

Crosswords and Information Theory

Peter Andreasen

December 17, 2000

Abstract

A gentle introduction to the wonders of the information theoretical concept of entropy through elementary calculation of the number of crosswords.

What is a crossword, really?

Most people have solved a crossword puzzle or played Scrabble. The existence of word-games like those are not to be taken for granted, though. As we are going to see, the existence of crosswords is entirely at the mercy of the underlying language. In fact, there is a connection between the information theoretic concept “entropy”, and the possibility of creating crossword puzzles!

Before we proceed, we need to get a few definitions straight. First, what *is* a crossword? Let us take a look at one:

g	e	m	□
a	r	e	□
m	a	t	h
e	□	□	e

What we see are rows and columns of words (single letters are accepted as words) separated by white squares.¹ Now, the words in a crossword need not be English as in the example above. We might want to create a Danish crossword or we might even want to have the columns and rows be quotes from Shakespeare’s sonnets. To be able to handle such complex rules for the creation of crosswords we make the following definition.

¹It is in the white squares you will normally find the hints needed to solve the puzzle – and in most crosswords the topmost row and the leftmost column are filled with these hints. For simplicity we will make no assumptions about the placement of the white squares.

Definition 1 A language L is a set of sequences of letters from an alphabet A (say, the letters ‘a’ through ‘z’ and the symbol ‘□’). A crossword of size n is a matrix with the dimensions $n \times n$ where all of the rows are sequences (of length n) from L and all of the columns are sequences (of length n) from L .

So if we want to make a really sophisticated crossword, we may let L be all the possible quotes from Shakespeare. In that case we should use an alphabet which included the letters as well as space and the various punctuation symbols. If we wanted to make a ‘classic’ crossword, we would have L be equal to any sequence you can make by taking words from a dictionary and gluing them together with one or more □’s inbetween. In this case the alphabet A would just be the letters and the special symbol □.

How many are there?

It is obvious, that very small crosswords are easily constructed. Especially crosswords which have only one row or one column. It is also easy to create a few very big but very dull crosswords: If you keep alternating the rows between ‘I□I□I...’ and ‘□I□I□...’, you certainly get a valid (and as big as you like) ‘Classic English’ crossword. So we want to not only consider the existence of big crosswords, but also check if there are many different of them. We are going to calculate the number of big crosswords now.

Assume we have chosen an alphabet A and a language L over which the crosswords must be made. We use the notation $|A|$ for the number of letters and symbols in the alphabet. Let us introduce the following number as well,

$$c_L(n) = \text{number of sequences from } L \text{ of length } n.$$

So for constructing a square crossword of size $n \times n$ over the language L , there are $c_L(n)$ possible choices for the first row. We will now use a small trick and for a moment employ a bit of probability theory: If we picked an absolutely random sequence of n letters from A , what are the chance that we got a 'valid' row, that is, a sequence from L ? The answer is

$$\frac{c_L(n)}{|A|^n}$$

because there are $c_L(n)$ valid sequences and $|A|^n$ possible sequences. An example: suppose we wanted to create a normal English crossword. In my dictionary, there are 1 word (namely 'I') of length 1 and 49 words of length 2. Valid sequences of length 2 are ' $\square\square$ ', ' $\square I$ ', ' $I\square$ ', and then the 49 words of length 2. A total of 52 sequences, that is, $c_L(2) = 52$. The total number of possible sequences of length 2 is $|A|^2 = 27 \times 27 = 729$ (note, that even though we only have 26 letters, the size of A is 27, because we need the symbol \square as well). And thus the probability of getting a valid sequence of length 2 would be $52/729 \sim 0.07$, in this example.

However, that was the probability of just *one* valid row. What about the rest? The probability of all n rows being valid equals the above probability multiplied with itself n times²,

$$\left(\frac{c_L(n)}{|A|^n}\right)^n = \frac{c_L(n)^n}{|A|^{n^2}}.$$

Now for the columns the situation is identical. And because the columns are as high as the rows are wide, the result is the same: The probability of all n columns being valid (that is, from L) equals

$$\frac{c_L(n)^n}{|A|^{n^2}}.$$

Now we may calculate the probability of a randomly selected matrix of $n \times n$ letters from A being in fact a crossword: We want *both* its rows *and* its columns to be valid, so we multiply:

$$\left(\frac{c_L(n)^n}{|A|^{n^2}}\right)^2 = \frac{c_L(n)^{2n}}{|A|^{2(n^2)}}$$

²This is basic probability theory. It is comparable to when we say that the probability of a coin landing heads up equals $\frac{1}{2}$ and then proceed to calculate the probability of two heads in a row as $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$. We multiply the probabilities when we want the probability of *both* events.

We now return to our original question: How many (big) crosswords are there? Well, we know the probability of a randomly selected matrix of $n \times n$ letters being a crossword, and there are a total of $|A|^{n \times n} = |A|^{n^2}$ possible $n \times n$ matrices, so we may write³

$$N_n = |A|^{n^2} \times \frac{c_L(n)^{2n}}{|A|^{2(n^2)}} = \frac{c_L(n)^{2n}}{|A|^{n^2}}.$$

This makes N_n our symbol for the number of crosswords of size $n \times n$.

Explosive numbers

To get to the core of the matter, we need to do a bit of mathematical wizardry, so now is the time to wear your pointed hat! First, we apply the logarithm⁴ to N_n :

$$\begin{aligned} \log N_n &= 2n \log c_L(n) - n^2 \log |A| \\ &= 2n^2 \left(\frac{\log c_L(n)}{n} - \frac{\log |A|}{2} \right) \\ &= \frac{2n^2}{\log |A|} \left(\frac{\log_{|A|} c_L(n)}{n} - \frac{1}{2} \right) \end{aligned}$$

The special symbol $\log_{|A|}$ is simply the logarithm to the base of $|A|$, that is, $|A|^{\log_{|A|} x} = x$. Recall that we are interested in the number N_n when n grows large. In the expression above, the first fraction,

$$\frac{2n^2}{\log |A|},$$

just grows towards infinity as n does the same. The second fraction,

$$\alpha_n = \frac{\log_{|A|} c_L(n)}{n}$$

is more interesting (so we name it α_n). The value of $c_L(n)$ must be between 0 and $|A|^n$ (that should be clear from the definition of $c_L(n)$). So (assuming

³This is another application of basic probability theory: The number of valid crosswords are calculated as the *probability* of a random matrix being valid *times* the total number of possible matrices.

⁴Recall, that taking the logarithm of a product yields a sum ($\log ab = \log a + \log b$), a logarithm of a fraction yields a difference ($\log a/b = \log a - \log b$) and the logarithm of a power turns into a product ($\log a^b = b \log a$).

$c_L(n) > 0$) we see that $\log_{|A|} c_L(n)$ is between 1 and n . Thus, when n grows large, value of α_n stays between 0 and 1. Let us assume that α_n in fact *converges*⁵ to some number α between 0 and 1. We may now conclude, that if $\alpha < \frac{1}{2}$ the value of $\log N_n$ goes towards negative infinity ($-\infty$) as n becomes big. To see this very clearly, consider the formula from above (this time written using the symbol α_n , but otherwise identical):

$$\log N_n = \frac{2n^2}{\log |A|} \left(\alpha_n - \frac{1}{2} \right).$$

Assuming $\alpha < \frac{1}{2}$ then α_n will also be less than $\frac{1}{2}$ (provided n is big enough) and hence we find that the value of $(\alpha_n - 1/2)$ becomes negative, while the first fraction as mentioned above grows towards infinity.

If, on the other hand, $\alpha > \frac{1}{2}$, the value of $\log N_n$ approaches positive infinity (∞), provided we make n big enough.

Now take off your pointed hat (the heavy math is over for now!) and consider the implications of α for N_n . If $\log N_n$ is growing unlimited when n grows, then it indicates that N_n is very big. If $\log N_n$ decreases to very large negative numbers when n grows, then N_n must be very close to zero⁶. Somehow the number α determines if the number of valid crosswords (N_n) gets very large as n (the size of the crosswords) gets large, *or* if the number of valid crosswords becomes almost zero!

It is rather clear, that the value α depends on L and only on L . Thus we will write α_L and say that the language L has α -value α_L .

For languages L made up from an English dictionary based on 26 letters plus \square , one find the value of α_L to be around 0.6, in other words, there are no imminent shortage of crosswords.

Curiouser and curiouser!

As if the madness will see no end, we now approach 3-dimensional crosswords, no, let us make it d -dimensional crosswords! The size of the crosswords is now measured by n^d where n is the generalized notion of height or width and d is the *dimension*

⁵This is by no means a trivial assumption. It is, however, beyond the scope of this article to look into these details.

⁶This is all due to the behaviour of the logarithm function.

(which was 2 before). That is, we consider cubic (for $d = 3$) or even hypercubic (for $d > 3$) crosswords. The probability of one dimension (think: row) of the crosswords being valid equals

$$\left(\frac{c_L(n)}{|A|^n} \right)^{n^{d-1}} = \frac{c_L(n)^{n^{d-1}}}{|A|^{n^d}}.$$

This is almost the same result as before, but note the exponent n^{d-1} . In the case $d = 3$, where we might imagine the crossword as a cube made up of 'sticks' of sequences from L , the exponent corresponds to the fact that in each dimension there are n^2 sticks. The probability of all dimensions (think: rows and columns) being valid equals

$$\left(\frac{c_L(n)^{n^{d-1}}}{|A|^{n^d}} \right)^d = \frac{c_L(n)^{dn^{d-1}}}{|A|^{dn^d}}.$$

Again, this should come as no shock. The total number of possible crosswords (think: any matrix) is multiplied with the probability and we find:

$$N_n^{(d)} = |A|^{n^d} \frac{c_L(n)^{dn^{d-1}}}{|A|^{dn^d}} = \frac{c_L(n)^{dn^{d-1}}}{|A|^{n^d(d-1)}}.$$

Applying logarithm yields

$$\log N_n^{(d)} = dn^{d-1} \log c_L(n) - n^d(d-1) \log |A|$$

and reorganizing the terms,

$$\log N_n^{(d)} = \frac{dn^d}{\log |A|} \left(\frac{\log_{|A|} c_L(n)}{n} - \frac{d-1}{d} \right). \quad (1)$$

The second fraction in the above expression is recognized from before. We recall that the number α is used to denote the limiting value of the fraction as n becomes very big. We find, that if, say, $d = 3$ the value of α must be at least $\frac{2}{3}$ if we want to have many, big crosswords. As the dimension of the crosswords grow, the languages L must have larger and larger α -value to sustain the notion of many crosswords.

It seems like α_L expresses something fundamental about the language L . So information theorists have a name for that value:

Definition 2 *Let L be a language. The entropy of L is defined as*

$$\tilde{H}(L) = \lim_{n \rightarrow \infty} \frac{\log_{|A|} c_L(n)}{n}.$$

We recognize the entropy as the same thing as we know as α_L . The little symbol above \tilde{H} is there to remind us that this is a special kind of entropy: The theory leading up this definition is not as concise and rigid as many information theorists would want. But we should not feel nothing has been accomplished: Our entropy captures some very deep aspects of the concept.

We may wonder what happens if, say, $\tilde{H}(L) = \frac{1}{4}$. What kind of crosswords are possible? Why, crosswords of dimension $d > \frac{4}{3}$ of course! If $d = \frac{4}{3}$ we have $(d-1)/d = \frac{1}{4}$. How to visualize a crossword in 1.333 dimensions is probably better left as an exercise to the reader!

A recapitulation

Let us briefly examine what we have learnt so far: We introduced the concept of language which is nothing but a set of sequences of letters. We have then made a satisfactory definition of what is a crossword over a language. Using elementary combinatorics and probability theory we have calculated the number of valid crosswords of size $n \times n$ (or, in the case of other dimensions, size n^d). This number depends on the size of the alphabet, $|A|$, as well as the special function $c_L(n)$. We then observed, that there are essentially two different cases: In the first case ($\alpha_L < \frac{d-1}{d}$), the number of crosswords becomes very small as the size, n , grows. In the second case ($\alpha_L > \frac{d-1}{d}$), the same number becomes infinitely big as the size grows.

This calls for a reformulation of our initial question: While we opened this paper asking about the existence of crosswords, we are now tempted to ask: "Given a language L , what is the greatest dimension d for which there are many (big) crosswords?". This move encourages us to consider non-integer values of d , and thus we have definitely left the realm of ordinary crossword puzzles, and maybe the real world as well! The answer to the new question is related to the entropy as we have just seen. In fact, combining the definition of entropy and formula (1), shows

$$\tilde{H}(L) = \frac{d_0 - 1}{d_0} \quad \text{and} \quad d_0 = \frac{1}{1 - \tilde{H}(L)},$$

where d_0 is exactly the largest dimension where it is possible to create (many) crosswords over L .

Note how d_0 may be arbitrarily big, even ∞ if the entropy equals 1. How is that for a crossword puzzle! Actually, if $\tilde{H}(L) = 1$ it is quite trivial to create crossword puzzles (in any dimension). An example of such a language L is that which is made up of every integer. The alphabet A is just the digits, and $c_L(n) = |A|^n = 10^n$ (because any sequence of length n which is made up from digits, is a valid number) so clearly $\tilde{H}(L) = 1$.

We have arrived at the concept of entropy by a quite unusual method. Aside from (hopefully) some pedagogical advantages there are other reasons for picking this approach: We now have an entropy concept defined on any *language* or, which is the same, any set of sequences made up of letters from A . This is not true for the traditional entropy which is introduced by the concept of information sources (which are also known as stochastic processes, and are based on a quite technical probability theoretic framework). In addition, while our entropy, in its current form, does not handle random languages (e.g. the language of all possible sequences of 0's and 1's created by flipping a coin), it is possible to refine our definitions to cover (and indeed generalize the probability theoretic entropy) these important cases as well.

Entropy

Something should probably be said about why entropy is so central to information theory. But where to start and where to end! We will look at only two aspects, one of somewhat philosophical nature and the other of very practical nature.

Entropy is often said to be a measure of how 'complex' or even 'caotic' things are. This corresponds nicely with the observations given above: A language made up of every possible integer is devoid of any form or structure. Anything goes. It is impossible to distinguish between the a sequence from the language and a sequence of completely random digits. This language, as explained above, has the entropy 1. On the other hand, a language made up of sequences of only one letter, say 'a', is completely structured. No room for choices. The function $c_L(n)$ is constantly 1 regardless of the value of n . This corresponds to the case where $\tilde{H}(L) = 0$.

The other use of entropy which we will touch upon is *data compression*. We mention the following

theorem in sketch form:

Theorem 1 (SHANNON) *Let L be a language over A . There exists an encoding such that any sequence $x \in L$ of length n may be encoded into a sequence of no more than $n(\tilde{H}(L) + \epsilon)$ letters. This holds for any positive number ϵ however small, provided the length n of x is large enough.*

This formulation only captures the essence of Shannon's theorem. What is important is the order in which things happen: First we choose the value ϵ as small as we want it. This determines how "close" to the entropy we want our encoding to be. Then the theorem tells us that there exists a number N and a code so that any sequence $x \in L$ which are at least N letters long can be encoded into just $|x|(\tilde{H}(L) + \epsilon)$ letters. So if the entropy of L is $\frac{1}{2}$ we can compress long sequences from L by a factor 2.

This concludes our tour. The connection between the complexity of a language and the ability to create crosswords may not come as a surprise. But that this connection leads directly to entropy, the cornerstone of information theory is, at the very least, rather neat.

Notes

This section contains some notes about the history of the results. It is probably most interesting to readers already familiar with the concepts in this paper. The idea of linking crossword puzzles and entropy is, in fact, as old as [Shannon, 1948], from which we quote the last paragraph of section 7:

The redundancy of a language is related to the existence of crossword puzzles. If the redundancy is zero any sequence of letters is a reasonable text in the language and any two-dimensional array of letters forms a crossword puzzle. If the redundancy is too high the language imposes too many constraints for large crossword puzzles to be possible. A more detailed analysis shows that if we assume the constraints imposed by the language are of a rather chaotic and random nature, large crossword puzzles are just possible when the redundancy is 50%. If the redundancy

is 33%, three-dimensional crossword puzzles should be possible, etc.

For a more detailed discussion as well as a bit of history on the result see the last part of [Immink et al., 1998]. The entropy \tilde{H} introduced in this paper is related to the Hartley entropy and Hausdorff dimensions of 'nice' subsets of A^∞ (considered as subsets of $[0, 1]$). Or, if one considers arbitrary subsets of A^∞ , the entropy \tilde{H} might be interpreted as a form of the box counting dimension, see e.g. [Falconer, 1990]. The connection between entropy and Hausdorff dimension is described in [Billingsley, 1965] and interesting results in this direction can be found in [Ryabko, 1986].

References

- Billingsley, P. *Ergodic theory and Information*. John Wiley & Sons, 1965.
- Falconer, K. *Fractal Geometry - Mathematical Foundations and Applications*. John Wiley & Sons, 1990.
- Immink, K.A.S., P.H. Siegel and J.K. Wolf. Codes for digital recorders. *IEEE Trans. Inform. Theory*, 44(6):2260–2299, 1998.
- Ryabko, B. Y. Noiseless coding of combinatorial sources, hausdorff dimension, and kolmogorov complexity. *Problems of Inform. Trans.*, 22(3): 170–179, 1986.
- Shannon, C.E. A mathematical theory of communication. Technical report, Bell System, 1948.