ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

NGUYỄN NGỌC KHƯƠNG

NGHIÊN CỨU CÁC MÔ HÌNH HỌC SINH CHUỖI TỪ CHUỖI SỬ DỤNG HỌC SÂU VÀ ỨNG DỤNG TRONG XỬ LÝ NGÔN NGỮ TỰ NHIÊN

Chuyên ngành: Khoa học máy tính
Mã số: 9480101.01

TÓM TẮT LUẬN ÁN TIẾN SĨ KHOA HỌC MÁY TÍNH

Hà Nội - 2022

Công trình được hoàn thành tại: Trường Đại học Công nghệ, Đại học Quốc Gia Hà Nội.

Người hướng dẫn khoa học:

1. PGS.TS.Nguyễn Việt Hà
2. PGS.TS. Lê Anh Cường

Mở đầu

Đối với các bài toán xử lý ngôn ngữ tự nhiên, một văn bản đầu vào chứa các mức độ ngữ nghĩa khác nhau như mức từ, mức câu, mức đoạn, mức toàn bộ văn bản. Hơn nữa các thành phần này trong văn bản có quan hệ với nhau rất đa nghĩa, ví dụ mỗi từ sẽ có ngữ nghĩa khác nhau khi ở trong các ngữ cảnh khác nhau. Vì vậy phát triển các mô hình học máy cho nhiệm vụ encoding một văn bản sao cho vec-tơ biểu diễn chứa đầy đủ và chính xác, phản ánh đúng văn bản đầu vào luôn là bài toán thách thức trong lĩnh vực nghiên cứu NLP. Đối với bộ giải mã, nhiệm vụ là sinh ra chuỗi đầu ra dựa trên một mục tiêu nhất định, ví dụ như sinh câu trả lời trong bài toán Chatbot sẽ khác trong bài toán tóm tắt văn bản. Một mô hình học máy tốt sẽ phải giải quyết vấn đề sử dụng một cách phù hợp thông tin đầu vào và thoả mãn nội dung đầu ra, vì vậy đây cũng luôn là vấn đề thách thức đối với bộ giải mã. Trong luận án này, chúng tôi tập trung phát triển các mô hình Seq2seq để giải quyết các vấn đề nêu trên.

Với mục tiêu đó, luận án tập trung nghiên cứu đều xuất các phương pháp nhằm tối ưu hoá việc mã hoá thông tin văn bản đầu vào, dựa trên việc mã hoá cấu trúc ngữ nghĩa phân cấp của văn bản. Chúng tôi cũng đồng thời phát triển mô hình sinh văn bản dựa trên việc sử dụng cơ chế chú ý (attention) kết hợp với mô hình hoá sự ràng buộc của chuỗi đầu ra. Chúng tôi phát triển các mô hình học sâu Seq2seq cho hai bài toán: bài toán thứ nhất là bài toán diễn giải (paraphasing) một văn bản đầu vào theo một cách diễn giải mới; bài toán thứ hai là tóm tắt văn bản theo tiếp cận tóm lược (abstractive text summarization).

Kết quả thực nghiệm cho bài toán diễn giải văn bản trên hai kho dữ liệu phổ biến cho thấy mô hình đã giải quyết được các giả thiết vai trò của biểu

diễn phân cấp có vai trò quan trọng đối với các văn bản dài trong bài toán diễn giải. Bên cạnh đó biểu diễn dữ liệu theo chiều sâu với các mức biểu diễn ngữ nghĩa khác nhau cũng đã chứng minh được tính hiệu quả trong quá trình sinh diễn giải của văn bản đầu vào. Đối với bài toán tóm tắt tóm lược, luận án đề xuất mô hình biểu diễn ngữ cảnh hai phía trong mối quan hệ mức từ và mức câu đối với văn bản đầu vào tại pha mã hoá để cải thiện chất lượng sinh tóm tắt tóm lược. Hiểu bản chất của văn bản đầu vào là yếu tố quan trọng quyết định đến chất lượng đầu ra của văn bản tóm tắt, cơ chế chú ý toàn cục chú trọng đến vai trò của từng thành phần trong văn bản đầu vào trên toàn bộ ngữ cảnh, trong khi đó cơ chế chú ý cục bộ đề cập đến vai trò của từng thành phần trong từng ngữ cảnh cụ thể. Luận án cũng để xuất mô hình kết hợp hai cơ chế chú ý trên để cải thiện chất lượng sinh tóm tắt tóm lược của mô hình đặc biệt đối với các văn bản đầu vào. Trong tóm tắt nói chung và tóm tắt tóm lược nói riêng, độ dài của bản tóm tắt là một yếu tố quan trọng khác trong phương diện nghiên cứu và ứng dụng. Chúng tôi cũng nghiên cứu đề xuất mô hình tích hợp ràng buộc độ dài tổng quát trong mô hình sinh chuỗi từ chuỗi thích hợp cho bài toán sinh tóm tắt tóm lược có giới hạn độ dài.

Chương 1

Tổng quan các vấn đề liên quan luận án

1.1 Bối cảnh

Bài toán sinh chuỗi $y_1, ..., y_m$ từ chuỗi $x_1, ..., x_n$ có thể được mô hình hoá thành hàm phân phối xác suất có điều kiện như sau:

$$p(y_1, ..., y_m | x_1, ..., x_n) = \prod_{j=1}^{m} p(y_i | y_1, ..., y_{j-1}, c) \tag{1.1.1}$$

Trong vế phải của công thức trên, mỗi phân bố $p(y_i|y_1, ..., y_{j-1}, c)$ mô tả xác suất xuất hiện của từ y_j với véc tơ đại diện cho câu đầu vào c và các từ trong chuỗi đầu ra đứng trước nó. Phân bố này được biểu diễn bằng một hàm **softmax** trên tất cả các từ trong tập từ vựng ở ngôn ngữ đích. Công thức trên có thể được viết lại thành dạng như sau:

$$\log p(x|y) = \sum_{j=1}^{m} \log p(y_j | y_{<j}, c) \tag{1.1.2}$$

Mỗi token y_j có xác suất xuất hiện được tính như sau:

$$p(y|y_{j<s}, s) = softmax(g(h_j)) \tag{1.1.3}$$

Trong đó g là hàm dùng để biến đổi trạng thái ẩn h_j của Giải mã tại bước giải mã tương ứng thành vector có kích thước bằng kích thước của tập từ vựng

trong ngôn ngữ đích. Trạng thái ẩn h_j được tính như sau:

$$h_j = f(h_{j-1}, s) \tag{1.1.4}$$

Trong đó f là hàm biểu diễn chung cho quá trình tính trạng thái ẩn tại bước hiện tại từ trạng thái ẩn đầu ra của bước trước bằng mạng nơ ron.

Mô hình sinh chuỗi từ chuỗi dựa trên kiến trúc mã hoá giải mã được trình bày ở trên tuy đã giải quyết bài toán chuyển hóa chuỗi đầu vào thành chuỗi đầu ra có độ dài khác nhau trên cùng hoặc khác ngôn ngữ, tuy nhiên nó tồn tại một số hạn chế như sau:

- Đầu tiên, dễ thấy nhất đó là việc sử dụng bộ mã hoá duyệt qua từng phần từ của chuỗi đầu vào và rồi lấy ra véc tơ trạng thái ẩn của mạng này ở thời điểm cuối cùng, và hi vọng rằng nó sẽ nhớ hết những thông tin cần thiết của chuỗi đầu vào trước khi chuyển hóa thành chuỗi đầu ra, điều này không phải là điều luôn khả thi. Với những chuỗi dài, sau khi duyệt qua hàng loạt các phần thì thông tin ở những phần đầu sẽ bị "quên", và đôi khi lại nhớ những thứ không cần nhớ.

- Thứ hai, các mô hình sinh chuỗi từ chuỗi dựa trên kiến trúc mã hoá giải mã sử dụng mạng nơ ron thường yêu cầu tài nguyên tính toán khá lớn để có thể huấn luyện để tối ưu mô hình.

- Kế tiếp, các hoạt động bên trong các mô hình sinh chuỗi từ chuỗi có thể khó diễn giải một cách tường minh, điều này có thể gây khó khăn trong việc giải thích lý do tại sao mô hình có thể sinh ra các trạng thái đầu ra nhất định.

- Bên cạnh đó, việc sử dụng các kỹ thuật huấn luyện mô hình dựa trên mạng nơ ron thường có đặc điểm quá khớp với dữ liệu đã huấn luyện nhưng thường kém hiệu quả trên dữ liệu mới.

- Thêm nữa, đối với các mô hình sinh nói chung và sinh văn bản nói riêng thì khó khăn trong việc xử lý các từ hiếm không có trong dữ liệu huấn luyện cũng là một thách thức đặt ra đối với các mô hình sinh chuỗi từ chuỗi.

1.2 Mục tiêu nghiên cứu

Trước những thách thức trên, ba câu hỏi nghiên cứu được đặt ra trong luận án bao gồm:
- Câu hỏi 1. Trong các mô hình học sinh chuỗi từ chuỗi, việc học biểu diễn của chuỗi đầu vào dựa trên các thành phần cơ sở là từ thì việc học biểu diễn các cụm, các câu, các đoạn trong chuỗi đầu vào đóng vai trò như thế nào với quá trình sinh ra chuỗi đầu ra?
- Câu hỏi 2. Quá trình sinh chuỗi đầu ra trong mô hình sinh chuỗi từ chuỗi, xem xét vai trò của các từ, câu, cụm trong chuỗi đầu vào trong phạm vi cục bộ hay toàn cục sẽ cho kết quả tốt hơn?
- Câu hỏi 3. Với mô hình sinh chuỗi từ chuỗi tổng quát có khả năng tích hợp yếu tố ràng buộc độ dài trong quá sinh ra chuỗi đầu ra với giới hạn độ dài khác nhau hay không?

Mục tiêu của luận án là nghiên cứu và đề xuất những kết quả lý thuyết cũng như các thuật toán, mô hình nhằm đưa ra những câu trả lời khẳng định cho ba câu hỏi trên.

1.3 Nhiệm vụ nghiên cứu

Để đạt được mục tiêu đề ra, nhiệm vụ nghiên cứu tập trung giải quyết các vấn đề chính sau đây:

- Khảo sát, nghiên cứu các mô hình học sinh chuỗi từ chuỗi, khả năng áp dụng của mô hình trong các ứng dụng thực tiễn trong lĩnh vực xử lý ngôn ngữ tự nhiên.

- Nghiên cứu, phân tích các phương pháp biểu diễn văn bản nói chung và mô hình véc tơ nói riêng, từ đó đề xuất mô hình biểu diễn văn bản trong mô hình học sinh chuỗi từ chuỗi cho hai bài toán sinh tóm tắt trừu tượng và sinh diễn giải văn bản.

- Nghiên cứu, đánh giá các cơ chế chú ý cho bài toán sinh văn bản, đề xuất kỹ thuật chú ý cho mô hình học sinh chuỗi từ chuỗi phù hợp với đặc trưng của bài toán sinh tóm tắt trừu tượng và sinh diễn giải văn bản.

- Khảo sát các mô hình ràng buộc độ dài trong bài toán giới hạn độ dài tóm tắt trừu tượng, đề xuất mô hình giới hạn mềm độ dài cho mô hình học sinh chuỗi từ chuỗi cho bài toán sinh tóm tắt trừu tượng.

- Triển khai thực nghiệm và đánh giá kết quả.

1.4 Đóng góp của Luận án

- Đề xuất phương pháp biểu diễn phân cấp văn bản trong mô hình học sinh chuỗi từ chuỗi cho bài toán sinh tóm tắt tóm lược. Đóng góp này được công bố trong kỷ yếu hội thảo Knowledge and Systems Engineering năm 2021.

- Đề xuất cơ chế chú ý trong mô hình học sinh chuỗi từ chuỗi cho bài toán sinh diễn giải văn bản. Đóng góp này được công bố trong kỷ yếu hội thảo International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making năm 2018.

- Đề xuất cơ chế chú ý phân cấp có điều kiện trong mô hình học sinh chuỗi từ chuỗi cho bài toán sinh diễn giải. Đóng góp này được công bố trong kỷ yếu hội thảo Multi-disciplinary International Conference on Artificial Intelligence năm 2018.
- Đề xuất cơ chế chú ý cục bộ thích hợp cho bài toán sinh tóm tắt tóm lược văn bản. Đóng góp này được trình bày tại hội thảo "Asia Pacific Information Technology Conference lần thứ 5"
- Đề xuất mô hình học sinh chuỗi từ chuỗi cho bài toán sinh tóm tắt tóm lược có ràng buộc độ dài. Đóng góp này được chấp nhận đăng trong tạp chí "Jounal of Intelligent Automation & Soft Computing"

Chương 2

Kiến thức cơ sở

2.1 Mạng nơ ron

Phần này cung cấp một cái nhìn tổng quan về mạng nơ-ron nhân tạo, với sự nhấn mạnh vào ứng dụng vào các nhiệm vụ phân loại và ghi nhãn.

2.2 Các biến thể của Mạng hồi quy

2.3 Mô hình ngôn ngữ dựa trên kỹ thật học sâu

2.4 Mô hình học sinh chuỗi từ chuỗi

2.4.1 Phát biểu bài toán

Mô hình học sinh chuỗi từ chuỗi sử dụng mạng nơron nhiều tầng là một mô hình học sâu với mục đích tạo ra một chuỗi đầu ra từ một chuỗi đầu vào (lưu ý độ dài của hai chuỗi này có thể khác nhau). Mô hình này được đề xuất bởi Sutskever [2] và cộng sự tại Google vào năm 2014. Cho dù mục đích ban đầu của mô hình này là để áp dụng cho bài toán dịch máy [6], tuy nhiên hiện nay mô hình này được áp dụng cho nhiều bài toán khác như: nhận dạng tiếng nói [4], tóm tắt văn bản [3], sinh diễn giải ảnh [8],... Một cách tổng quát, mô

hình học sinh chuỗi từ chuỗi có thể được phát biểu như sau:

Cho chuỗi đầu vào $x = x_1, x_2, ..., x_n$ và chuỗi đầu ra $y = y_1, y_2, ..., y_m$, trong đó $x_t \in S_x, y_u \in S_y$, và S_x, S_y là tập các khả năng có thể cho mỗi cặp x_t và y_t tương ứng. Giả sử, đầu vào và đầu ra của mô hình là các biến ngẫu nhiên, các giá trị n và m phụ thuộc vào từng cặp chuỗi đầu vào, đầu ra cụ thể.

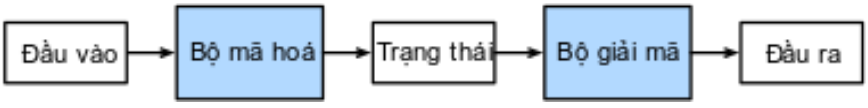
Giả sử, mô hình sinh luôn sinh được chuỗi y cho mỗi chuỗi x dựa trên phân phối xác suất có điều kiện $p(y|x)$, ký hiệu $y = f(x)$. Nhiệm vụ của quá trình huấn luyện các mô hình học sinh chuỗi từ chuỗi là tìm được hàm θ để cực đại hoá xác suất có điều kiện $p(y|x) : y' = \arg\max_y p(y|x, \theta)$.

Với mỗi hệ thống sinh ngôn ngữ dựa trên mô hình học sinh chuỗi từ chuỗi, chúng ta cần trả lời ba câu hỏi sau:

- Mô hình hoá $p(y|x, \theta)$
- Cách tìm tham số θ
- Cách sinh ra đầu ra y

2.4.2 Kiến trúc mã hoá - giải mã chuẩn

Kiến trúc mã hoá - giải mã là mô hình hoá tiểu chuẩn cho các tác vụ sinh chuỗi từ chuỗi. Kiến trúc mã hoá - giải mã tổng quát được mô tả trong hình 2.1. Các thành phần chính của mô hình bao gồm:



Hình 2.1: Mô hình sinh chuỗi từ chuỗi tổng quát.

- Bộ mã hoá được sử dụng để ánh xạ chuỗi token trong ngôn ngữ nguồn đầu vào thành một vector có kích thước cố định. Tại mỗi bước mã hóa, bộ mã hoá sẽ nhận vector tương ứng với mỗi token trong chuỗi đầu vào

để tạo ra vector biểu diễn trung gian đại diện cho chuỗi đầu vào tại bước mã hóa cuối cùng.

- Bộ giải mã sử dụng vector biểu như khởi tạo cho trạng thái ẩn đầu tiên và tạo ra chuỗi các token ở ngôn ngữ đích tại mỗi bước giải mã. Do đó, hàm xác suất có điều kiện có thể được phân tích như sau:

$$p(y_1, ..., y_m|x_1, ..., x_n) = \prod_{j=1}^{m} p(y_i|y_1, ..., y_{j-1}, c) \tag{2.4.1}$$

Công thức trên có thể được viết lại thành dạng như sau:

$$\log p(x|y) = \sum_{j=1}^{m} \log p(y_j|y_{<j}, c) \tag{2.4.2}$$

Mỗi token y_j có xác suất xuất hiện được tính như sau:

$$p(y|y_{j<s}, s) = softmax(g(h_j)) \tag{2.4.3}$$

Trong đó g là hàm dùng để biến đổi trạng thái ẩn h_j của Giải mã tại bước giải mã tương ứng thành vector có kích thước bằng kích thước của tập từ vựng trong ngôn ngữ đích. Trạng thái ẩn h_j được tính như sau:

$$h_j = f(h_{j-1}, s) \tag{2.4.4}$$

Trong đó f là hàm biểu diễn chung cho quá trình tính trạng thái ẩn tại bước hiện tại từ trạng thái ẩn đầu ra của bước trước bằng mạng RNN hoặc bằng những cải tiến khác như LSTM và GRU. Trong mô hình của Sutskever và cộng sự [7] vector s đại diện cho câu nguồn chỉ được sử dụng một lần để làm trạng thái ẩn đầu tiên cho bộ giải mã. Trong mô hình của tác giả Bahdanau và cộng sự [1] và của tác giả Luong và cộng sư [5] s là một vector đặc biệt được sử dụng xuyên suốt tại mỗi bước trong quá trình giải mã.

Chương 3

Mô hình học sinh chuỗi từ chuỗi cho bài toán sinh diễn giải

3.1 Cơ chế chú ý toàn cục cho bài toán diễn giải văn bản

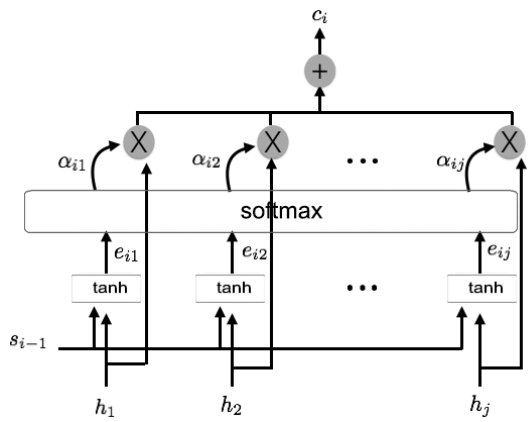
3.1.1 Mô hình đề xuất

Đề xuất của chúng tôi lấy ý tưởng trong nghiên cứu để giải quyết bài toán dịch máy của Wu và cộng sự [9]. Mô hình gồm 03 thành phần: bộ mã hoá, bộ giải mã và mạng chú ý. Bộ mã hoá sử dụng 4 lớp mạng LSTM trong đó có 1 lớp mạng LSTM hai hướng và 2 lớp mạng LSTM một hướng. Lớp mạng LSTM hai hướng được đặt là lớp đầu tiên để có thể biểu diễn dữ liệu đầu vào theo hai hướng. Véc tơ chú ý c_t (trong hình 3.1) được tính thông qua điểm liên quan α_{ti} trên mỗi trạng thái ẩn h_i như sau:

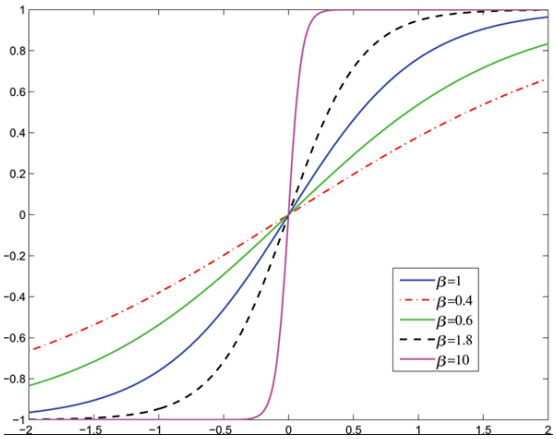
$$c_t = \sum_{i=0}^{T} \alpha_{ti} h_i \tag{3.1.1}$$

Giá trị của α_{ti} thể hiện mức độ liên quan của các đơn vị trong văn bản nguồn tới quá trình sinh các thành phần trong văn bản đầu ra và được tính như sau:

$$\alpha_{ti} = \frac{exp(e_{ti})}{\sum_{k=0}^{T} exp(e_{tk})} \tag{3.1.2}$$



Hình 3.1: Cơ chế chú ý.



Hình 3.2: Hình dáng hàm *tanh*.

trong đó e_{ti} được gọi là trọng số chú ý và được tính thông qua mạng nơ ron f như sau:

$$e_{ti} = f((W_a * s_{t-1} + U_a * h_i)) \tag{3.1.3}$$

trong đó f thường sử dụng hàm *tanh* với tham số mặc định β là 1 như đường liền màu xanh nước biển trong hình 3.2.

Trên thực tế, với bài toán sinh diễn giải đôi khi chỉ đơn giản là việc diễn đạt lại hoặc thay thế một số từ có vai trò quan trọng trong văn bản nguồn ta đã có thể sinh ra một bản diễn giải mới cho văn bản đáp ứng được yêu cầu, mong muốn đặt ra. Điều đó có nghĩa một số từ, cụm từ, thành phần trong văn bản nguồn không có vai trò trong quá trình sinh diễn giải văn bản. Để giải quyết vấn đề này, chúng tôi thêm vào một tham số mới β cho hàm *tanh* được sử dụng theo công thức sau:

$$e_{ti} = f(\beta * (W_a * s_{t-1} + U_a * h_i)) \tag{3.1.4}$$

Mục tiêu của tham số β là loại bỏ vai trò của một số từ, cụm từ hoặc thành phần trong văn bản đầu vào không có ý nghĩa trong quá trình diễn giải văn bản(tương ứng với giá trị của hàm*tanh* là -1). Điều này làm trực tiếp thay đổi giá trị của trọng số chú ý e_{ti}, ta gọi β là hệ số phạt (Penalty Coefficient) và giá trị chú ý dựa trên việc bổ sung hệ số phạt β được gọi là Hệ số phạt chú

ý(Penalty Coefficient Attention - PCA).

3.1.2 Thực nghiệm

Bảng 3.1: Kết quả thực nghiệm trên kho dữ liệu PPDB

		Kích thước Beam = 5				*Kích thước Beam = 10*			
Số lớp	**Mô hình**	**BLEU**	**METEOR**	**Emb Greedy**	**TER**	**BLEU**	**METEOR**	**Emb Greedy**	**TER**
2	Sequence to Sequence	12.50	21.30	32.55	**82.90**	12.90	20.50	32.65	**83.00**
	Seq2Seq with Attention	13.00	21.20	32.95	82.20	13.80	20.60	32.29	81.90
4	Sequence to Sequence	18.30	**23.50**	33.18	82.70	18.80	23.50	33.78	82.10
	Bi-directionalLSTM	19.20	23.10	34.39	77.50	19.70	23.20	34.56	84.40
	Seq2Seq with Attention	19.90	23.20	34.71	83.80	20.20	22.90	**34.90**	77.10
	Mạng LSTM thặng dư	20.30	23.10	34.77	77.10	21.20	23.00	34.78	77.00
	PCA-LSTM	**20.57**	23.30	**34.82**	76.60	**21.65**	**23.60**	34.80	76.40

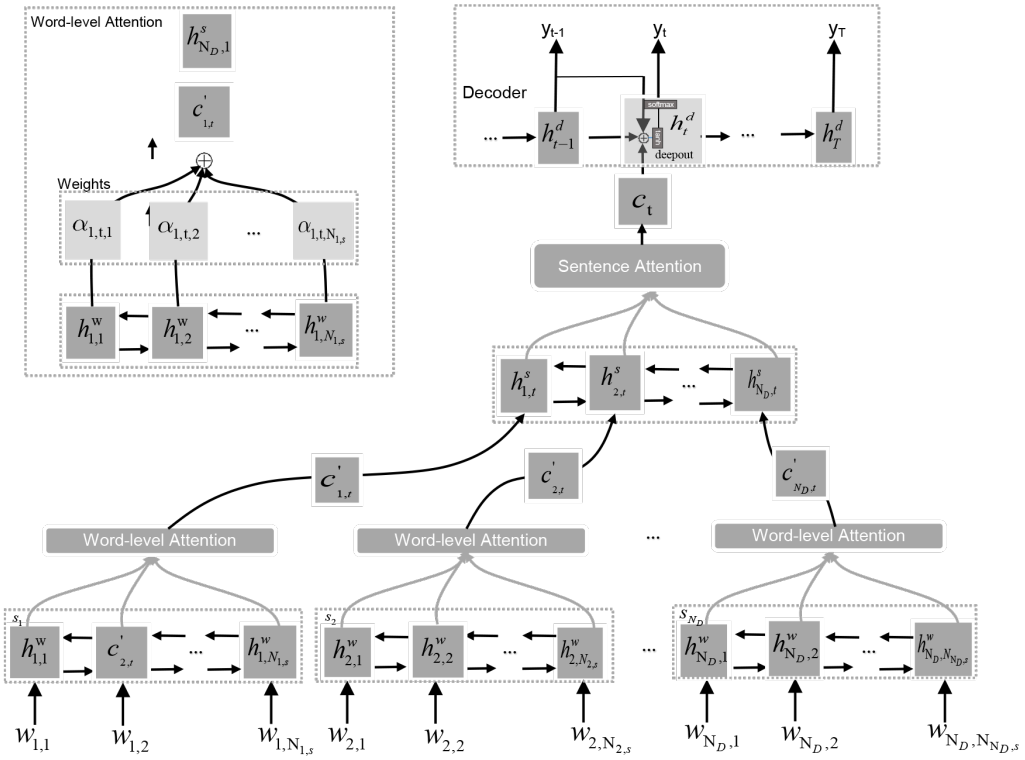
Bảng 3.2: Kết quả thực nghiệm trên kho dữ liệu WikiAnswer

		Kích thước Beam = 5				*Kích thước Beam = 10*			
Số lớp	**Mô hình**	**BLEU**	**METEOR**	**Emb Greedy**	**TER**	**BLEU**	**METEOR**	**Emb Greedy**	**TER**
2	Sequence to Sequence	19.20	26.10	62.65	**35.10**	19.50	26.20	62.95	34.80
	Seq2Seq with Attention	21.20	22.90	63.22	37.10	21.20	23.00	63.50	**37.00**
4	Sequence to Sequence	33.20	29.60	73.17	28.30	33.50	29.60	73.19	28.30
	Bi-directionalLSTM	34.00	30.80	73.80	27.30	34.30	30.70	73.95	27.00
	Seq2Seq with Attention	34.70	31.20	73.45	27.10	34.90	31.20	73.50	27.10
	Mạng LSTM thặng dư	37.00	**32.20**	75.13	27.00	37.20	**32.20**	75.19	27.80
	PCA-LSTM	**37.23**	32.16	**75.85**	26.85	**37.80**	31.60	**76.25**	27.70

3.2 Cơ chế chú ý phân cấp cho bài toán diễn giải văn bản

3.2.1 Mô hình đề xuất

Kiến trúc tổng quát của mô hình đề xuất được trình bày trong hình 3.3 với đầu vào là văn bản D. Kiến trúc tổng quát bao gồm ba thành phần chính: bộ mã hoá văn bản D bao gồm 2 bộ mã hoá tương ứng với hai mức từ và câu; thành phần chú ý phân cấp cho mức từ và mức câu; bộ giải mã. Trong đó dạng biểu diễn đầu của bộ mã hoá và trạng thái ẩn hiện tại của bộ giải mã được sử dụng để tính xác suất có điều kiện $p(y_t|D, Y_{t-1})$ trong đó $Y_{t-1} = (y_1, y_2, ..., y_{t-1})$



Hình 3.3: Kiến trúc mạng với cơ chế chú ý phân cấp.

Mã hoá và cơ chế chú ý phân cấp: Mô hình đề xuất được thể hiện trong hình 3.3, bao gồm hai bộ mã hoá có quan hệ với nhau. Bộ mã hoá mức từ có chức năng chuyển hoá chuỗi đầu vào các từ trong văn bản đầu vào ($w_{i,1}$,..., $w_{i,N_{i,s}}$) thành chuỗi trạng thái ẩn mức từ ($h^w_{i,1}$,..., $h^w_{i,N_{i,s}}$). Bộ mã hoá thứ hai có chức năng chuyển các biểu diễn mức câu ($c^{'}_{1,t}$,..., $c^{'}_{N_D,t}$) thành chuỗi các trạng thái ẩn mức câu ($h^s_{1,t}, ..., h^s_{N_D,t}$), chuỗi trạng thái ẩn này sẽ được sử dụng để xác định các trọng số chú ý mức và mức câu trong mô hình đề xuất.

3.2.2 Thực nghiệm

Kết quả thực nghiệm cho thấy hiệu suất của mô hình HCANN tốt hơn hầu hết các mô hình trong thực nghiệm ở các độ đo(BLEU, TER and METEOR). Đặc biệt ở độ đo Emb Greedy, mô hình đề xuất cho kết quả tốt hơn đáng kể so với các mô hình khác, ngoại trừ PCA-LSTM khi kích thước tìm kiếm beam

là 5. Kết quả thu được có thể được giải thích thông qua việc quan sát độ dài các cặp dữ liệu trên các kho dữ liệu. Trong khi đối với kho dữ liệu `PPDB 2.0` chủ yếu chứa các cụm diễn giải ngắn và kho dữ liệu `WikiAnswer` chủ yếu chứa các cụm diễn giải dài, cơ chế chú ý phân cấp biểu diễn được mối quan hệ giữa các thành phần cơ sở của văn bản từ từ, đến cụm đếm câu nên cho kết quả khả quan hơn các mô hình khác đặc biệt trên các kho dữ liệu chứa các cụm diễn giải dài. Kết quả thực nghiệm cho thấy mô hình HCANN phù hợp với các các văn bản chứa cụm diễn giải dài. Trên hết, những kết quả thử nghiệm này cũng cho thấy rằng hiệu suất của các mô hình dựa trên cơ chế chú ý phân cấp có điều kiện đáp ứng tốt hơn đối với các bài toán sinh văn bản.

3.3 Kết luận chương

Mô hình đề xuất cùng với các kết quả thực nghiệm đã được cống bố trong kỷ yếu hội thảo quốc tế IUKM 2018. Vớ cơ chế chú ý HCANN thích hợp cho việc biểu diễn ngữ cảnh mức từ và mức câu trong chuỗi đầu vào trong bài toán sinh diễn giải văn bản đã được cống bố trong kỷ yếu hội thảo quốc tế MIWAI 2018.

Chương 4

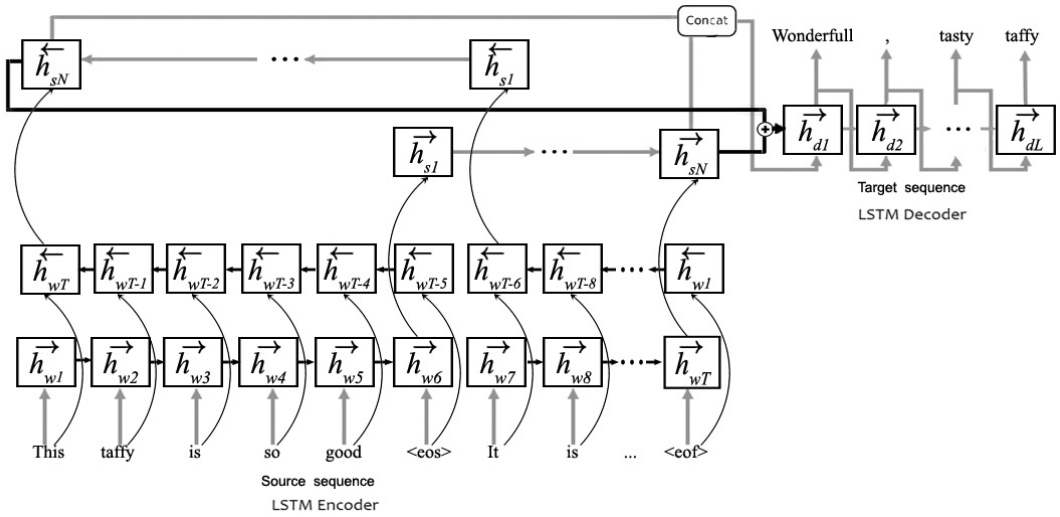
Mô hình học sinh chuỗi từ chuỗi cho bài toán sinh tóm tắt tóm lược

4.1 Mô hình biểu diễn phân cấp cho bài toán tóm tắt tóm lược

4.1.1 Mô hình đề xuất

Chúng tôi đề xuất mô hình gồm hai thành phần chính: bộ mã hoá và bộ giải mã dựa trên mạng LSTM thành phần và được minh hoạ chi tiết trong hình 4.1. Trong đó bộ mã hoá được thiết kế với nhiều lớp LSTM được xếp chồng lên nhau nhắm thực hiện các chắc năng mã hoá cho các đối tượng khác nhau trong văn bản. Cụ thể với mỗi mức biểu diễn của văn bản nguồn(văn bản đầu vào), chúng tôi mô hình hoá như sau:

- h_t^{ew} và h_t^{es} lần lượt là trạng thái ẩn mức từ và mức câu trong bộ mã hoá; h_t^{dw} là trạng thái ẩn mức từ trong bộ giải mã, ở bước thời gian t.
- x_t^{ew} và x_t^{es} lần lượt là véc tơ nhúng mức từ và mức câu trong bộ mã hoá tại vị trí t.
- y_t^{dw} là véc tơ nhúng mức từ ở vị trí t trong bộ giải mã.



Hình 4.1: Mô hình biểu diễn phân cấp.

Bảng 4.1: Kết quả thực nghiệm trên kho dữ liệu GigaWord

Model/Datasets	smaller than 150 words			larger than 150 words		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
ABS+ (Rush et al., 2015)	29.86	12.65	28.34	27.92	11.75	27.15
RAS-Eleman (Chopra et al., 2016)	33.78	**15.97**	31.15	32.28	14.28	30.75
NMT (Luong et al., 2015)	33.10	14.45	30.71	31.35	13.23	29.79
Hierarchical seq2seq	33.55	15.60	31.78	33.40	15.45	31.16
Our Model	**34.08**	15.90	**32.80**	**34.25**	**16.20**	**32.80**

Bảng 4.2: Kết quả thực nghiệm trên kho dữ liệu Amazon Reviews

Model/Datasets	smaller than 150 words			larger than 150 words		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
ABS+ (Rush et al., 2015)	80.95	38.46	79.45	80.25	37.67	78.65
RAS-Eleman (Chopra et al., 2016)	**84.77**	**41.76**	81.36	**84.01**	41.15	80.50
NMT (Luong et al., 2015)	80.15	38.65	80.05	79.90	37.89	79.58
Hierarchical seq2seq	83.95	42.10	81.65	82.86	41.67	80.78
Our Model	**84.71**	**43.28**	**83.21**	**83.61**	**42.49**	**82.09**

Tại pha mã hoá: Mỗi câu sẽ được chèn thêm ký hiệu kết thúc câu "eos" và kết thúc văn bản sẽ được chèn thêm ký hiệu "oef". Lần lượt từng từ của câu sẽ được truyền vào dưới dạng véc tơ nhúng cho lớp mạng LSTM thứ nhất để học dạng biểu diễn câu, sau đó một lớp mạng LSTM khác sẽ được sử dụng để học biểu diễu văn bảng với thành phần cơ sở là các dạng biểu diễn câu.

4.1.2 Thực nghiệm

4.2 Cơ chế chú ý cục bộ cho bài toán tóm tắt tóm lược

4.2.1 Mô hình đề xuất

Chúng tôi đề xuất mô hình kết hợp giữa cơ chế chú ý cục bộ và cơ chế chú ý toàn cục nhằm khai thác được đầu đủ thông tin vai trò của các thành phần trong chuỗi trong quá trình sinh chuỗi đầu ra thông qua thuật toán sau:

Thuật toán 1 Cơ chế chú ý toàn cục trên mạng thặng dư.

Đầu vào: Véc tơ trạng thái ẩn của bộ giải mã h_t^{raa} và tất cả các véc tơ trạng thái ẩn của bộ mã hoá h_s^{raa}.

Đầu ra: Véc tơ chú ý c_t tại mỗi bước thời gian t ở phía bộ giải mã.

- Bước 1: Tính điểm chú ý. Với mỗi véc tơ trạng thái ẩn của bộ mã hoá thì ta cần tính điểm thể hiện sự liên quan với vector trạng thái ẩn h_t^{raa} của bộ giải mã. Cụ thể, ta sẽ áp dụng một phương trình tính điểm "chú ý" với đầu vào là véc tơ trạng thái ẩn của bộ giải mã - h_t^{raa} và một véc tơ trạng thái ẩn của bộ mã hoá - h_s^{raa} và trả về một giá trị vô hướng $score(h_t^{raa}, h_s^{raa})$.

- Bước 2: Tính trọng số chú ý. Áp dụng hàm softmax với đầu vào là điểm chú ý.

$$\alpha_{ts} = \frac{\exp(score(h_t^{rra}, h_s^{rra}))}{\exp() \sum_{s'=1}^{S} score(h_t^{rra}, \overline{h_{s'}^{rra}})} \tag{4.2.1}$$

- Bước 3: Tính toán véc tơ ngữ cảnh c_t là tổng của các trọng số chú ý nhân với véc tơ trạng thái ẩn của bộ giải mã tại bước thời gian tương ứng.

$$c_t = \sum_{s'=1}^{S} \alpha_{ts} \overline{h_{s'}} \tag{4.2.2}$$

Trong đó véc tơ c_t là véc tơ ngữ cảnh tại bước giải mã t chứa đầy đủ thông tin ngữ cảnh hai chiều với các thông tin chú ý cục bộ và toàn cục trong quá

trình sinh đầu ra của mô hình sinh chuỗi từ chuỗi cho bài toán sinh tóm tắt tóm lược.

4.2.2 Thực nghiệm

Bảng 4.3: Kết quả dữ liệu thực nghiệm trên kho dữ liệu Gigaword

	Gigaword					
	Smaller than 150 words			Larger than 150 words		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
ABS+	29.86	12.65	28.34	27.92	11.75	27.15
RAS-Eleman	33.78	15.97	31.15	32.28	14.28	30.75
Sequence-to-sequence RNNs	33.68	15.45	31.71	32.35	15.23	30.79
Pointer-Generator Networks	33.65	**16.60**	31.78	33.65	15.45	31.56
Generative Adversarial Network	34.15	16.25	**31.80**	33.75	15.55	31.90
Our proposed model (LRRA)	**34.01**	15.95	31.04	**34.10**	**15.80**	**31.95**

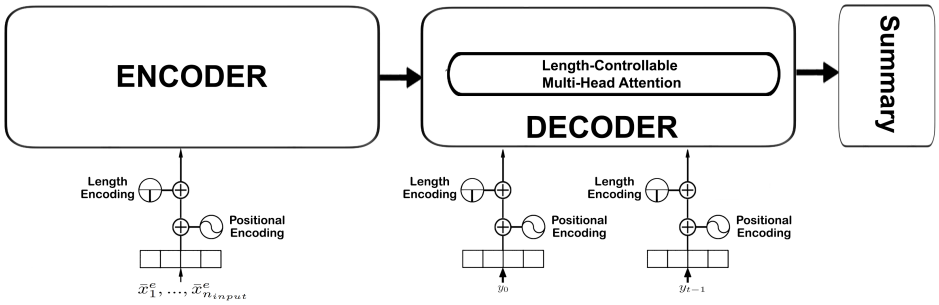
Bảng 4.4: Kết quả dữ liệu thực nghiệm trên kho dữ liệu Amazon Review

	Amazon Review					
	Smaller than 150 words			Larger than 150 words		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
ABS+	80.95	38.46	79.45	80.25	37.67	78.65
RAS-Eleman	**84.77**	41.76	81.36	84.01	41.15	80.50
Sequence-to-sequence RNNs	81.75	42.65	81.05	80.90	38.89	80.25
Pointer-Generator Networks	84.25	42.10	81.85	82.98	42.17	81.18
Generative Adversarial Network	84.55	**43.20**	82.05	83.56	**42.56**	81.78
Our proposed model (LRRA)	84.60	43.15	**82.68**	**84.21**	42.50	**81.88**

4.3 Mô hình sinh tóm tắt văn bản tóm lược có ràng buộc độ dài

4.3.1 Mô hình đề xuất

Kiến trúc mô hình tích hợp thông tin độ dài mong muốn vào mô hình sinh chuỗi từ chuỗi sử dụng kiến trúc transfomer được mô tả chi tiết trong hình 4.2.



Hình 4.2: Mô hình đề xuất đầy đủ.

Length	Model	ROUGE-1	ROUGE-2	ROUGE-L
30	**Mô hình đề xuất đầy đủ**	**36.98**	**17.88**	**32.94**
	Mô hình đề xuất 1	34.28	15.78	32.64
	Mô hình đề xuất 2	35.98	16.28	31.74
50	**Mô hình đề xuất đầy đủ**	42.34	**21.53**	39.16
	Mô hình đề xuất 1	39.64	19.13	38.86
	Mô hình đề xuất 2	41.14	21.33	37.36
70	**Mô hình đề xuất đầy đủ**	41.23	19.89	34.87
	Mô hình đề xuất 1	39.73	19.59	33.67
	Mô hình đề xuất 2	40.83	19.69	34.27

Bảng 4.5: Kết quả thực nghiệm của các mô hình đề xuất trên kho dữ liệu CNN/DM

4.5.2.1 Tích hợp ràng buộc độ dài vào bộ mã hoá

4.5.2.2 Tích hợp ràng buộc độ dài vào bộ giải mã

4.3.2 Thực nghiệm

Để tạo ra các tham chiếu so sánh với các nghiên cứu liên quan, tương tự các nghiên cứu trước đây cho bài toán sinh tóm tắt có ràng buộc độ dài, chúng tôi tiến hành thực nghiệm với độ dài đầu ra mong muốn với các kích thước khác nhau, cụ thể lần lượt là 30, 50 và 70.

Length	Model	ROUGE-1	ROUGE-2	ROUGE-L
30	**Mô hình đề xuất đầy đủ**	**40.75**	20.04	**37.19**
	Mô hình đề xuất 1	38.65	18.54	35.09
	Mô hình đề xuất 2	39.85	19.74	36.89
50	**Mô hình đề xuất đầy đủ**	**39.04**	19.32	**37.78**
	Mô hình đề xuất 1	36.64	16.92	35.08
	Mô hình đề xuất 2	38.44	18.52	36.18
70	**Mô hình đề xuất đầy đủ**	**38.10**	23.77	**33.54**
	Mô hình đề xuất 1	36.50	23.37	31.74
	Mô hình đề xuất 2	37.50	23.17	32.34

Bảng 4.6: Kết quả thực nghiệm của các mô hình đề xuất trên kho dữ liệu NEWROOMS

4.4 Kết luận chương

Trong chương này, luận án đã trình bày ba mô hình đã đề xuất cho bài toán sinh tóm tắt tóm lược văn bản.

Mô hình thứ nhất là mô hình học biểu diễn phân cấp dữ trên kiến trúc học sinh chuỗi từ chuỗi đã được cống bố trong kỷ yếu hội thảo quốc tế KSE 2021.

Mô hình thứ hai đó là mô hình kết hợp cơ chế chú ý cục bộ mà cơ chế chú ý toàn cục đã được cống bố trong kỷ yếu hội thảo quốc tế APIT 2022.

Đặc biệt chúng tôi đã đề xuất và kiểm nghiệm giả thiết mô hình sinh tóm tắt tóm lược có ràng buộc độ dài dựa trên mô hình sinh chuỗi từ chuỗi đã được chấp nhận đăng trong tạp chí IASC 2023(thuộc danh mục SCIE).

Chương 5

Kết luận

5.1 Các đóng góp của luận án

Trong luận án này, chúng tôi tập trung phát triển các mô hình Seq2seq để giải quyết các vấn đề nêu trên thông quá đó Luận án đã được các kết quả và đóng góp bao gồm:

- Cấu trúc biểu diễn phân cấp của văn bản có vai trò trong việc biểu diễn ngữ nghĩa của các thành tố trong văn bản đầu vào, luận án đã tập trung nghiên cứu, đề xuất phương pháp biểu diễn phân cấp văn bản trong mô hình học sinh chuỗi từ chuỗi cho bài toán sinh tóm tắt tóm lược.

- Với lớp các bài toán sinh ngôn ngữ, vai trò của của các thành tố trong văn bản đầu vào có vai trò quyết định đến chất lượng mô hình sinh, tuy nhiên mỗi lớp bài toán sẽ có những lớp đặc trưng riêng nên Luận án tập trung nghiên cứu, đề xuất cơ chế chú ý trong mô hình học sinh chuỗi từ chuỗi cho bài toán sinh diễn giải văn bản.

- Ngoài vai trò của các từ, thì trong văn bản vai trò của các cụm, các câu sẽ có ảnh hưởng nhất định đến việc hiểu ngữ nghĩa của văn bản đầu vào. Luận án cũng đã tập trung nghiên cứu, đề xuất cơ chế chú ý phân cấp có điều kiện trong mô hình học sinh chuỗi từ chuỗi cho bài toán sinh diễn giải.

- Đối với mô hình cơ chế chú ý toàn cục thường cho những kết quả khả thi hơn

các mô hình chú ý cục bộ nhưng đối với các văn bản dài thì mô hình chú ý toàn cục sẽ có gặp nhiều thách thức ở độ phức tạp tính toán nên Luận án cũng đã tập trung ghiên cứu, đề xuất cơ chế chú ý cục bộ kết hợp mạng nơ ron thặng dư để có thể tạo ra cơ chế chú ý toàn cục thích hợp cho bài toán sinh tóm tắt tóm lược văn bản.

- Cuối cùng, ràng buộc trong các mô hình sinh văn bản dạng end-to-end là một bài toán mới mở và có nhiều thách thức, luận án cũng đã nghiên cứu, đề xuất mô hình học sinh chuỗi từ chuỗi cho bài toán sinh tóm tắt tóm lược có ràng buộc độ dài.

5.2 Hướng phát triển

Trong nghiên cứu của chúng tôi, hiện tại chúng tôi tập trung vào việc đánh giá vai trò các thành tố trong văn bản nguồn, và mức độ ảnh hưởng đến quá trình sinh ra văn bản đích trong các mô hình học sinh chuỗi từ chuỗi. Việc xây dựng cơ chế chú ý cho các thành tốt này còn sử dụng một số các tham số mang tính thử sai. Một trong những hướng nghiên cứu tiếp theo là xây dựng mô hình học tham số hướng đến các hệ thống end-to-end hoàn chỉnh.

Đối với bài toán ràng buộc trong quá trình sinh trên mô hình học sinh chuỗi từ chuỗi chúng tôi mới dừng lại xem xét trên các yếu tố định lượng như độ dài ngoài ra các yếu tố định tính như kiểu, hay lứa tuổi,... cũng là những hướng nghiên cứu hứa hẹn trong các bài toán sinh văn bản trong xử lý ngôn ngữ tự nhiên.

Danh mục công trình khoa học của tác giả liên quan đến luận án

[1] Ngoc-Khuong Nguyen, Viet-Ha Nguyen, Dac-Nhuong Le and Anh-Cuong Le. "A Method of Integrating Length Constraints into Encoder-Decoder Transformer for Abstractive Text Summarization",Jounal of Intelligent Automation & Soft Computing(2022) - Accepted.

[2] Ngoc-Khuong Nguyen, Anh-Cuong Le and Viet-Ha Nguyen. "A Local Attention-based Neural Networks for Abstractive Text Summarization",5th Asia Pacific Information Technology Conference (APIT 2022) - Submitted.

[3] Khuong Nguyen-Ngoc, Anh-Cuong Le and Viet-Ha Nguyen. "A Hierarchical Encoder-Decoder Long Short-Term Memory Model for Abstractive Summarization",13th International Conference on Knowledge and Systems Engineering (KSE 2021),pp 281-286.

[4] Khuong Nguyen-Ngoc, Anh-Cuong Le and Viet-Ha Nguyen. "A Hierarchical Conditional Attention-based Neural Networks for Paraphrase Generation",the 12th Multi-disciplinary International Conference on Artificial Intelligence (MIWAI), 2018, pp 161 - 174, DOI:10.1007978-3-030-03014-8_14.

[5] Khuong Nguyen-Ngoc, Anh-Cuong Le and Viet-Ha Nguyen. "An Attention-based Long-Short-Term-Memory Model for Paraphrase Generation ",the 6th International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making(IUKM), 2018, pp.166-178, DOI:10.1007978-3-319-75429-1_14.

Tài liệu tham khảo

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[2] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1724–1734. ACL, 2014.

[3] Sumit Chopra, Michael Auli, and Alexander M. Rush. Abstractive sentence summarization with attentive recurrent neural networks. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 93–98. The Association for Computational Linguistics, 2016.

[4] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. In *Advances in neural information processing systems*, pages 577–585, 2015.

[5] Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. Multi-task sequence to sequence learning. *arXiv preprint arXiv:1511.06114*, 2015.

[6] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1412–1421. The Association for Computational Linguistics, 2015.

[7] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[8] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[9] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016.