Text Generation using Markov chain

(A Natural Language Generation (NLG))

Md Ismail Hossain

Spring 2022: MATH-486-01-Intro to Stochastic Processes
Department of Mathematics
New Mexico Institute of Mining and Technology

May 2, 2022

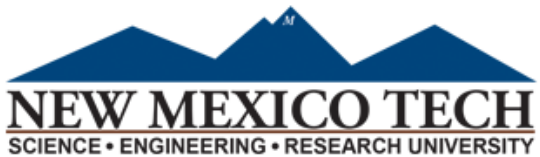

Table of Contents

1. Introduction
2. Objectives and Data
3. Methodology
4. Findings
5. Conclusion

NEW MEXICO TECH
SCIENCE • ENGINEERING • RESEARCH UNIVERSITY

Introduction

Some Application of NLG:

- Chatbots.
- Siri, Alexa, Google Assistant.
- When we type a email we noticed suggestion for next word or sometime sentences in Gmail.

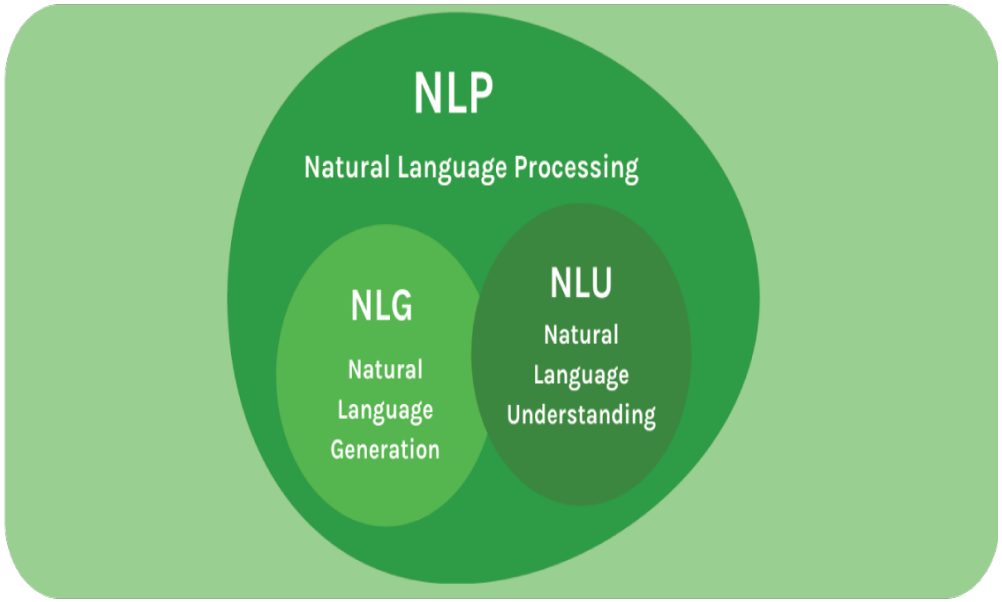

Figure: Language analytics domain.

Objectives and Data

Objective

- Would calculate the Entropy of the train data to measure it's predictability.[Shannon, 1951]
- The main objective is to use Markov chain to generate next words.[J, 2018, Strika, 2019]

Analysis Tools: Python

Data

Used three different author works and applied the methodologies to predict the next word by Markov chain:

- Shakespeare [Macbeth] ▸ Data Source
- J.K. Rowling [The Philosopher's Stone] ▸ Data Source
- George R. R. Martin [Game of Thrones] ▸ Data Source

NEW MEXICO TECH
SCIENCE • ENGINEERING • RESEARCH UNIVERSITY

Methodology

Analysis steps:

- Calculate the Entropy of the entire document. (word level)
- Construct the transition probability matrix (TPM).
- Predict the next word or words using the TPM.

NEW MEXICO TECH
SCIENCE • ENGINEERING • RESEARCH UNIVERSITY

Entropy

Entropy inherent the average uncertainty to the variables possible outcomes. It can be expressed using the function:

$$H(X) = -\sum_{i=1}^{n} P(x_i) log_b P(x_i)$$

where,

X = Discrete random variables with possible outcome $x_1, x_2, x_3, ..., x_n$

$P(x_i)$ = The probability of occurring the i-th character.

b = The base of the logarithmic function.

For our case we would use b = length of the $x_1, x_2, x_3, ..., x_n$ to get the $H(X)$ value within 0 to 1.

Methodology

Markov Property

In our case, x_t = process are in t-th character or word. So, using the Markov property, the x_{t+1}-th word or character only depends on x_t-th, not the previous words!

Transition Probability Matrix

By using the counts of the individual words or characters we can create the transition probability from i-th to j-th state:

$$p_{ij} = \mathbb{P}(X_{t+1} = j | X_t = i)$$

So, the transition probability matrix can be expressed as:

$$P = (p_{ij})$$

NEW MEXICO TECH
SCIENCE • ENGINEERING • RESEARCH UNIVERSITY

Entropy results

Table: Entropy values within 0 to 1

Book	**Author**	**Entropy**
Macbeth	William Shakespeare	0.71
The Philosopher's Stone	J. K. Rowling	0.65
Game of Thrones	George R. R. Martin	0.57

NEW MEXICO TECH
SCIENCE • ENGINEERING • RESEARCH UNIVERSITY

Macbeth

Next 5 words after Macbeth

'Macbeth Macb A foolish thought honest'

Next 10 words after Macbeth

'Macbeth Rosse Ile be commanded heeres a TraitorAnd must be all'

Next 20 words after Macbeth

'Macbeth MacbethBeware MacduffeBeware the Charme the Liars and pristine HealthI would be acted ere they comming on the heat of that'

NEW MEXICO TECH
SCIENCE • ENGINEERING • RESEARCH UNIVERSITY

The Philosopher's Stone

Next 5 words after Harry
'Harry nodded vigorously. "Stop!'

Next 10 words after Harry
'Harry? I have a grin.. Professor Dumbledore gently'

Next 20 words after Harry
'Harry found an excellent adventure was another word on it had just one another point beating wouldn't be the board,'

Game of Thrones

Next 5 words after Jon Snow

'Jon Snow took a bloody cloak.'

Next 10 words after Jon Snow

'Jon Snow with a blanket of snow crunched beneath his feet.'

Next 20 words after Jon Snow

"Jon Snow, why is the King's Justice, you had a mind needs books as a whisper, and so it"

NEW MEXICO TECH
SCIENCE • ENGINEERING • RESEARCH UNIVERSITY

Conclusion

In conclusion we can say that there is scope of improvement of the study:

- Can clean the data in a better way. Because some time I found two word contacting together.
- Use several other Machine learning model to compare the performance of our model.

NEW MEXICO TECH
SCIENCE • ENGINEERING • RESEARCH UNIVERSITY

Thank You

NEW MEXICO TECH
SCIENCE • ENGINEERING • RESEARCH UNIVERSITY

References I


J, A. M. (2018).
Next Word Prediction using Markov Model.
https://medium.com/ymedialabs-innovation/next-word-prediction-using-markov-model-570fc0475f96.
[Online; accessed 16-March-2018].


Shannon, C. E. (1951).
Prediction and entropy of printed english.
The Bell System Technical Journal, 30(1):50–64.


Strika, L. (2019).
Markov Chains: How to Train Text Generation to Write Like George R. R. Martin.
https://www.kdnuggets.com/2019/11/markov-chains-train-text-generation.html.
[Online; accessed 29-November-2019].

NEW MEXICO TECH
SCIENCE • ENGINEERING • RESEARCH UNIVERSITY