Limitations of the entropy measure in n-gram language modelling

Michael Bauwens

UCLL HOGESCHOOL

RESEARCH & EXPERTISE

What are the most **predictable sentences** given their own (small) dataset?

Which sentences have the **lowest entropy** given a statistical n-gram LM of their dataset?

email dataset

- 913 complaint emails to a fictive company about a delay in the delivery of a smartphone order
- 6k sentences
- 69k trigrams
- 51k unigrams

MLE language model

- trigram model *
- vocab cutoff 2 (1.6k <UNK>)
- via nltk

* all sentences include <s> <s> and </s> </s> pads

1. Shannon entropy

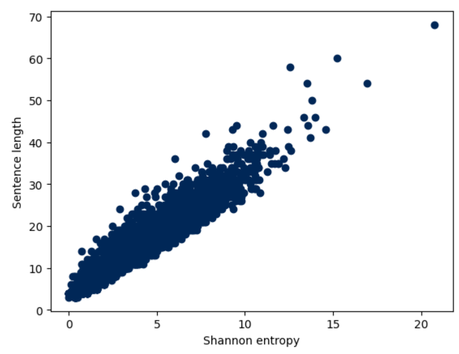
$$H_1(X) = \sum_{i=1}^{n} P(x_i)\ log_2(P(x_i))$$

This formula was introduced in Shannon's information theory. When using a log with base 2, it represents the amounts of bits required to encode a message. This gives us information about the uncertainty of a probability distribution (Shannon, 1948).

Sentences with lowest entropy	
a) <UNK> <UNK> (e.g. "Timothy Verstraeten")	0.001998
b) <UNK> (e.g. "Bart")	0.001998
c) Hartelijk dank	0.001998

Shorter sentences, lower entropy

Because of the exponential effect of the formula (due to multiplication), longer sentences always have a higher entropy (ρ=0.94 ***). This is therefore not a good measure for predictability for sentences with varying length.



2. Length normalisation

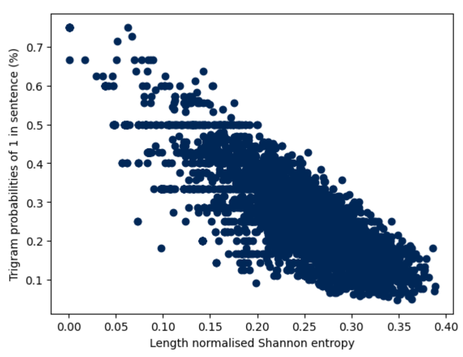
$$H_2(X) = \frac{H_1(X)}{length(X)}$$

In order to cancel out the influence of sentence length, we can instead calculate the average Shannon entropy by normalising by information length. Note: the Shannon-McMillan-Breiman Theorem also exists, which calculates entropy as the negative average log probability (Jurafsky and Martin, 2021).

Sentences with lowest entropy	
a) <UNK> <UNK> (e.g. "Timothy Verstraeten")	0.0004
b) Hartelijk dank	0.0005
c) <UNK> (e.g. "Bart")	0.000666

Disproportionately high probabilities

Correlation with sentence length has diminished (ρ=0.48 ***) but sentences with low entropy are often still short. This is because they contain more unique combinations with a probability of 1 (ρ=-0.82 ***). However, this is not an *intuitive* measure for predictability (nor for typicality) because those trigrams almost always have an extremely low relative frequency.



3. Relative frequency weighting

$$H_3(X) = \frac{\sum_{i=1}^{n} P(x_i)_w\ log_2(P(x_i)_w)}{length(X)}$$

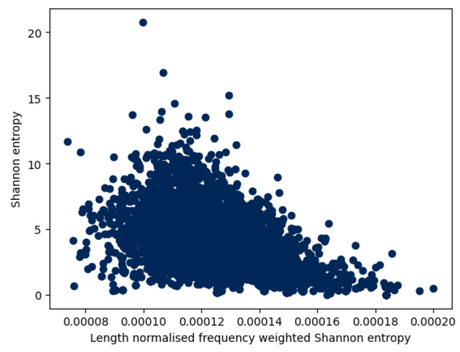
where w is relative frequency of x_i in corpus

The impact of the perfect probabilities in case of low frequency trigrams can be mitigated by adding the weight of its relative frequency. When applying this as is, however, <UNK> tokens receive a large boost because of their collective, high frequency mass. To avoid this, <UNK> tokens can be weighted by the lowest relative frequency in the dataset. This relative frequency weighted version of entropy is then also normalised by length (cf. step 2).

Sentences with lowest entropy	
a) Aangezien ik mijn smartphone vaak nodig heb voor het bijhouden van mijn agenda zou ik jullie willen vragen mij zo snel mogelijk te laten weten wanneer de smartphone geleverd zou kunnen worden.	0.000074
b) Maar dit pakketje heeft al reeds een vertraging van 10 dagen.	0.000076
c) Vriendelijke groeten	0.000076

Typical sentences

Because this modified version penalises characteristics typical to the original Shannon entropy formula, it now correlates negatively (ρ=-0.56 ***). In this use case, it does identify sentences which are more typical for the entire dataset. Among some other typical phrases, the emails contain sentences about smartphones (not) being delivered (a) and about the delay (b), and they include typical genre markers such as "Vriendelijke groeten" (c).



Discussion

- Are these modifications applicable to other datasets? This formula is customised to the characteristics of the measure in this specific email dataset.
- Are these limitations of entropy typical for smaller datasets? What about larger corpora?
- What's the impact on the entropy measure used as intrinsic evaluation? Entropy (and its derived perplexity) are often used to gauge performance of a language model on a test set. If these limitations also apply in this context, this might impact the interpration and validity.

Michael Bauwens
Language technology 🧑‍💻 Soft Skills 🎓
Int'l student recruitment 🌎

Check out code via (link in LinkedIn post)

jupyter

Jurafsky, D., & Martin, J. (2021). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423.

CLIN32 Tilburg 2022
Computational Linguistics In the Netherlands

michael.bauwens@ucll.be

research-expertise.ucll.be

@ucll_re

@ucll.re

/company/ucleuvenlimburg-re