IA225 -
NGD-BASED WORD PREDICTION ALGORITHM FOR CONTEXTUALLY RELEVANT SUGGESTIONS

THALIS ROCHA PESTANA

PALAISEAU
2023

1. INTRODUCTION

Researchers like Bennett, Vitanyi, Cilibrasi, and others have made important contributions to using Kolmogorov complexity for classification and measuring the shared information between objects. Their work has yielded promising results in different fields, including natural language.

The idea behind their approach is to compare pairs of objects and determine how much information they have in common. This is done by assigning a distance value that reflects the similarity between binary representations of the objects. If two objects share a lot of common information, their distance is small, indicating they are close. Conversely, if two objects have less shared information, their distance is larger, suggesting they are more independent [2].

For instance, when comparing two identical words, their distance is zero because they have all the information in common. On the other hand, two completely unrelated words would have a distance close to 1 in a normalized scale of distances between 0 and 1. The researchers' goal is to measure the similarity or dissimilarity between objects by quantifying the amount of shared information. This provides a way to classify objects based on their common characteristics, regardless of the specific domain they belong to.

In this context, the researchers introduced three metrics: NID (Normalized Information Distance), NCD (Normalized Compression Distance), and NGD (Normalized Google Distance). These approaches are aligned with Kolmogorov's concept of defining a numerical measure of information content of words, i.e. a measure of their randomness.

The objective of this research is to explore the effectiveness of utilizing specific distance measures in the prediction of suitable words to complete given incomplete phrases. To achieve this, an interface has been developed in Python. This interface accepts an incomplete sentence as input and feeds it into a pre-trained language model, GPT-2. The model is then asked to produce a certain number of word predictions, as specified by the user, to complete the sentence. Subsequent calculations are performed using Normalized Google Distance (NGD) and Normalized Compression Distance (NCD) measures to determine the "distance" between the input phrase and each of the generated predictions. Following these calculations, the system presents the top 5 predictions.

2. METHODOLOGY

This study is based on the notions of Normalized Information Distance (NID) which measures the normalized difference in information content between two objects. NID is based on conditional complexity $K(x|y)$ and measures the amount of information contained in x that is not present in y.

Bennet and al. defined information distance between to words x and y as the size of the shortest program which maps x to y and y to x. An alternative definition can be given as follows:

$$ID'(x,y) = max\{K(x|y),\ K(y|x)\} \quad (1)$$

The equation 1 states that the shortest program which computes x from y takes into account all similarities between x and y. By using NID as a distance measure, it is possible to define a metric space where the distance between objects is determined by their information content. Therefore, NID satisfies the metric axioms:

$(M1)\ for\ any\ x\ \in\ X,\ d(x,x)\ =\ 0$ $\quad (identity)$
$(M2)\ for\ any\ x,\ y\ \in X,\ if\ x \neq y,\ then\ d(x,y) \neq 0$
$(M3)\ for\ any\ x,\ y\ \in X,\ d(x,y)\ = d(y,x)$ $\quad (symmetry)$
$(M4)\ for\ any\ x,\ y,\ z\ \in X,\ d(x,y)\ \leq d(x,z)\ +\ d(z,y)$ $\quad (triangle\ inequality)$

From the axioms it is possible to develop the Normalized Information Distance as formulated in the equation 2.

$$NID(x,y) = \frac{K(x,y) - min[K(x), K(y)]}{max[K(x),K(y)]} \quad (2)$$

The NID remains an abstract notion, since Kolmogorov complexity is not computable [1]. Consequently, NID was proposed as an ideal distance which can be approximated by replacing the Kolmogorov function K by computable compression algorithms. Vitanyi proposed two other metrics : Normalized Compression Distance (NCD) and Normalized Google Distance (NGD).

NCD is obtained by replacing K by a compressor such as “zip” as shown in equation 3.

$$NCD(x,y) = \frac{Z(x,y) - min[Z(x), Z(y)]}{max[Z(x), Z(y)]} \quad (3)$$

NGD is obtained by replacing K by $log_2(1/f(x))$, where $f(x)$ is the observed frequency of x on the Web, or by $log_2(N/g(x))$, where N is the corresponding total number of indexed pages and $g(x)$ denotes the number of pages containing x. The equation 4 shows the NGD.

$$NGD(x,y) = \frac{max[log(f(x)), log(f(y))] - log(f(x,y))}{log\ N - min[log(f(x)), log(f(y))]} \quad (4)$$

2.1 Pre-processing

The initial idea of applying NGD directly to make predictions had limitations due to the computational feasibility of calculating NGD for each word in a vocabulary with several words. To overcome this challenge, the GPT-2 model was utilized to make predictions and the NGD was applied to rank the predictions, returning the “best” 5 words to complete a phrase given by the user.

The GPT-2 model is a powerful language model that has been trained on a large corpus of text data from the internet. It has the ability to generate coherent and contextually relevant text based on the patterns it has learned from the training data. Despite being a powerful tool, the GPT-2 was doing identical predictions for simple sentences with contextual information. For instance, the phrase: “I’m tired, I need…” received the word “help” several times. As the main objective was to provide a word predictor with different word suggestions, a pre-treatment of the information needed to be done before the computation of the distances in order to avoid multiple identical predictions.

An additional challenge we encountered involved conducting searches on Google. Ideally, using Google's API is the most appropriate method to extract search results from Google. However, this API has very limited quotas. To avoid problems with these limits, two other libraries were experimented: *BeautifulSoup* and *Selenium*. However, Google's built-in measures to prevent automated scraping resulted in inconsistent search result numbers when executing the script. Therefore, the decision was made to utilize the Google API, even with its limitations.

Another issue arose in connection to how the input phrase was queried on the search engine. Without enclosing the search term in quotation marks, the engine returned results containing all words from the phrase, but not in the specified order. By incorporating quotation marks, phrases like "I'm tired" only returned results with all words in the exact stated order.

2.2 Method evaluation

After the pre-processing, two approaches were applied. The first approach consists in calculating the NGD between the entire input phrase and each word predicted (without any tokenization) and the second approach consists in calculating the mean NCD in an analogous way.

To test the different methods, some phrases were used as inputs for the system, and the predictions it made were collected for review. These phrases covered a wide variety of common situations and topics to check how flexible the system is. The phrases used and the expected predictions were:

“I am tired, I need…” (sleep)
“Last week I went to fast foods everyday, I love…” (eat/hamburguer)
“I play football every week. Football is my…” (passion/hobby/favorite)
“There is a movie showing, want to go with…” (me?/ us?)
“Spiderman is a…” (superhero/ hero)

3. RESULTS

The table 1 shows for each phrase tested the predictions from GPT-2, the rank of words using NGD and the rank using NCD. For a posterior comparison, the word predictions from an iPhone 11 keyboard were noted. In the third topic of the References section is provided a link with some images from the prompt interface with the words predicted by GPT-2 and the suggestions based in NGD and NCD.

Table 1 - Input phrases, predictions and suggestions.

	GPT-2 Predictions	NGD ranking	NCD ranking	iPhone keyboard
“I’m tired, I need…”	['another', 'rest', 'to', 'more', 'a', 'you', 'your', 'some', 'sleep', 'help']	1 - 'rest'; 2 - 'sleep'; 3 - 'another'; 4 - 'some'; 5 - 'more'	1 - 'another'; 2 - 'rest'; 3 - 'to'; 4 - 'more'; 5 - 'a'	1 - to; 2 - a; 3 - some
“Last week I went to fast foods everyday, I love…”	['it', 'my', 'to', 'their', 'burgers', 'fast', 'being', 'food', 'the', 'them']	1 - 'burgers'; 2 - 'fast'; 3 - 'food'; 4 - 'being'; 5 - 'them'	1 - 'fast'; 2 - 'food'; 3 - 'it'; 4 - 'my'; 5 - 'to'	1 - it; 2 - the; 3 - ❤️
“I play football every week. Football is my…”	['life', 'team', 'highlight', 'hobby', 'friends', 'body', 'number', 'passion', 'go', 'big']	1 - 'hobby'; 2 - 'passion'; 3 - 'highlight'; 4 - 'friends'; 5 - 'team'	1 - 'life'; 2 - 'team'; 3 - 'highlight'; 4 - 'hobby'; 5 - 'friends'	1 - first; 2 - favorite; 3 - best
“There is a movie showing, want to go with…”	['with,', 'TR', 'me', 'it', 'them', 'with?"', 'a', 'the', 'that', 'some']	1 - 'with,'; 2 - 'them'; 3 - 'some'; 4 - 'that'; 5 - 'it'	1 - 'with,'; 2 - 'with?"'; 3 - 'that'; 4 - 'TR'; 5 - 'me'	1 - the; 2 - us; 3 - a
“Spiderman is a…”	['big', 'member', 'superhero', 'real', 'master', 'monster', 'good', 'super', 'fantastic', 'former']	1 - 'superhero'; 2 - 'fantastic'; 3 - 'former'; 4 - 'monster'; 5 - 'master'	1 - 'big'; 2 - 'member'; 3 - 'superhero'; 4 - 'real'; 5 - 'master'	1 - cute; 2 - nice; 3 - cool

Source: Produced by the author.

4. DISCUSSION AND CONCLUSIONS

We can observe that, with the exception of the fourth phrase, the NGD ranking aligns with our expected predictions. NCD also performs well but omits some significant words, such as 'passion' in the third phrase, 'burgers' in the second, and 'sleep' in the first. The iPhone keyboard prediction is less successful in these instances.

However, when examining the fourth input phrase, NCD and the iPhone keyboard provided superior predictions. NCD included "us" among its top predictions, while the iPhone keyboard suggested "me". In this case, the words predicted by the GPT-2 model are largely functional or "closed-class" words. This class includes pronouns, determiners, prepositions, and conjunctions, which are infrequently updated with new words—hence, they are "closed." These words typically serve grammatical or structural functions and are common in a variety of contexts, thereby reducing their distinctiveness. This high frequency and diversity block the specificity of the NGD measure, making it challenging to effectively rank these words.

Another issue lies not with the NGD function, but with the predictions generated by the GPT-2 model. The GPT-2 model occasionally produced predictions that did not align with the expected outcomes. For instance, the phrase "Spiderman is a…" might result in predictions such as: powerful; unique; new; small; serious; good; Super; mutant; prominent; p—without any mention of the word "hero". However, it does demonstrate the model's capability to understand the context of the words since GPT-2 consistently returns "hero" or "superhero" for the input phrase: "Dr. Octopus is a villain. Spiderman is a…".

It's important to note that we need more sophisticated methods for evaluating the model's performance. Testing a larger number of input phrases with a more extensive vocabulary of expected predictions of similar meanings would allow the computation of more comprehensive metrics, such as precision, recall, or F1-score. But as the Google API, the only method found to provide consistent search results, imposes a daily quota, such extensive testing is not feasible.

Another significant point is the unsuitability of this model for real-time applications. In real-world scenarios where people need to compose messages quickly, the model's average response time of 3 minutes is impractical for regular conversation in a messaging application.

In conclusion, while the Normalized Google Distance (NGD) ranking generally aligns well with the expected predictions, it falters when dealing with high-frequency, closed-class words. In addition, the model is limited by the GPT-2 predictions that do not always align with the ground truth. More sophisticated evaluation methods are required for better model assessment and the response time renders it unsuitable for real-time applications such as instant messaging. Despite its shortcomings, the model still demonstrates an impressive capacity for contextual understanding.

REFERENCES

1 - Ferbus-Zanda, Marie, and Serge Grigorieff. **"Kolmogorov Complexity in perspective."** arXiv preprint arXiv:0801.0354 (2008).

2 - **"FCI - Chapitre 2: Codage et Compression."** 2023. AI225. Télécom Paris. Accessed June 13, 2023. https://aicourse.r2.enst.fr/FCI/Chapitre2.html.

3 - **"Results Folder"** Google Drive, accessed June 13, 2023, https://drive.google.com/drive/folders/1JEq2gva_yyAUsWOjuag-i692rwmFn9g9?usp=sharing.