TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM
TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG
KHOA CÔNG NGHỆ THÔNG TIN

DỰ ÁN CÔNG NGHỆ THÔNG TIN

Phát triển mô hình và xây dựng ứng dụng sinh thơ tiếng Việt sử dụng kỹ thuật Natural Language Generation (NLG)

Người hướng dẫn: **GS.TS. LÊ ANH CƯỜNG**

Người thực hiện: **NGUYỄN CHÍ ANH – 520H0335**

TRẦN VĂN HUY – 520H0538

Khoá **: 24**

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM
TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG
KHOA CÔNG NGHỆ THÔNG TIN

DỰ ÁN CÔNG NGHỆ THÔNG TIN

Phát triển mô hình và xây dựng ứng dụng sinh thơ tiếng Việt sử dụng kỹ thuật Natural Language Generation (NLG)

Người hướng dẫn: **GS.TS. LÊ ANH CƯỜNG**
Người thực hiện: **NGUYỄN CHÍ ANH**
TRẦN VĂN HUY
Khoá : **24**

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024

LỜI CẢM ƠN

(Sẽ được cập nhật sau)

.....
.....
.....
.....
.....
.....
.....
.....
.....

TP. Hồ Chí Minh, ngày 2 tháng 1 năm 2024
Tác giả
(Ký tên và ghi rõ họ tên)

Nguyễn Chí Anh

Trần Văn Huy

CÔNG TRÌNH ĐƯỢC HOÀN THÀNH TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG

Chúng em xin cam đoan đây là công trình nghiên cứu của riêng nhóm và được sự hướng dẫn khoa học của GS.TS. Lê Anh Cường. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong bài báo cáo còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

Nếu phát hiện có bất kỳ sự gian lận nào chúng em xin hoàn toàn chịu trách nhiệm về nội dung bài báo cáo của mình. Trường Đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do chúng em gây ra trong quá trình thực hiện (nếu có).

TP. Hồ Chí Minh, ngày 2 tháng 1 năm 2024

Tác giả

(Ký tên và ghi rõ họ tên)

Nguyễn Chí Anh

Trần Văn Huy

TÓM TẮT

Hiện nay lĩnh vực sinh văn bản tự động đang nổi lên như một hiện tượng với sự xuất hiện của các chat bot thông minh như Chat-GPT, Bard,… giúp con người trả lời mọi thắc mắc về các lĩnh vực trong cuộc sống mà không cần tốn công trà và đọc trên google như ngày trước. Điều này đã cho chúng em ý tưởng về tạo ra một công cụ sinh thơ mà ở đó con người chỉ cần yêu cầu sinh ra một bài thơ dựa trên những yêu cầu đầu vào theo ý muốn là có thể có cho mình một bài thơ hoàn chỉnh. Hiện nay, đã có không ít các nghiên cứu về mô hinh sinh thơ đem lại hiệu quả cao và nhanh chóng. Tuy nhiên, các nghiên cứu ngày nay vẫn còn hạn chế về ngôn ngữ được sử dụng cho bài toán và ngữ cảnh của bài thơ. Hầu như các nghiên cứu phát triển mô hinh sinh thơ bằng tiếng Việt vẫn còn hạn chế nhưng sự ra đời của các nghiên cứu này đã đem lại hướng tích cực và đặt nền tảng để các mô hình sinh thơ đã và đang phát triển sau này hoàn thiện hơn.

Trong dự án này, chúng em sẽ nghiên cứu cụ thể về sinh thơ Lục Bát với kỹ thuật sinh văn bản tự động(Natural Language Generation – NLG), sử dụng mô hình có sẵn là GPT-2. Để phục vụ việc huấn luyện mô hình thì chúng em cào dữ liệu về thơ lục bát ở nhiều nguồn khác nhau như trang thivien.net, facebook và bộ dữ liệu thơ lục bát trên github.

MỤC LỤC

DANH MỤC HÌNH VẼ

DANH MỤC BẢNG BIỂU

DANH MỤC CÁC CHỮ VIẾT TẮT

NLG	Natural Language Generation
RNN	Recurrent Neural Network
LSTM	Long Short Term Memory

CHƯƠNG 1 – MỞ ĐẦU

Trong chương này, sẽ giới thiệu về tổng quan đề tài, lý do chọn đề tài, mục tiêu thực hiện đề tài, phân tích đối tượng và phạm vị nghiên cứu, đưa ra phương pháp nghiên cứu, ý nghĩa thực tiễn của đề tài mang lại và cuối cùng là khái quát về bố cục nội dung của đề tài

1.1 Lý do chọn đề tài

Chắc hẳn chúng ta ai cũng đã được tiếp xúc với nền văn học Việt Nam trong suốt quá trình trưởng thành, đặc biệt là với kho tàng thơ đồ sộ của ông cha ta để lại. Chẳng hạn những mẩu truyện Kiều của Nguyễn Du từ thời phong kiến xa xưa cho đến những bài thơ cách mạng thời kháng chiến chống Pháp, Mỹ, ít nhiều ta cũng được tiếp xúc với nó qua lời ru của mẹ, của ngoại hay qua những tiết ngữ văn thời còn là học sinh cấp 2, cấp 3. Tuy nhiên như chúng ta đã biết để sáng tác ra một bài thơ hoàn chỉnh và hay, một bài thơ mà để lại dư vị trong lòng người thưởng thức như ông cha ta ở thời nay là rất khó do nó đòi hỏi tác giả phải là người có ngòi bút bay bổng, tâm hồn thơ ca, và đặc biệt là phải bỏ rất nhiều thời gian tâm huyết để xây dựng bài thơ, điều mà có lẽ trong cuộc sống "chạy đua với thời gian" hiện đại ngày nay chúng ta ít ai có được.

Chính vì thế đã có các nghiên cứu liên quan đến việc phát triển mô hình sinh thơ bằng cách áp dụng các kỹ thuật tiên tiến về trí tuệ nhân tạo nhằm tạo ra những bài thơ đảm bảo chính xác về vần điệu và giọng điệu và thể thơ mà vẫn tiết kiệm thời gian hơn so với việc phải tự bỏ công sức vào để sáng tác. Tuy nhiên, các nghiên cứu này vẫn còn tồn đọng yếu tố khách quan và chủ quan liên quan đến vấn đề ngôn ngữ cũng như là ngữ cảnh của bài thơ được sinh ra trong quá trình huấn luyện.

Từ vấn đề được đặt ra ở trên cùng với những lý do khách quan và chủ quan từ phía người nghiên cứu, chúng em lựa chọn đề tài "Phát triển mô hình sinh thơ tiếng Việt sử dụng NLG" để tiến hành nghiên cứu, phân tích, làm rõ vấn đề. Bài nghiên cứu này chủ yếu thực hiện trên tập dữ liệu là ngôn ngữ Tiếng Việt với thể thơ là lục bát là chính.

1.2 Mục tiêu thực hiện đề tài

Mục tiêu của đề tài là thử nghiệm việc xây dựng các mô hình dựa trên mô hình Transformer đó là mô hình GPT-2. Việc huấn luyện sẽ dựa trên các khía cạnh như giọng điệu, vần của thơ lục bát từ đó đề xuất những mô hình theo các cách tiếp cận khác nhau bằng cách so sánh và đánh giá độ chính xác giữa các mô hình.

Kết quả đầu ra kỳ vọng cho dự án nghiên cứu là các mô hình được đề xuất cho từng cách tiếp cận khác nhau dựa trên việc so sánh, đánh giá và phân tích các mô hình thực nghiệm

1.3 Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu của đề tài là mô hình có sẵn GPT-2 và ứng dụng mô hình này cho bài toán sinh thơ.

Đối với phạm vi nghiên cứu, bài toán này tập trung nghiên cứu về thể thơ lục bát sao cho mô hình có thể sinh ra đúng bài thơ lục bát với ngữ nghĩa rõ ràng. Các tập dữ liệu về thơ lục bát được lấy từ trang thivien.net, facebook và github. Nội dung của tập dữ liệu được sử dụng cho mục đích huấn luyện và thực nghiệm mô hình.

1.4 Phương pháp nghiên cứu

Dự án được nghiên cứu theo hướng thực nghiệm trên các mô hình khác nhau có thể phát triển cho bài toán sinh thơ và ứng dụng mô hình học sẵn có của mô hình Transformer là GPT-2 để học giọng điệu, vần, thể thơ lục bát và sau đó fine-tune lại mô hình để có thể sinh ra một bài thơ có một ngữ cảnh rõ ràng, gắn kết và có đầy đủ tính chất của thể thơ lục bát.

1.5 Ý nghĩa thực tiễn đề tài

Bài dự án nghiên cứu hướng đến việc đề xuất một mô hình có khả năng sinh ra một bài thơ chuẩn theo thể thơ lục bát với văn phong đậm chất dân tộc với ngữ cảnh rõ ràng và gắn kết. Mặc dù, đã có không ít những nghiên cứu đề xuất mô hình sinh thơ , nhưng đa số đều hướng tới sử dụng văn phong thơ nước ngoài như tiếng

Anh để huấn luyện và còn khá ít các nghiên cứu sử dụng tập dữ liệu là thơ tiếng Việt, đặc biệt là đối với thể thơ lục bát. Do đó, khi đề xuất một số hướng tiếp cận trong bài dự án này cũng là hướng để cho các nghiên cứu khoa học sau này phát triển thêm.

Đối với thực tiễn, hiện nay có khá ít các sản phẩm liên quan đến việc tạo ra một bài thơ, và hầu như các sản phẩm này đều phục vụ cho các phong cách thơ nước ngoài. Có số ít các sản phẩm sinh được ra thơ Việt Nam nhưng văn phong không đậm chất dân tộc, ngữ cảnh rời rạc, không mang nhiều ý nghĩa và chỉ sinh ra được thể thơ tự do. Chính vì thế, việc phát triển một sản phẩm có khả năng sinh thơ đậm chất dân tộc với phong cách truyền thống cho người Việt là một vấn đề khá là mới mẻ nhưng khi hoàn chỉnh thì sẽ có thể thu hút được sự chú ý của đông đảo cộng đồng văn học Việt Nam hay những người chuyên chơi thơ như một thú vui trong cuộc sống.

1.6 Bố cục đề tài

Cấu trúc của dự án gồm 6 chương:

Chương 1: Mở đầu

Ở chương 1 sẽ giới thiệu về lý do chọn đề tài, mục tiêu thực hiện, đối tượng và phạm vi nghiên cứu, phương pháp nghiên cứu và ý nghĩa thực tiễn của đề tài.

Chương 2: Cơ sở lý thuyết và các nghiên cứu có liên quan

Chương 2 sẽ giải thích các lý thuyết được sử dụng trong đề tài này.

Chương 3: Mô hình đề xuất

Chương 3 sẽ đi sâu vào lý thuyết và giải thích về mô hình đề xuất, nhấn mạnh các phương pháp đã sử dụng trong dự án này.

Chương 4: Thực nghiệm

Trong chương này sẽ trình bày các phương pháp sử dụng và nêu rõ độ đo accuracy cho từng phương pháp.

Chương 5: Xây dựng ứng dụng

Chương 5 sẽ đi vào giới thiệu trang web, framework sử dụng và những thông tin hiển thị trên trang web.

Chương 6: Kết luận

Chương này sẽ tổng kết và thảo luận về kết quả của đề tài, phân tích những mặt hạn chế và đưa ra hướng phát triển trong tương lai.

CHƯƠNG 2 – CƠ SỞ LÝ THUYẾT VÀ CÁC NGHIÊN CỨU LIÊN QUAN

2.1 Kỹ thuật sinh ngôn ngữ tự nhiên (Natural Language Generation)

2.1.1 Giới thiệu chung về NLG

Sinh ngôn ngữ tự nhiên(Natural Language Genration – NLG) - một nhiệm vụ trong lĩnh vực xử lý ngôn ngữ tự nhiên(Natural Language Procressing – NLP) là quá trình tạo ra các cụm từ và câu có ý nghĩa dưới dạng ngôn ngữ tự nhiên với dữ liệu từ tri thức và thông tin được cung cấp trong một biểu diễn logic. Về bản chất, nó tự động tạo ra các đoạn văn có nội dung mô tả, tóm tắt hoặc giải thích lại một đoạn văn khác với văn phong giống với con người với tốc độ hàng nghìn trang mỗi giây.

NLG là một lĩnh vực nghiên cứu và công nghệ mới mẻ với nhiều ứng dụng thực tế. Một số ứng dụng phổ biến của NLG bao gồm:

- Tạo văn bản sáng tạo: NLG có thể được sử dụng để tạo ra các định dạng văn bản sáng tạo khác nhau, chẳng hạn như thơ, mã, kịch bản, bản nhạc, email, thư, v.v.
- Dịch ngôn ngữ: NLG có thể được sử dụng để dịch văn bản từ một ngôn ngữ sang ngôn ngữ khác.
- Tạo nội dung văn bản: NLG có thể được sử dụng để tạo nội dung văn bản cho các ứng dụng khác nhau, chẳng hạn như trang web, ứng dụng di động hoặc email.

Nếu như mà để phân biệt với hiểu ngôn ngữ tự nhiên(Natural Language Understanding – NLU) - một nhánh khác trong lĩnh vực xử lý ngôn ngữ tự nhiên thì trọng tâm của NLG đó là các sự lựa chọn còn NLU là quản lý giả thuyết.

Đối với NLU, mục tiêu của nó là có thể hiểu được ý nghĩa, ý định của văn bản hay giọng nói bất kỳ. Để đạt được điều này thì NLU sinh ra và quản lý nhiều giả thuyết về những gì người dùng có thể đang cố gắng diễn dạt. Sau đó, hệ thống đánh

giá và tinh chỉnh các giả thuyết này dựa trên nhiều yếu tố khác nhau như bối cảnh, ngữ pháp và kiến thức trước đó để hiểu được ý định trong câu văn hay giọng nói của người dùng.

Còn đối với NLG mục tiêu là tạo ra ngôn ngữ hay văn bản nên việc này khiến nó phải quan tâm đến với các lựa chọn như sau:

- Lựa chọn nội dung(Content Selection): Lựa chọn nội dung đề cập đến quá trình thông tin nào sẽ được chọn và loại trừ khỏi văn bản được tạo. Quá trình này tránh khiến người dùng choáng ngợp với những chi tiết thừa và đảm bảo thông tin được được hiểu dễ dàng, từ đó nâng cao trải nghiệm người dùng.
- Lựa chọn từ vựng (Lexical Selection): Hệ thống phải chọn từ vựng phù hợp nhất để diễn đạt các khái niệm cụ thể. Việc lựa chọn từ vựng đặc biệt quan trọng khi hệ thống NLG tạo ra văn bản đầu ra trong nhiều ngôn ngữ.
- Tổng hợp (Aggregation): Tổng hợp đề cập đến quá trình kết hợp hai hoặc nhiều đơn vị văn bản nhỏ hơn thành một đơn vị duy nhất, ngắn gọn hơn và mang tính thông tin hơn. Mỗi đơn vị văn bản có thể được diễn đạt trong các câu riêng biệt, nhưng trong nhiều trường hợp việc kết hợp này sẽ giúp dễ đọc hiểu hơn.
- Lập kế hoạch diễn đạt (Discourse Planning): Đây là một bước quan trọng vì nó quyết định đến cấu trúc đầu ra của văn bản được sinh. Nói đây là bước quan trọng vì cách tổ chức của văn bản đầu ra có thể ảnh hưởng đến đến cách người dùng đọc hiểu văn bản. Ví dụ như một văn bản đầu ra có sự gắn kết về nội dung, bố cục rõ ràng, thì người dùng sẽ đọc hiểu văn bản dễ hơn.
- Biểu thức tham chiếu (Referring Expression): Hệ thống phải xác định cách tham chiếu đến các đối tượng được thảo luận. Quá trình tạo ra biểu thức đề cập liên quan chặt chẽ đến việc lựa chọn từ vựng, vì nó cũng

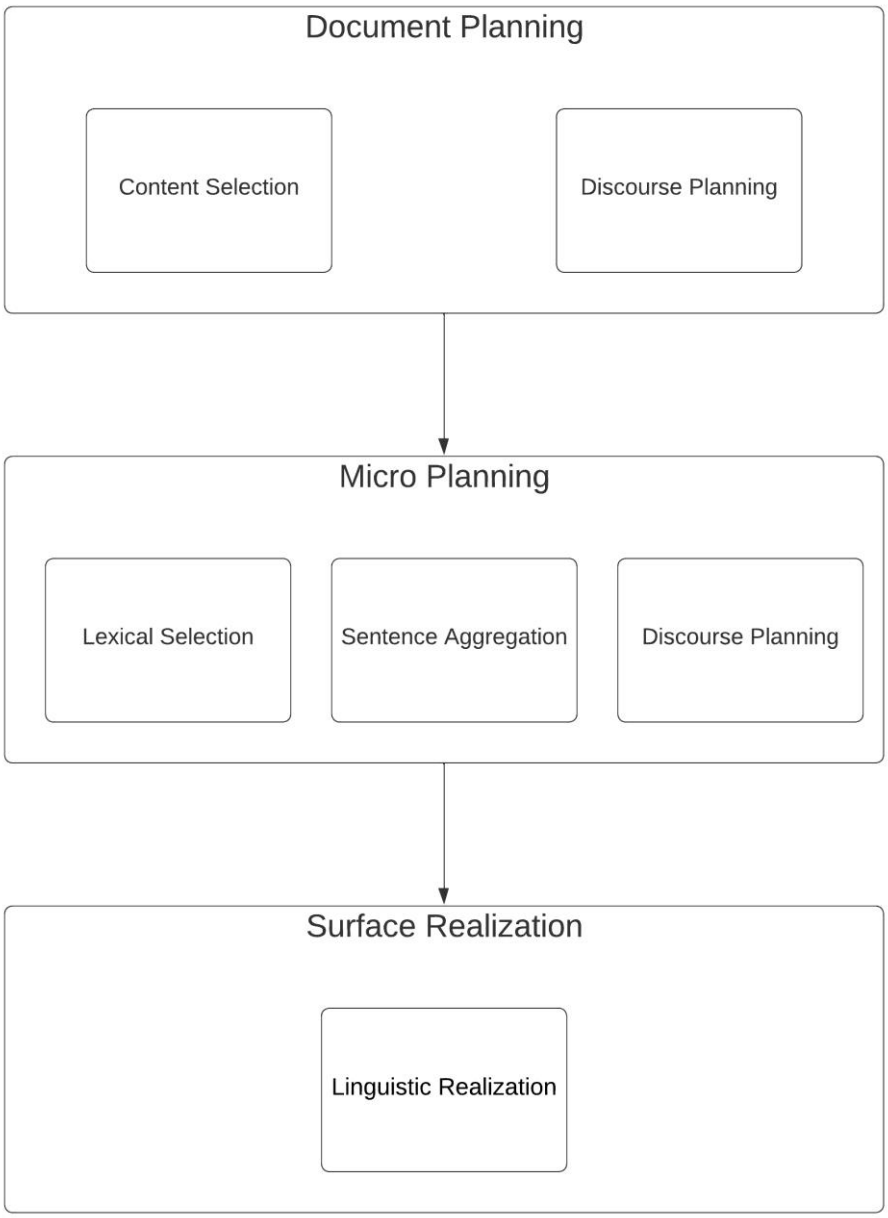
liên quan đến việc tạo ra các hình thức ngôn ngữ bề mặt nhằm xác định các yếu tố lĩnh vực.

- Hiện thực hóa ngôn ngữ (Linguistic Realization): Hiện thực hoá ngôn ngữ là quá trình áp dụng các quy tắc ngữ pháp để tạo ra một văn bản có đúng cú pháp, hình thái và chính tả.

2.1.2 Kiến trúc của NLG

Hệ thống NLG được phát triển với mục đích nhằm tái tạo các câu thoại giao tiếp tự nhiên, bắt chước ngữ điệu giọng nói của con người. Để đạt được điều này thì vào năm 2000, Reiter và Dale đã đề xuất một pipeline để mô tả kiến trúc của NLG bao gồm 3 công đoạn:

1. Lập kế hoạch tài liệu (Document planning): là một bước quan trọng trong kiến trúc NLG, đóng vai trò quyết định nội dung và là một bản phác thảo cho văn bản được tạo ra. Nó xác định cấu trúc và nội dung của văn bản, bao gồm các phần, đoạn, câu và từ với các công đoạn như lựa chọn nội dung(Content Selection) và lập kế hoạch diễn đạt (Discourse Planning).
2. Lập kế hoạch vi mô (Micro planning): Giai đoạn microplanning xác định nội dung cụ thể của từng phần, đoạn và câu bao gồm việc lựa chọn từ ngữ, cấu trúc câu và ngữ pháp. Cụ thể như lựa chọn từ vựng(Lexical Selection), tổng hợp câu(Sentence Aggregation), biểu thức tham chiếu(Referring Expression).
3. Hiện thực hóa (Surface realization): là giai đoạn cuối cùng trong kiến trúc NLG, chịu trách nhiệm chuyển các kế hoạch tài liệu (document plans) thành văn bản có thể đọc được. Nó bao gồm việc lựa chọn ngữ pháp phù hợp để thể hiện các ý tưởng của kế hoạch tài liệu.



Hình 2.1. Kiến trúc của NLG

Hình 1 cho thấy sơ đồ kiến trúc NLG gồm 3 giai đoạn. Document Planning tạo ra kế hoạch văn bản, được đưa cho Micro Planning. Micro Planning tạo ra kế hoạch đối thoại và Surface Realization áp dụng các quy tắc ngữ pháp để tạo ra đầu ra ngôn ngữ tự nhiên hợp lệ.

2.2 Các mô hình ngôn ngữ về Language Model Generation

2.2.1 N-gram

N-gram là một mô hình học máy dựa trên việc tính xác suất của các từ và các chuỗi để đưa ra dự đoán về từ hay câu tiếp theo chính xác nhất. Tên gọi N-gram ở đây có nghĩa là một chuỗi gồm n từ. N-gram có các loại chuỗi như chuỗi 2 từ được gọi là 2-gram hoặc bigram, ví dụ như "please turn", "turn your", "your homework", chuỗi 3 từ được gọi là 3-gram hoặc trigram, ví dụ như "I love you", "fill in blanks".

Trong thực tế, chúng ta có thể gặp mô hình N-gram trong tính năng gợi ý từ của các bàn phím điện thoại thông minh hay gợi ý từ gmail. Bằng cách phân tích dữ liệu từ người dùng, mô hình này được sử dụng để tạo ra các gợi ý cho từ tiếp theo trong câu.

Để làm rõ hơn về ý tưởng của N-gram. Lấy ví dụ như ta muốn tính xác xuất của từ "the" xuất hiện trong câu "its water is so transparent that" thì ta phải tìm trong bộ dữ liệu xem có tất cả bao nhiêu câu "its water is so transparent that" và bao nhiêu trong số đó có từ "the" đằng sau, sau đó tính xác suất:

$$P(\text{the} \mid \text{its water is so transparent that}) = \frac{C(\text{its water is so transparent that the})}{C(\text{its water is so transparent that})}$$

Nhưng đối với phương pháp đếm như này nếu như áp dụng với bộ dữ liệu rất lớn thì không khả thi. Hay ví dụ ta muốn tính xác suất của một chuỗi "its water is so transparent" thì ta phải chia số lần xuất hiện của chuỗi trên với tất cả số lần xuất hiện của một chuỗi 5 chữ bất kỳ trong bộ dữ liệu, điều này hoàn toàn không khả thi nếu bộ dữ liệu rất lớn.

Cho nên N-gram tính xác suất sử dụng quy tắc dây chuyền(chain rule), cho phép tính toán xác suất của một chuỗi từ, được xác định bởi các xác suất của các n-gram hay chuỗi n từ riêng lẻ tạo nên chuỗi đó, quy tắc này được biểu diễn như sau:

$$P(w_1 \ldots w_n) = P(w_1)P(w_2|w_1)P(w_3|w_1w_2)\ldots P(w_n|w_1 \ldots w_{n-1})$$

trong đó $P(w_1 \ldots w_n)$ là xác suất của chuỗi từ $w_1 \ldots w_n$,
$P(w_n|w_1 \ldots w_{n-1})$là xác suất của từ w_i cho biết lịch sử $w_1 \ldots w_{n-1}$.

Ví dụ, để tính xác suất của chuỗi từ “its water is so transparent”, ta áp dụng chain rule như sau:

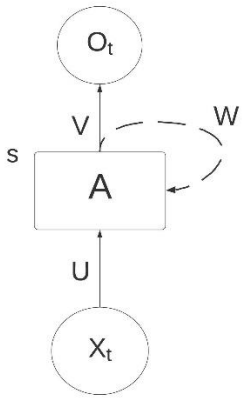
P(“its water is so transparent”) = P(“its)P(“water|its”)P(“is|water its”)P(“so|is water its”)P(“transparent|so is water its”).

Tuy nhiên, chain rule cũng có một số hạn chế. Ví dụ, chain rule giả định rằng các từ trong một chuỗi là độc lập với nhau. Điều này không phải lúc nào cũng đúng, vì các từ có thể phụ thuộc lẫn nhau theo nhiều cách khác nhau.

2.2.2 RNN

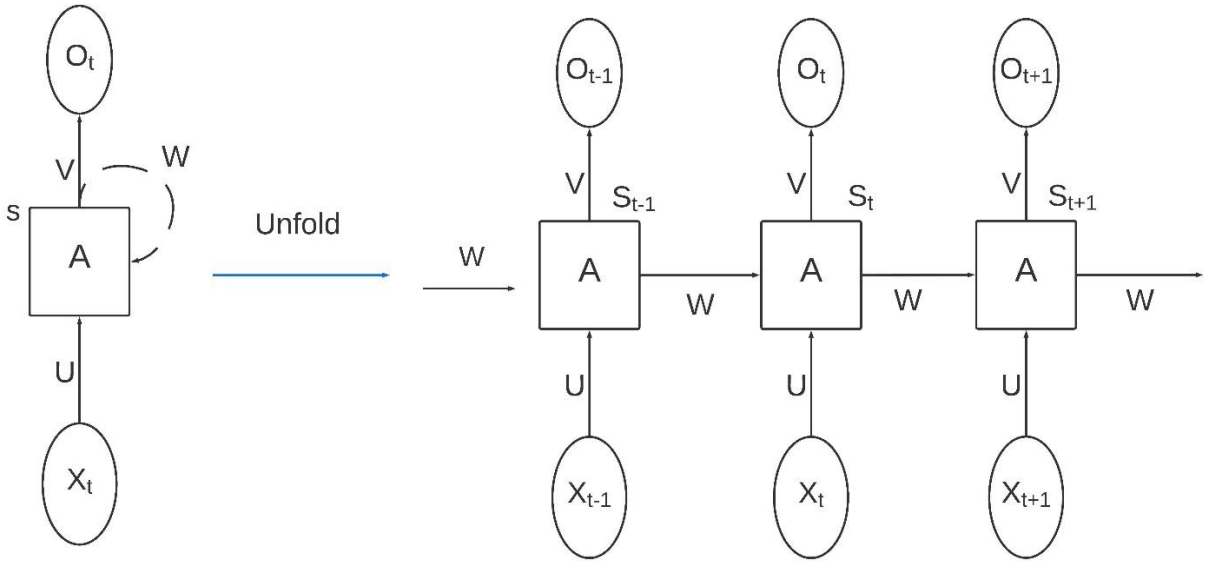
Trong các mạng nơ-ron truyền thống tất cả các đầu vào và cả đầu ra là độc lập với nhau, tức là chúng không liên kết thành chuỗi với nhau. Tuy nhiên nếu xét từng từ một đứng riêng lẻ ta không thể hiểu được nội dụng của toàn bộ câu mà phải dựa trên những từ xung quanh mới có thể hiểu được trọn vẹn một câu nói, vậy nên mạng nơ ron truy hồi hay Recurrent Neural Network đã được thiết kế đặc biệt để giải quyết vấn đề này.

Nói về định nghĩa chung, một mạng nơ ron truy hồi là bất kỳ một hệ thống mạng nào tồn tại một chu trình trong các kết nối mạng của nó, nghĩa là giá trị của một đơn vị nào đó phụ thuộc trực tiếp hoặc gián tiếp vào đầu ra trước đó của chính nó như một đầu vào tiếp theo.



Hình 2.2. Mạng nơ ron truy hồi với vòng lặp

Hình 2 trên biểu diễn kiến trúc của một mạng nơ ron truy hồi. Chu trình A ở thân mạng nơ ron là điểm mấu chốt trong nguyên lý hoạt động của mạng neural truy hồi. Đây là chu trình sao chép nhiều lần của cùng một kiến trúc nhằm cho phép các thành phần có thể kết nối liền mạch với nhau theo mô hình chuỗi. Đầu ra của vòng lặp trước chính là đầu vào của vòng lặp sau. Nếu trải phẳng thân mạng nơ ron A ta sẽ thu được một mô hình dạng:



Hình 2.3. Cấu trúc trải phẳng của mạng nơ ron hồi quy

Mô hình trên mô tả phép triển khai nội dung của một RNN. Việc tính toán bên trong RNN được thực hiện như sau:

- X_t là đầu vào tại bước t. Ví dụ, X_1 là một vec-tơ one-hot tương ứng với từ thứ 2 của câu bất kỳ.
- S_t là trạng thái ẩn tại bước t. S_t được tính toán dựa trên các trạng thái ẩn phía trước và đầu vào tại bước đó. $S_t = f(U_{X_t} + W_{S_{t-1}})$. Hàm f thường là một hàm phi tuyến tính như hàm tanh hoặc ReLU.
- O_t là đầu ra tại bước t, ta có thể coi O_t như một véc-tơ xác suất các từ có thể xuất hiện lấy ví dụ bài toán dự đoán từ tiếp theo ở trong một câu.

Kiến trúc mạng nơ ron truy hồi được sử dụng rộng rãi trong lĩnh vực xử lý ngôn ngữ tự nhiên như dịch máy, tạo văn bản, nhận diện giọng nói,…

2.2.3 LSTM

2.2.4 Bi-LSTM

2.2.5 Transformer

CHƯƠNG 3 – MÔ HÌNH ĐỀ XUẤT

CHƯƠNG 4 – THỰC NGHIỆM

CHƯƠNG 5 – XÂY DỰNG ỨNG DỤNG

CHƯƠNG 6 – KẾT LUẬN

(Các phần trống sẽ được cập nhật sau xuyên suốt quá trình thực hiện dự án)

TÀI LIỆU THAM KHẢO

PHỤ LỤC

Phần này bao gồm những nội dung cần thiết nhằm minh họa hoặc hỗ trợ cho nội dung luận văn như số liệu, biểu mẫu, tranh ảnh. . . . nếu sử dụng những câu trả lời cho một *bảng câu hỏi thì bảng câu hỏi mẫu này phải được đưa vào phần Phụ lục ở dạng nguyên bản* đã dùng để điều tra, thăm dò ý kiến; **không được tóm tắt hoặc sửa đổi**. Các tính toán mẫu trình bày tóm tắt trong các biểu mẫu cũng cần nêu trong Phụ lục của luận văn. Phụ lục không được dày hơn phần chính của luận văn