zadeh, Probability measures of fuzzy events, *J. of Mathematical, Analysis and Applications*, 23(2)(1986)421-427.
- [38] L. A. Zadeh, Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic, *Fuzzy Sets and Systems*, 90(1997)111-127.

Welcome to comment or publish this paper.