Entropy as a Quality Measure of Correlations between n Information Sources in Multi-agent Systems

G. Enee[1] and J. Collonge[2]
[1]*ISEA - EA 7484, UNC, Campus de Nouville, Noumea, New Caledonia*
[2]*Atout Plus Groupe, Noumea, New Caledonia*

Keywords: Entropy, Multi-agent Systems, Agent Communication Languages.

Abstract: Shanon's entropy has been widely used through different Science fields, as an example, to measure the quantity of information found in a message coming from a source. In real world applications, we need to measure the quality of several crossed information sources. In the specific case of language creation within multi-agent systems, we need to measure the correlation between words and their meanings to evaluate the quality of that language. When sources of information are numerous, we are willing to make correlations between those different*s* sources. Considering those n sources of information are put together in a matrix having n dimensions, we propose in this paper to extend Shanon's entropy to measure information quality in $\mathbb{R}^{2+}$ *and* then in $\mathbb{R}^{n+}$.

1 INTRODUCTION

Entropy introduced by (Shanon C.E. and al., 1949) can be used to measure uncertainty or randomness in a flux coming from a source. The more a source is uncertain, the more it brings novelty and thus the highest is the measure of entropy. On the opposite, the more a source repeats the same pattern, the less it brings new information and the lower is the entropy of such a source. The main focus of our article is to adapt entropy to multiple sources of information. Thus we will first describe the measure itself as it was presented by (Shanon C.E. and al., 1949). Then we will demonstrate how to deal with two sources of information. To enhance interpretation, we propose a new measure of information quality that will be sustained by an example of the emergence of a language within a multi-agent system. In the fourth part, we generalize the entropy measure to n sources of information and finally we conclude our work.

2 ENTROPY TO MEASURE UNCERTAINTY

Information theory and thus entropy has been used in computer science mainly to optimize the transmission of data through a medium of communication. Entropy gives precise bit size to use to transmit a particular serie of data. It also measures the uncertainty in a flux coming from a source giving an evaluation of transmission error. We will focus here on the description of the measure itself applied to information transmission. We will shortly describe the behavior of the measure while data bring uncertainty or not.

2.1 Entropy for T, a Transmitted Message

Let's M be the $1 - dimension$ matrix describing the transmitted message T. The message contains n different values. Each of the n boxes of matrix M is filled with the number of times each symbol of T appears. Since the transmitted message is one information source, the matrix M is mono-dimensional too.

Thus p_i represents the probability of having the i^{th} particular symbol among n others and is calculated as follow:

$$p_i = \frac{M_i}{\sum_j M_j} \tag{2.1}$$

Thanks to p_i, we are now able to measure the quantity of uncertainty in the transmitted message T:

$$H = -\sum_i p_i \times \log_2(p_i) \tag{2.2}$$

Finally that measure brings useful information about the transmitted message:

Enee, G. and Collonge, J.
Entropy as a Quality Measure of Correlations between n Information Sources in Multi-agent Systems.
DOI: 10.5220/0007684802810287
In *Proceedings of the 11th International Conference on Agents and Artificial Intelligence (ICAART 2019)*, pages 281-287
ISBN: 978-989-758-350-6

- the least number of bits needed to transmit that message M upon a perfect medium of communication.
- the uncertainty level in transmitted message T.

In fact, entropy can be easily bounded in order to evaluate the distance of the result to maximal uncertainty or to maximal certainty.

2.2 Bounding Entropy

The measure of entropy is naturally bounded . Thus uncertainty will be maximum when entropy is maximal too. Maximum uncertainty happens when every p_i reaches uniformity i.e. when $\forall i M_i = c$ and thus $p_i = \frac{c}{c \times n} = \frac{1}{n}$. Entropy will then be:

$$\begin{aligned} H_{uniformity} &= -\sum_{i=1}^{n} \frac{1}{n} \times \log_2(\frac{1}{n}) \\ &= -n \times \frac{1}{n} \times \log_2(\frac{1}{n}) \\ &= -(\log_2(1) - \log_2(n)) \\ &= log_2(n) \qquad (2.3) \end{aligned}$$

While uncertainty is maximum when H reaches $H_{uniformity}$ value, certainty will be maximal when only one p_i has a value i.e. when $p_i = 1$ or $equivalently\ if$ $\exists i, M_i = \frac{c}{c} = 1, \forall j \neq i M_j = 0$. That situation describes a message containing a repeated sequence of the same symbol while n different symbols were expected. Thus message exists but finaly brings no information. Entropy will be:
[1]

$$\begin{aligned} H_{certainty} &= -1 \times \log_2(1) - (n-1) \times 0 \times \log_2(0) \\ &= 0 \qquad (2.4) \end{aligned}$$

So H will tend to reach 0 while message brings certainty.

Now that we have described how to measure uncertainty with a one-dimension source of information that is a transmitted message, let's see how we can deal with two sources of information.

3 ENTROPY IN $\setminus^{2+}$

Entropy can be useful when we deal with two different sources of information and we want to demonstrate correlations between those sources. To picture what we are dealing with, we propose here to sustain our demonstration with the formation of a lexicon matrix(MacLennan B.J. and al., 1994) that emerges from agents or group of agents communicating. That matrix, we call M, contains on one hand the word used to communicate and on the other hand, the meaning of the word when it is used. Thus we are willing to show if in a lexicon matrix each word has a unique meaning or not. First let's describe how to adapt entropy mesure to two sources of information.

3.1 Entropy with Two Sources of Information

When we deal with two sources of information, the transmitted information T_1 indicates for each value it can take, the direct correlation with information source T_2 in the matrix M. Thus two dimensional matrix M measures the quantity of correspondence (i.e. correlation) between those two sources of information. Entropy will help us to measure quality of the lexicon, i.e. Level of certainty.

Let's suppose that source of information T_1 produces n different values and that source of information T_2 produces m different values: the matrix M will then be of size (n,m) to capture any correlation between T_1 and T_2.

To measure the reality of a correlation between the two sources of information, we must adapt the calculus of p[2]:

$$p_{ij} = \frac{M_{ij}}{\sum_k M_{kj} + \sum_l M_{il} - M_{ij}} \qquad (3.1)$$

Entropy is evaluated the same way:

$$H = -\sum_{ij} p_{ij} \times \log_2(p_{ij}) \qquad (3.2)$$

From discussion started in 2.2, we can evaluate maximal entropy to occur when every p_{ij} has the same value, i.e. when $\forall i, j M_{ij} = c$. As a consequence, $H_{uniformity}$ will be:

[1] $\lim_{x \to 0} x \times \log_2(x) = \lim_{x \to +\infty} \frac{-\log_2(x)}{x} = 0$

[2] p is the weight of M_{ij} compared to all values in the same column and in the same line.

$$
\begin{aligned}
H_{uniformity} &= -\sum_{ij} \frac{c}{c \times (n+m-1)} \times \log_2\left(\frac{c}{c \times (n+m-1)}\right) \\
&= -(n \times m) \times \frac{1}{n+m-1} \times \log_2\left(\frac{1}{n+m-1}\right) \\
&= -\frac{n \times m}{(n+m-1)} \times (\log_2(1) - \log_2(n+m-1)) \\
&= \frac{n \times m}{(n+m-1)} \log_2(n+m-1) \quad (3.3)
\end{aligned}
$$

Thanks to the choice of p_{ij} we affirm that entropy will be minimal when there is only one positive value of p_{ij} for each row and each column in matrix M. Thus we can only have q such p_{ij} considering that $q = \min(m, n)$. As a consequence $H_{certainty}$ will be:

$$
\begin{aligned}
H_{certainty} &= -q \times 1 \times \log_2(1) - (n \times m - q) \times 0 \times \log_2(0) \\
&= -q \times \log_2(1) \\
&= 0 \quad (3.4)
\end{aligned}
$$

Note that from a technical point of view, the maximal entropy is given by

$$H_{max} = \frac{n \times m}{\exp(1) \times \ln(2)} \quad (3.5)$$

However it occurs if and only if $p_{ij} = \frac{1}{e}$ (all i, j), which in impossible in our setting since the $n \times m$ elements M_{ij} are integers. For the reader's convenience, let us show the real $\frac{n \times m}{\exp(1) \times \ln(2)}$ is our entropy function upper bound.

Let $n, m \in \mathbb{N} \setminus \{0\}$.

Since the $\mathbb{R}$-vector space of matrix of size $n \times m$ and the set $\left(\mathbb{R}_+ \setminus \{0\}\right)^{n \times m}$ are isomorph, the mappig H defined above by $H(P) = -\sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij} \log_2(p_{ij})$ for any matrix $P = (p_{ij}) \in \mathcal{M}_{n,m}(\mathbb{R}_+ \setminus \{0\})$ can be identified to the mapping

$$
\begin{aligned}
\tilde{H} : \left(\mathbb{R}_+ \setminus \{0\}\right)^{n \times m} &\to \mathbb{R}, \\
x &\mapsto -\sum_{k=1}^{n \times m} x_k \frac{\ln(x_k)}{\ln(2)}
\end{aligned}
$$

where $x_k = p_{r,k-(r-1)m}$ for each $k \in \left\{(r-1)m + 1, (r-1)m + 2, \ldots, rm\right\}$ when $r \in \{1, 2, \ldots, n\}$ and $p_{\cdot,\cdot}$ has been introduced above.

This is the reason why we note H instead of $\tilde{H}$ hereafter. We recall that any proper real-valued function which is coercive and strictly concave admits a unique global maximizer.

On one hand, it is clear that the mapping H is proper (the set $\left\{x \in \left(\mathbb{R}_+ \setminus \{0\}\right)^{n \times m} \middle| H(x) > -\infty\right\}$ is non-empty) and coercive ($\lim_{||x|| \to +\infty} H(x) = -\infty$, where $|| \cdot ||$ is an arbitrary norm on $\mathbb{R}^{n \times m}$).

On the other hand, the Hessian matrix $\nabla^2 H$ of H is the $(n \times m) \times (n \times m)$ negative definite matrix whose components are given for all$x \in \left(\mathbb{R}_+ \setminus \{0\}\right)^{n \times m}$ by $\left(\nabla^2 H(x)\right)_{k,k} = -\frac{1}{\ln(2) x_k}$ and $\left(\nabla^2 H(x)\right)_{k,k'} = 0$ for $k \neq k'$, which prove H strictly concave on $\left(\mathbb{R}_+ \setminus \{0\}\right)^{n \times m}$.

To conclude the proof we compute

$$
\begin{aligned}
\nabla H : \left(\mathbb{R}_+ \setminus \{0\}\right)^{n \times m} &\to \left(\mathbb{R}_+ \setminus \{0\}\right)^{n \times m}, \\
x \mapsto \nabla H(x) &= \frac{-1}{\ln(2)} \left(\ln(x_1) + 1, \ldots, \ln(x_{nm}) + 1\right)
\end{aligned}
$$

and we apply Fermat's rule for concave function. We get

$$H(x) \leq \frac{n \times m}{\exp(1) \times \ln(2)} \text{ for all } x \in \left(\mathbb{R}_+ \setminus \{0\}\right)^{n \times m}. \quad (3.6)$$

We can observe[3] that if $(n + m - 1) = e$:

$$
\begin{aligned}
H_{uniformity} &= \frac{n \times m}{(n+m-1)} \log_2(n+m-1) \\
&= \frac{n \times m}{\ln(2)} \times \frac{ln(e)}{e} \\
&= \frac{n \times m}{e \times \ln(2)} \\
&= H_{max}
\end{aligned}
$$

Now that we have bounded entropy measure dealing with two different sources of information, let's show through an example how it can be efficiently used.

3.2 Application to Lexicon Quality Evaluation

As described in the introduction, we will now focus our attention upon an example (Enee and al., 2002) to sustain our demonstration. As a first step to study a language structure, we can fill a lexicon matrix that

[3] Which is impossible since n and m are integers.

Table 1: Matrix of a perfect language.

$Word\,(T_1)\backslash Meaning\,(T_2)$	1	2	3	4	5	6	7	8
1	c	0	0	0	0	0	0	0
2	0	c	0	0	0	0	0	0
3	0	0	c	0	0	0	0	0
4	0	0	0	c	0	0	0	0
5	0	0	0	0	c	0	0	0
6	0	0	0	0	0	c	0	0
7	0	0	0	0	0	0	c	0
8	0	0	0	0	0	0	0	c

will indicate for each word i.e. T_1, their meaning i.e. T_2. Thus, each time a word i is used, we add one in the matrix to the corresponding meaning j: $M_{ij} = M_{ij} + 1$. While original entropy can capture the redondancy of words or of meanings, it won't be able to capture if a language is well shaped. Let's describe a simple lexicon composed with 8 words and 8 meanings. For comprehension matter, the matrix M will be diagonally filled and meanings or words will be symbolized by numbers. Thus a perfect language will have a matrix looking like table 1.

As a purpose of simplification, we consider that we have the same value c for each unique word / meaning correspondance. If we compare the original measure of entropy $H_{origins}$ (cf. equation 2.1) and the new calculus H_{new} (cf. equation 3.1), we will find:

$$\begin{aligned} H_{origins} &= -\sum_{ij} p_{ij} \log_2(p_{ij}) \\ &= -8 \times \frac{c}{8 \times c} \log_2(\frac{c}{8 \times c}) \\ &\quad -56 \times \frac{0}{8 \times c} \log_2(\frac{0}{8 \times c}) \\ &= -\log_2(\frac{1}{8}) \\ &= \log_2(8) \\ &= 3 \times \log_2(2) \\ &= 3 \end{aligned} \tag{3.7}$$

$$\begin{aligned} H_{new} &= -\sum_{ij} p_{ij} \log_2(p_{ij}) \\ &= -8 \times \frac{c}{c} \log_2(\frac{c}{c}) - 56 \times \frac{0}{2 \times c} \log_2(\frac{0}{2 \times c}) \\ &= -8 \log_2(1) \\ &= 0 \end{aligned} \tag{3.8}$$

It is obvious that the original measure is unable to take into account the two dimensional aspect of a lexicon formation. It indicates that matrix contains uncertainty while new measure describes the matrix as perfectly weighted.

There exists another matrix configuration where $H_{origins}$ offers confusing results (see table 2).

The two measures will then be:

$$\begin{aligned} H_{origins} &= -\sum_{ij} p_{ij} \log_2(p_{ij}) \\ &= -8 \times \frac{c}{8 \times c} \log_2(\frac{c}{8 \times c}) \\ &\quad -56 \times \frac{0}{8 \times c} \log_2(\frac{0}{8 \times c}) \\ &= -\log_2(\frac{1}{8}) \\ &= \log_2(8) \\ &= 3 \times \log_2(2) \\ &= 3 \end{aligned} \tag{3.9}$$

$$\begin{aligned} H_{new} &= -\sum_{ij} p_{ij} \log_2(p_{ij}) \\ &= -8 \times \frac{c}{8 \times c} \log_2(\frac{c}{8 \times c}) - 56 \times \frac{0}{c} \log_2(\frac{0}{c}) \\ &= -\log_2(\frac{1}{8}) \\ &= 3 \end{aligned} \tag{3.10}$$

Equation 3.10 and equation 3.9 shows the same results. H_{new} proves this matrix is confusing regarding language understanding, while $H_{origins}$ indicates that this matrix is as confusing as the perfectly weighted one. We conclude that changing p_{ij} calculus in entropy is the key to measure correlations between two sources of information since the new calculus takes into account the two-dimensional aspect of the data.

As entropy is now well understood in $\mathbb{R}^{2+}$, we propose to introduce a new way to measure correlations between different sources of information thanks to entropy.

Table 2: Matrix of a fully confusing language.

$Word\,(T_1) \backslash Meaning\,(T_2)$	1	2	3	4	5	6	7	8
1	c	0	0	0	0	0	0	0
2	c	0	0	0	0	0	0	0
3	c	0	0	0	0	0	0	0
4	c	0	0	0	0	0	0	0
5	c	0	0	0	0	0	0	0
6	c	0	0	0	0	0	0	0
7	c	0	0	0	0	0	0	0
8	c	0	0	0	0	0	0	0

4 INTRODUCING A MEASURE OF QUALITY

The maximal value of entropy is named $H_{uniformity}$ and the minimal value is named $H_{certainty}$. Equation 3.4 proves that $H_{certainty}$ always equals 0 in $\mathbb{R}^{2+}$. We propose to introduce a new simple calculus of the linear distance between $H_{calculated}$ to ideal matrix called dH:

$$\begin{aligned} dH &= \frac{H_{calculated}}{H_{uniformity} - H_{certainty}} \\ &= \frac{H_{calculated}}{H_{uniformity}} \end{aligned} \tag{4.1}$$

Maximal value of dH is therefore normalized to 1since $H_{certainty} \leq H_{calculated} \leq H_{uniformity}$.

We can now evaluate the quality of an entropical matrix by introducing Q_H, a percentage of quality of the matrix for the measured entropy:

$$Q_H = (1 - dH) \times 100 \tag{4.2}$$

Q_H is 0% when $H_{calculated}$ worthes $H_{uniformity}$. By opposition, Q_H is 100% when $H_{calculated}$ is $H_{certainty}$. Thus quality reflects the lack of diversity in the matrix and as a consequence, quality indicates the strength of correlation between information sources.

Results for table 1 and table 2 are respectively:

$$\begin{aligned} Q_H &= (1 - \frac{H_{calculated}}{H_{uniformity}}) \times 100 \\ &= (1 - \frac{0}{\frac{8\times 8}{(8+8-1)}\log_2(8+8-1)}) \times 100 \\ &= 100\% \end{aligned} \tag{4.3}$$

$$\begin{aligned} Q_H &= (1 - \frac{H_{calculated}}{H_{uniformity}}) \times 100 \\ &= (1 - \frac{3}{\frac{8\times 8}{(8+8-1)}\log_2(8+8-1)}) \times 100 \\ &= (1 - \frac{3}{\frac{64}{15}\log_2(15)}) \times 100 \\ &= (1 - \frac{45}{64\log_2(15)}) \times 100 \\ &\approx 82\% \end{aligned} \tag{4.4}$$

We find that perfect lexicon matrix has 100% quality while confusing lexicon matrix has an 82% level of quality.

That last level of quality should awake researcher's curiosity by analysing further more the confusing lexicon matrix. We can observe in the 2 lexicon matrix that all words have the same meaning: they are all synonyms. To measure homonymy and synonymy in a lexicon matrix, we only have to little adapt the p_{ij} calculus in H:

$p_{ij} = \frac{M_{ij}}{\sum_k M_{ik}}$ for synonymy

$p_{ij} = \frac{M_{ij}}{\sum_l M_{lj}}$ for homonymy

Studying synonymy or homonymy is about to study each dimension of the matrix separately.

We propose to use conversly Q_H as a level of noise L_H. Using the fitted p_{ij}, L_H will be:

$$
\begin{aligned}
L_{H_{synonymy}} &= dH \times 100 \\
&= \left(\frac{-8 \times \frac{c}{8\times c}\log_2(\frac{c}{8\times c})}{8 \times \log_2(8)}\right) \times 100 \\
&= \left(\frac{-\log_2(\frac{1}{8})}{8 \times 3 \times \log_2(2)}\right) \times 100 \\
&= \left(\frac{3 \times log_2(2)}{24}\right) \times 100 \\
&= \frac{1}{8} \times 100 \\
&= 12,5\% \qquad (4.5)
\end{aligned}
$$

$$
\begin{aligned}
L_{H_{homonymy}} &= dH \times 100 \\
&= \left(\frac{-8 \times \frac{C}{C}\log_2(\frac{c}{c})}{8 \times \log_2(8)}\right) \times 100 \\
&= \left(\frac{0}{24}\right) \times 100 \\
&= 0\% \qquad (4.6)
\end{aligned}
$$

Analyzing L_H measure reveals that it is perfectly corresponding to the matrix as there is one column filled with noisy values: synonyms. It is $\frac{1}{8}$ of the whole matrix or $12,5\%$. On the other hand the level of homonymy is 0% as it should be while the matrix contains no single word having different meanings.

Q_H and L_H offers two ways to analyze correlations matrix containg different sources of information. We propose now to generalize our work to n sources of information thus to matrix in $\mathbb{R}^{n+}$.

5 ENTROPY IN $\setminus^{n+}$

While working with n different sources of information to correlate, entropy will thus be extracted from a n dimensional matrix. The p calculus will then modified as follow:

$$
p_{a_1 \dots a_n} = \frac{M_{a_1 \dots a_n}}{(\sum_{j=1}^{n}\sum_{k=1}^{s_j} M_{a_1 \dots k \dots a_n}) - (n-1) \times M_{a_1 \dots a_n}} \qquad (5.1)
$$

Where s_j is the size of the j^{th} dimension of the matrix and a_i is the is the index in the matrix of the i^{th} dimension. Entropy calculus will remain the same:

$$
H_{calculated} = -\sum_{a_1 \dots a_n} p_{a_1 \dots a_n} \times \log_2(p_{a_1 \dots a_n}) \qquad (5.2)
$$

Maximum entropy is reached while $\forall a_1, \dots, a_n \ M_{a_1,\dots,a_n} = c$. For calculus simplification matter, we assert that a matrix filled with the same value $c \varepsilon \mathbb{R}$ brings the same information quality as a matrix filled with 1 i.e. every single value divided by c.

$$
\begin{aligned}
H_{uniformity} &= -\sum_{a_1,\dots,a_n} \frac{1}{a_1 + \dots + a_n - 1} \\
&\quad \times \log_2\left(\frac{1}{a_1 + \dots + a_n - 1}\right) \\
&= -(a_1 \times \dots \times a_n) \times \frac{1}{a_1 + \dots + a_n - 1} \\
&\quad \times \log_2\left(\frac{1}{a_1 + \dots + a_n - 1}\right) \\
&= -\frac{(a_1 \times \dots \times a_n)}{a_1 + \dots + a_n - 1} \\
&\quad \times \log_2\left(\frac{1}{a_1 + \dots + a_n - 1}\right) \\
&= -\frac{(a_1 \times \dots \times a_n)}{a_1 + \dots + a_n - 1} \\
&\quad \times (\log_2(1) - \log_2(a_1 + \dots + a_n - 1)) \\
&= \frac{(a_1 \times \dots \times a_n)}{a_1 + \dots + a_n - 1} \\
&\quad \times \log_2(a_1 + \dots + a_n - 1) \qquad (5.3)
\end{aligned}
$$

Entropy reaches its minimum while the biggest identity square matrix would be represented in the whole matrix i.e. when we have $p_{a_1 \dots a_n} = 1$ for each unique "a" position.

If we consider q as $min(a_1, \dots, a_n)$[4], the calculus of minimal entropy becomes:

$$
\begin{aligned}
H_{certainty} &= -q \times 1 \times \log_2(1) \\
&\quad -(a_1 + \times \dots \times a_n - q) \times 0 \times \log_2(0) \\
&= -q \times \log_2(1) \\
&= 0 \qquad (5.4)
\end{aligned}
$$

The dH measure calculus remains the same as the calculus of Q_H and L_H. L_H would be extended to find out why the Q_H does not reach 100%. The principle is still the same to adapt L_H to n dimensional matrix: fix one or more column in the $p_{a_1 \dots a_n}$ variable and then give a meaning to the L_H calculus like we did in 4.

This last assertion concludes our work.

6 CONCLUSION AND FURTHER WORK

Entropy has been used for decades in computer science but not only. While it offers clear evaluation of

[4] in order to get the biggest square identity matrix

the quality of the transmission of an information, until now, it was not used to correlate different sources of information in a simple way. That modified entropy offers clear and efficient measure to correlate interactions between agents and multi-agent systems.

Next step is to implement an algorithm to make the calculus of H in n dimensional matrix with a reasonable complexity.

REFERENCES

Shannon, C.E. and W. Weaver (1949). *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, Ill.

MacLennan, B.J. and Burghardt, G.M. (1994). Synthetic ethology and the evolution of cooperative communication. Adaptive Behavior, 2(2), Fall 1993, pp. 161-188. MIT Press.

Enee, G. and Escazut, C. (2002). A Minimal Model of Communication for a Multi-Agent Classifier System. Advances in Learning Classifier Systems. LNAI 2321 (Lecture Notes in Artificial Intelligence), Pier Luca Lanzi, Wolfgang Stolzmann, Stewart W. Wilson (Eds.). Springer-Verlag Berlin Heidelberg 2002.