

Митник Денис Олександрович, здобувач вищої освіти КПП ім. Ігоря Сікорського, Україна
Науковий керівник: Гавриленко Олена Валеріївна, кандидат фізико-математичних наук, доцент, доцент кафедри інформаційних систем та технологій ФІОТ КПП ім. Ігоря Сікорського, Україна

ПОРІВНЯННЯ ЛАНЦЮГІВ МАРКОВА ТА НЕЙРОМЕРЕЖ ДЛЯ ВИРІШЕННЯ ЗАДАЧІ ГЕНЕРАЦІЇ ВІРШІВ

Анотація. В даній статті обумовлено важливість задачі генерації віршів. Описано алгоритми генерації віршів за допомогою ланцюгів Маркова та нейромереж. Охарактеризовано переваги та недоліки кожного з підходів та здійснено їх порівняння.

КЛЮЧОВІ СЛОВА: генерація віршів, ланцюги маркова, нейромережі, natural language processing.

Abstract. This article states the importance of the task of generating poems. Algorithms for generating poems using Markov chains and neural networks are described. The advantages and disadvantages of each of the approaches are characterized and their comparison is made.

KEYWORDS: verse generation, Markov chains, neural networks, natural language processing.

1. Вступ

Задача генерації тексту та віршів є однією з найбільш актуальних в сфері Natural Language Processing. Існує досить велика кількість підходів та моделей для генерації тексту, проте багато з них мають досить велику різницю в порівнянні зі справжньою людською мовою, отже постає задача контексту, тобто логічно звязаного змісту згенерованого тексту. Задача генерації віршів розкриває генерацію не лише з боку логічного змісту, а й ритму, рими та емоцій, розв'язання цієї задачі дозволяє зробити продукти які використовують згенеровану мову більш привабливими для користувачів, такі як голосові помічники або чат-боти. Також генерація віршів може бути одним зі способів генерації синтетичних даних для тренування мовних моделей. Однією з проблем які виникають в творчості є проблема авторського права і вона також може бути вирішена за допомогою генерації віршів. З теоретичної точки зору ця задача відкриває нові грані розуміння мов з боку computational linguistics, а саме пояснює яке формальне представлення мови та які методи є найбільш доречним для моделювання її комп'ютером.

2. Ланцюги Маркова

Ланцюг Маркова це ймовірнісна модель яка описує послідовність подій і ймовірність

наступної події залежить лише від поточного стану моделі [1]. У випадку генерації віршів ланцюги Маркова можуть використовуватися для знаходження наступного слова за наявними вже у вірші.

На першому кроці алгоритму необхідно розділити тренувальний корпус на пари вікно (слова які передують) та слово яке слідує за цим вікном. В результаті ми отримуємо структуру даних наступного вигляду [2]:



Рисунок 1. Структура даних для збереження вікон та ймовірностей наступних слів

Обираємо початкову послідовність для генерації, після цього виділяємо з послідовності вікно зі слів, знаходимо ймовірні наступні слова та обираємо з ймовірностями наступне слово, даних алгоритм продовжується поки вірш не буде мати необхідну довжину. Наведемо псевдокод алгоритму:

ПОЧАТОК

- 1) Розбити корпус на пари вікно - наступне слово та розрахувати ймовірності наступного слова
- 2) Отримати початкову послідовність
- 3) Поки довжина згенерованої послідовності менша за очікувану:

- a) Отримати вікно з послідовності – Input Gate
- b) Обрати наступне слово за ймовірностями з корпусу – Forget Gate
- c) Додати наступне слово до згенерованої послідовності – Output Gate

КІНЕЦЬ

Наведений алгоритм є найпростішою реалізацією ланцюга Маркова для генерації віршів, проте даний алгоритм можна модифікувати аби він краще зберігав риму у вірші. Наприклад під час отримання наступного слова за вікном можна надавати більше перевагу слова, які римуються відповідно до обраної схеми римування, також в такому випадку необхідно обмежувати кількість слів у одному рядку вірша, аби розуміти яке слово останнє в рядку [3].

Перевагами даного алгоритму є швидкість імплементації і досить непогана імітація стилю вірша, проте простий статистичний підхід не завжди може зберегти ритм та риму вірша.

3. Нейромережевий підхід

Одним з основних напрямків NLP є методи на основі нейронних мереж. Задача генерації віршів також може вирішуватися цим підходом, однією з найбільш ефективних архітектур є рекурентні нейронні мережі.

Рекурентні нейронні мережі особливі тим, що в них зв'язки вузлів можуть бути з самими собою, що дозволяє даному типу мереж дуже якісно обробляти будь-які послідовності [4]. Однією з найефективніших реалізацій блоків рекурентної нейромережі є LSTM (Long Short Term Memory).

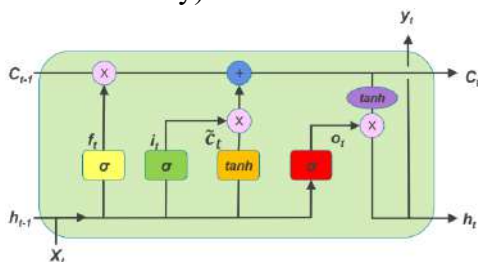


Рисунок 2. Структура LSTM блоку[5]

Цей блок особливий тим що він здатний вивчати довготривалі залежності, що в задачі генерації віршів є надзвичайно важливим для збереження ритму та змісту.

Блок LSTM складається з наступних частин:

Рівняння Gate:

$$i_t = \sigma(w_i[h_{t-1}, x_t] + b_i)$$

$$f_t = \sigma(w_f[h_{t-1}, x_t] + b_f)$$

$$o_t = \sigma(w_o[h_{t-1}, x_t] + b_o)$$

де

- i_t – input gate
- f_t – forget gate
- o_t – output gate
- σ – сигмоїд-функція
- w_x – ваги нейронів відповідного gate
- h_{t-1} – вихідні дані попереднього lstm блоку
- x_t – вхідні дані в момент часу t
- b_x – bias для відповідного gate

Рівняння для стану:

$$c_t^{\wedge} = \tanh(w_c[h_{t-1}, x_t] + b_c)$$

$$c_t = f_t * c_{t-1} + i_t * c_t^{\wedge}$$

$$h_t = o_t * \tanh(c_t)$$

Реалізувати генерацію віршів за допомогою lstm нейромережі можна наступним чином [6]:

ПОЧАТОК

- 1) Розбити тренувальний корпус на пари вікно-наступне слово
- 2) За допомогою даного датасету тренувати методом градієнтного спуску нейромережу
- 3) Зберегти ваги
- 4) Задати початкову послідовність
- 5) Поки довжина згенерованої послідовності менша за бажану:

- a) Отримати вікно з послідовності
- b) Подати вікно на вхід нейромережі та отримати наступне слово

- с) Додати наступне слово до згенерованої послідовності

КІНЕЦЬ

Перевагами використання нейронних мереж для генерації віршів є їх більш складна лінійна трансформація параметрів, що дозволяє вивчити залежності в тексті та

зберігати контекст [6]. Вони показують кращий результат та є дуже гнучкими щодо мотифікацій. серед недоліків цього підходу варто зазначити досить довгий час тренування та підбору гіпер параметрів моделі та досить нетривіальну реалізацію усіх компонентів моделі.

Висновки. Було описано такі алгоритми генерації віршів як ланцюги Маркова та нейромережі. В результаті порівняння алгоритмів було виявлено, що статистичні методи не завжди можуть вловити ритм та риму віршу, в той час як рекурентні нейронні мережі здатні якісно вловлювати контекст, ритм та риму віршу. Використання нейромереж для задачі генерації віршів при правильному підборі архітектури може досягати майже людських результатів.

Список інформаційних джерел

1. Manuel Brenner (2022). Understanding Markov Chains [Електронний ресурс] // Режим доступу <https://towardsdatascience.com/understanding-markov-chains-cbc186d30649>
2. Mehrab Jamee (2018). Making a Markov Chain Poem Generator in Python [Електронний ресурс] // Режим доступу: <https://medium.com/upperlinecode/making-a-markov-chain-poem-generator-in-python-4903d0586957>
3. Gabriele Barbieri (2013). Markov constraints for generating texts with style [Електронний ресурс] // Режим доступу http://amsdottorato.unibo.it/5685/1/barbieri_gabriele_tesi.pdf
4. Alex Sherstinsky (2020). Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network [Електронний ресурс] // Режим доступу <https://arxiv.org/pdf/1808.03314.pdf>
5. Divyanshu Thakur (2018). LSTM and its equations [Електронний ресурс] // Режим доступу <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>
6. Paul Mooney (2017). Poetry Generator (RNN Markov) [Електронний ресурс] // Режим доступу <https://www.kaggle.com/code/paultimothymooney/poetry-generator-rnn-markov>