Entropy-based, text essence picking strategy

Adam Dohojda

October 2023

1 Introduction

This technique will be based on simple metric called Entropy. It can be explained as expected value of surprise when sampling from population. Don't want to delve too deep into the mathematical side of it, but the more homogenous population, the smaller the surprsie, and the more diverse-the bigger surprise upon each picking.

2 Goal

Goal is to reduce set of unique words from any text, let it be $\{W\}$ to set of the most surprising words $\{S\}$, where the raw text will be treated as a set $\{T\}$ where words do repeat according to the original text. The phrase *most surprising* is arbitrary, and lacks solid mathematical explanations, but still can be used due to its utility. Actually explaining my interpretation and approach to this concrete problem is the main subject of this paper.

3 Method theory

Let us have $Entropy = E[Surprise(w)] = \sum_{w \in W} \log_2(\frac{1}{P(w)}) * P(w)$. This will be a mean value of surprise for random pick of a word from T. Now according to Shannon information theory, the rarer the outcome, the more information it brings. So I propose treating surprise for each word as it's rarity metric. And as it grows as words are less probable, it seems to be a good candidate for that job. But how will we decide what word is rare, and what is not rare? I came up with idea of using standard deviation of surprise, which will be defined as follows:

$$\sqrt{D^2(surprise(w))} = \sqrt{\sum_{w \in W} (\log_2(\frac{1}{P(w)}) - Entropy)^2 * P(w)}$$

This will allow us to treat words more like observations from random distribution (which in fact they are). And the prefered words with more valuable information will share certain similarities with outliers in sampling theory. There

will be words with *typical surprise* and those with *non-typical surprises*. Now the set $\{S\}$ can take form as follows:

$$\{S\} = \{w \in W : surprise(w) \in (entropy + k \cdot \sqrt{D^2(surprise(w))}, +\infty\}$$

For my goal, words with lower than typical surprise are not points of my interest. That is why I don't account the words that have surprise below typical levels.

4 Methodology of picking the right k

Let us see what cutoff point makes sense, and what it's usefulness depends on.

Simple fact for that matter will be that the maximum surprise that is possible to achieve on given set $\{Q\}$ is $\log_2(|Q|)$ when using classical probability.

proof: As Entropy is undefined for impossible events (P(X) = 0), it's maximum value is reached for smallest possible probability in a set. And as it is obvious, the smallest amount of occurrence of the element in a set is 1. So it gives us probability equal to $\frac{1}{|Q|}$ because $|Q|$ is number of all elements we are picking amongst. Thus, $\forall q \in \{Q\}, surprise(q) \in (0, \log_2((\frac{1}{|Q|})^{-1}))$. QED.

On this basis we can state that cutoff point should suffice the condition:

$$entropy + k \cdot \sqrt{D^2(surprise(w))} < \log_2((\frac{1}{|T|})^{-1}), k \in \mathbb{N}$$

thus:

$$k < \frac{\log_2(|T|) - entropy}{\sqrt{D^2(surprise(w))}}, k \in \mathbb{N}$$

Where all components of the equation can be calculated using information available from $\{T\}$ (set of words from original text): set $\{W\}$ of unique words, probabilities assigned to each word and $|T|$.

5 Possible further improvements

Of course this is just a basic proposition but I think theory (and some practical tinkering with texts) render it promising. What can have an impact on amount of information (how well we get the essence) out of text T will depend on how well we prepare $\{W\}$. Some obvious interventions include:

1. removing conjunctions as they carry small information,
2. transforming words into their basic form

6 References

1. Working example in form of Python script: nlp_entropy.py