Tham khảo project:

<https://drive.google.com/drive/folders/1ab6ct5xDF8oJmki8-0RJZl3V4mha6yew>

<https://github.com/hoangducsn00/Da-poet>

Background:

Thơ lục bát là thể thơ đặc trưng của tiếng Việt, có quy luật cơ bản rất đơn giản:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Câu 6 | - | Bằng | - | Trắc | - | **Bằng** |  |  |
| Câu 8 | - | Bằng | - | Trắc | - | **Bằng** | - | Bằng |

2 tiếng Bằng có vần với nhau

Câu 8: tiếng 6 và tiếng 8 dù đều là vần bằng, nhưng 1 tiếng có thanh huyền, một tiếng có thanh ngang

Ví dụ:

Thôn đoài ngồi nhớ thôn Ðông

Một người chín nhớ mười mong một người

Gió mưa là bệnh của trời

Tương tư là bệnh của tôi yêu nàng.

Nguyễn Bính

Bỗng dưng buồn bã không gian

Mây bay lũng thấp giăng màn âm u

Nai cao gót lẫn trong mù

Xuống rừng nẻo thuộc nhìn thu mới về.

Huy Cận

Ngoại lệ: Có hai ngoại lệ trong thơ lục bát [[see](https://sites.google.com/site/baclieu32/-viet-bai-huong-dan-lam-tho/huongdanlamtholucbatvasongthatlucbat)]

Problem statement:

* input: câu lục
* output: câu bát

dataset: không tìm thấy data sẵn về thơ lục bát tiếng việt? => phải tự xây dựng

Tiếp cận:

* dataset chỉ chấp nhận thơ lục bát chuẩn (không phải song thất lục bát, không phải lục bát biến thể, ưu tiên ca dao để đồng nhất về phong cách sáng tác?, nếu chỉ ca dao gây thiếu data thì về sau add thêm tác phẩm khác: truyện Kiều, Lục vân tiên...)
* coi mỗi cặp lục bát là 1 quan sát độc lập: vd một bài thơ 4 câu, 6-8-6-8 thì tách thành 2 bài lục bát riêng
* x = [w1 w2 … w6]; y = [w1 w2 … w8]
* tokenize ở mức tiếng (không phải mức từ)

keywords: seq2seq, gpt2, lstm/gru, bidirectional, attention?

Project stages:

1. Sinh được câu 8 từ câu 6: chưa cần đúng vần, chưa cần đúng luật
2. Câu 8 (đứng một mình) có nghĩa
3. Câu 8 có vần với câu 6
4. Câu 8 đúng luật Bằng Trắc (how?)
5. Câu 8 có ngữ nghĩa ngữ cảnh giống câu 6 (how?)