TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN CUỐI KÌ**

**NHẬP MÔN HỌC MÁY**

**Tóm tắt văn bản tiếng việt tự động**

*Người hướng dẫn*: **TS BÙI THÀNH HÙNG**

*Người thực hiện*: **NGUYỄN MINH NHỰT – 518H0545**

**ĐẶNG THỌ DUY – 518H0168**

Lớp **: 18H50203**

Khoá  **: 22**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN CUỐI KÌ**

**NHẬP MÔN HỌC MÁY**

**Tóm tắt văn bản tiếng việt tự động**

*Người hướng dẫn*: **TS BÙI THÀNH HÙNG**

*Người thực hiện*: **NGUYỄN MINH NHỰT – 518H0545**

**ĐẶNG THỌ DUY – 518H0168**

Lớp **: 18H50203**

Khoá  **: 22**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021**

# LỜI CẢM ƠN

Lời đầu tiên, chúng chúng em xin được gửi lời cảm ơn chân thành đến với Khoa CNTT. Chúng em xin chân thành cảm ơn sự hỗ trợ, giúp đỡ trong quá trình thực hiện bài báo cáo môn Nhập môn học máy này.

Chúng em xin chân thành cảm ơn thầy Bùi Thanh Hùng - thầy đã trực tiếp hướng dẫn và chỉ bảo chúng chúng em hoàn thành bài báo cáo này.

Chúng tôi xin chân thành cảm ơn những người bạn đang học tập và làm việc tại trường Đại học Tôn Đức Thắng và gia đình đã đồng hành, động viên, tạo điều kiện và giúp đỡ chúng em trong suốt quá trình này.

Do năng lực thực tế của chúng em còn non yếu nên chúng em đảm bảo còn nhiều thiếu sót, mong quý thầy cô bỏ qua. Đồng thời chúng em mong nhận được nhiều ý kiến đóng góp từ nhiều nguồn khác nữa, điều này sẽ giúp chúng em tích lũy thêm kinh nghiệm để hoàn thành bài báo cáo tốt nghiệp sắp tới đạt kết quả tốt hơn.

**ĐỒ ÁN ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Tôi xin cam đoan đây là sản phẩm đồ án của riêng tôi / chúng tôi và được sự hướng dẫn của TS Bùi Thanh Hùng. 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 đồ án 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 tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình.** Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày 21 tháng 04 năm 2021*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

**

*Nguyễn Minh Nhựt*

**

*Đặng Thọ Duy*

PHẦN ĐÁNH GIÁ CỦA GIẢNG VIÊN

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

# TÓM TẮT

Tóm tắt văn bản tiếng Việt tự động là một chú ý không thể bỏ qua khi nhắc tới các nghiên cứu xử lý ngôn ngữ tự nhiên. Đây là một vấn đề được đông đảo các nhà khoa học tham gia nghiên cứu và hiện giờ đã có khá nhiều những mô hình giải quyết cho bài toán này.

Trong bài báo cáo này, chúng em sẽ giải quyết bài toán tóm tắt văn bản tiếng Việt tự động theo hướng tóm lược – abstractive based summarization, ứng dụng mô hình học sâu Seq2Seq.

Với đánh việc sử đụng độ đo ROUGE – một thang đo phổ biến được đưa ra để đánh giá các mô hình, chúng em nhận thấy mô hình của nhóm còn non trẻ và cần nhiều sự cố gắng hơn nữa.

# MỤC LỤC

[LỜI CẢM ƠN i](#_Toc70371277)

[TÓM TẮT iv](#_Toc70371278)

[MỤC LỤC 1](#_Toc70371279)

[DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT 3](#_Toc70371280)

[DANH MỤC CÁC HÌNH VẼ 4](#_Toc70371281)

[DANH MỤC CÁC BẢNG 5](#_Toc70371282)

[PHẦN I 7](#_Toc70371283)

[PHÂN TÍCH YÊU CẦU CỦA BÀI TOÁN 7](#_Toc70371284)

[1.1 Giới thiệu về bài toán 7](#_Toc70371285)

[1.2 Phân tích yêu cầu của bài toán 8](#_Toc70371286)

[1.2.1 Khái quát về bài toán 8](#_Toc70371287)

[1.2.2 Yêu cầu của bài toán 8](#_Toc70371288)

[PHẦN II 10](#_Toc70371289)

[PHƯƠNG PHÁP GIẢI QUYẾT 10](#_Toc70371290)

[2.1 Các phương pháp giải quyết bài toán 10](#_Toc70371291)

[2.1.1 Hướng trích chọn – Extraction 10](#_Toc70371292)

[2.1.2 Hướng tóm lược – Abstraction 15](#_Toc70371293)

[2.2 Phương pháp đề xuất giải quyết bài toán 17](#_Toc70371294)

[2.3 Phương pháp giải quyết bài toán 18](#_Toc70371295)

[2.3.1 Mô hình tổng quát 18](#_Toc70371296)

[2.3.2 Đặc trưng của mô hình đề xuất 20](#_Toc70371297)

[2.3.2.1. Cào dữ liệu 20](#_Toc70371298)

[2.3.2.2 Trích xuất đặc trưng 20](#_Toc70371299)

[PHẦN III 25](#_Toc70371300)

[THỰC NGHIỆM 25](#_Toc70371301)

[3.1 Dữ liệu 25](#_Toc70371302)

[3.1.1 Nguồn gốc dữ liệu 25](#_Toc70371303)

[3.1.2 Mô tả dữ liệu 25](#_Toc70371304)

[3.2 Xử lý dữ liệu 27](#_Toc70371305)

[3.3 Công nghệ sử dụng 28](#_Toc70371306)

[3.4 Cách đánh giá 29](#_Toc70371307)

[3.5 Kết quả đạt được 29](#_Toc70371308)

[PHẦN IV 34](#_Toc70371309)

[KẾT LUẬN 34](#_Toc70371310)

[4.1 Kết quả đạt được 34](#_Toc70371311)

[4.2 Hướng phát triển 34](#_Toc70371312)

[TÀI LIỆU THAM KHẢO 36](#_Toc70371313)

[TỰ ĐÁNH GIÁ 37](#_Toc70371314)

# DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT

**CÁC CHỮ VIẾT TẮT**

LSTM : Long Short Term Memory

Seq2Seq: Sequence to Sequence

ROUGE: Recall-Oriented Understudy for Gisting Evaluation

# DANH MỤC CÁC HÌNH VẼ

[Hình 1: Yêu cầu của bài toán tóm tắt văn bản 8](#_Toc70371081)

[Hình 2: Highlighter = Hướng trích chọn 11](#_Toc70371082)

[Hình 3: Ví dụ về sự kiện và các thành phần trong câu 13](#_Toc70371084)

[Hình 4: Mô hình trích chọn sự kiện 13](#_Toc70371085)

[Hình 5: Mô hình tóm tắt sự kiện 14](#_Toc70371086)

[Hình 6: Bộ dữ liệu cho mô hình 14](#_Toc70371087)

[Hình 7: Đánh giá mô hình phát hiện sự kiện 15](#_Toc70371088)

[Hình 8: Đánh giá mô hình trích chọn sự kiện 15](#_Toc70371089)

[Hình 9: Viết thường = Hướng tóm lược 15](#_Toc70371090)

[Hình 10: Mô hình giải quyết bài toán 17](#_Toc70371092)

[Hình 11: Điểm ROUGE sau khi đánh giá mô hình 17](#_Toc70371093)

[Hình 12: Mô hình tổng quát 20](#_Toc70371094)

[Hình 13: Mô hình Seq2Seq cộng cơ chế Attention 22](#_Toc70371095)

[Hình 14: Ví dụ ứng dụng 23](#_Toc70371096)

[Hình 15: 1 khối nhớ LSTM tiêu biểu 24](#_Toc70371097)

[Hình 16: Cây thư mục – Folder tree tập dữ liệu 26](#_Toc70371098)

[Hình 17: Trích 1 đoạn nhỏ trong văn bản gốc – plaintext 27](#_Toc70371099)

# DANH MỤC CÁC BẢNG

[Bảng 1: Bảng đánh giá điểm số ROUGE mô hình 34](#_Toc70371267)

# PHẦN I

# PHÂN TÍCH YÊU CẦU CỦA BÀI TOÁN

## 1.1 Giới thiệu về bài toán

Ngày nay công nghệ thông tin và máy tính phát triển mạnh mẽ đi theo đó là khối lượng tri thức, tài liệu văn bản khổng lồ được tạo ra với nhiều mục đích sử dụng khác nhau làm cho việc đọc hiểu và trích lọc các thông tin quan trọng trong khối tri thức đồ sộ này tốn rất nhiều thời gian và chi phí (đặc biệt là chi phí cho hạ tầng và truyền dẫn thông tin đáp ứng yêu cầu cho một số lượng ngày càng nhiều các thiết bị cầm tay). Để tăng hiệu quả trong việc tiếp nhận thông tin ấy, nhiều nghiên cứu về khai phá dữ liệu và xử lý ngôn ngữ tự nhiên đã được thực hiện. Một trong những nghiên cứu quan trọng đóng vai trò then chốt đó tóm tắt văn bản tự động.

Tóm tắt văn bản là một vấn đề nhận được nhiều sự quan tâm của cộng đồng nghiên cứu với rất nhiều các bài báo khoa học xuất hiện tại các hội nghị lớn như: ACL, DUC, TAC, COLING, EMNLP...

Hiện trên thị trường đã có nhiều mô hình thực hiện và giải quyết bài toán tóm tắt văn bản nhưng đa số đó là văn bản tiếng Anh vì thế tóm tắt văn bản nói riêng và xử lý ngôn ngữ tự nhiên nói chung áp dụng cho tiếng Việt gặp nhiều thách thức hơn. Sở dĩ là vì tiếng Việt với đặc trưng như là tiếng đơn âm và có thanh điệu nên việc tách từ, tách các

thành phần ngữ nghĩa trong câu tiếng Việt đòi hỏi xử lý phức tạp hơn so với xử lý câu trong tiếng Anh, song số lượng kho dữ liệu (dataset) bằng tiếng Việt được chuẩn hóa còn hạn hẹp và tính chính thức cũng chưa cao như tiếng Anh.

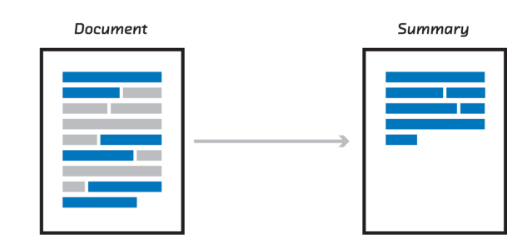
## 1.2 Phân tích yêu cầu của bài toán

### 1.2.1 Khái quát về bài toán

Đầu tiên, tóm tắt được định nghĩa như là một văn bản dược tạo từ một hoặc nhiều văn bản, truyền đạt các thông tin quan trọng từ các văn bản gốc, văn bản tóm tắt không dài hơn 50% độ dài văn bản gốc và thông thường bản tóm tắt có độ dài khá ngắn, ngắn hơn nhiều so với 50% độ dài văn bản gốc.

### 1.2.2 Yêu cầu của bài toán

Tất cả những gì chúng ta cần cho bài toán này là một tóm tắt chính xác và hợp ngữ pháp trong khi vẫn giữ được các thông tin chính và ý nghĩa của văn bản gốc.



#### Hình 1: Yêu cầu của bài toán tóm tắt văn bản

Mục tiêu chính của bài toán là tạo ra bản tóm tắt giống như cách con người tóm tắt, vì thế mà đây là bài toán được đánh giá là đầy thách thức. Khi chính con ngườ thực hiện tóm tắt một văn bản, chúng ta càn phải đọc toàn bộ nội dung rồi dựa trên sự hiểu biết và tiếp thu của bản thân để viết lại một đoạn tóm tắt nhằm làm nổi bật các ý chính của văn bản gốc.

Và chúng ta cũng đã biết, tất cả những gì máy tính hiểu là dãy bit 0 và 1 vì thế nó khó có thể có được tri thức và khả năng cảm thụ ngôn ngữ như của con người, nên việc thực hiện tóm tắt văn bản tự động là một công việc phức tạp.

# PHẦN II

# PHƯƠNG PHÁP GIẢI QUYẾT

## 2.1 Các phương pháp giải quyết bài toán

Theo khảo sát của chúng em, ở đây có khá nhiều hướng để tiếp cận cho bài toán tóm tắt văn bản tiếng việt. Theo tác giả Nguyễn Mạnh Cường [3], thầy đã đề ra các giải pháp như sau:

* Tóm tắt theo hướng trích chọn – extractive based summarization
* Tóm tắt theo hướng tóm lược – abstractive based summarization
* Tóm tắt bằng đồ thị hóa
* Tóm tắt bằng các mô hình học máy

Nhưng chung quy, chúng ta có 2 hướng chính để ta giải quyết bài toán này, đầu tiên là giải quyết theo hướng trích chọn – extraction và tóm lược – abstraction.

### 2.1.1 Hướng trích chọn – Extraction

Đầu tiên là giải quyết theo hướng trích chọn – extraction, chúng ta sẽ thực hiện đánh giá các phần quan trọng của văn bản, do đó, phương pháp này chỉ phụ thuộc vào việc trích chọn các câu từ văn bản gốc dựa trên việc xếp hạng mức độ liên quan của các cụm từ để chỉ chọn những cụm từ liên quan nhất tới nội dung của tài liệu gốc hay được gọi là khối lượng từ trọng tâm – Weight text.

Nói một cách dễ hiểu hơn là hướng tóm lược đóng vai trò như 1 cây viết dạ quang – highlighter, thứ mà sẽ làm nổi bật lên những ý chính trong văn bản gốc



#### Hình 2: Highlighter = Hướng trích chọn

Ví dụ như sau:

**Văn bản gốc**: Peter và Elizabeth bắt taxi để đi dự buổi tiệc tối nay ở thành phố Duo Flora. Trong khi đang dự tiệc, Elizabeth đột nhiên bất tỉnh và phải nhập viện khẩn cấp.

**Văn bản tóm tắt**: Peter và Elizabeth dự buổi tiệc ở thành phố Duo Flora, Elizabeth nhập viện khẩn cấp

#### **2.1.1.1 Phương pháp minh họa – Trích chọn hướng sự kiện:**

Như trích dẫn bên dưới, mục [8], tạm dịch là “Tóm tắt hướng trích chọn dựa trên trích xuất sự kiện” của nhóm tác giả Maofu Liu, Wenje Li, Mingli Wu, Quin Lu.

Khái quát thì tóm tắt sự kiện là trích xuất và tố chức các đoạn tóm tắt theo các điều kiện của sự kiện có trong câu muốn tóm tắt. Cụ thể là chúng ta sẽ đi theo quy luật phân chia sự kiện như sau: “[Who] did [What] to [Whom] [When] and [Where]” and “did [What]”, tạm dịch là “[Ai] làm [Cái gì] với [Ai] [Khi nào] và [Ở đâu]” và “đã làm [Cái gì]”

Hiện nay cũng có khá nhiều những công cụ tóm tắt sử dụng hướng đi này ứng dụng nhiều miền dữ liệu khác nhau như:

* Tóm tắt sự kiện trên mạng xác hội Twitter như là NSS09, CP11, SKW13, CA13
* Tóm tắt các sự kiện tin tức như là WLL07, WWL10, LLW07, Gi13

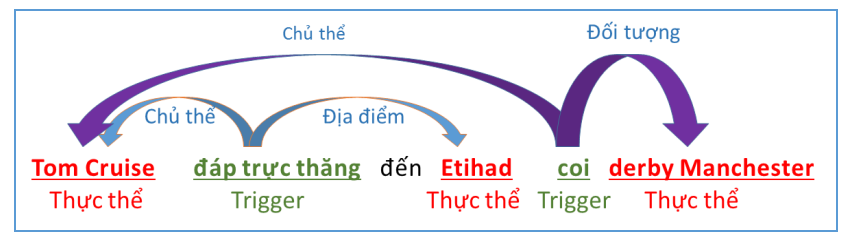
Qua bài báo cáo của tác giả Trần Mai Vũ [9] thì bà định nghĩa sự kiện trong văn bản tiếng Việt như sau:

* Sự kiện tin tức trong tiếng Việt mô tả một hành vi cụ thể của một chủ thể tại một thời gian và địa điểm xác định.
* Hành vi này được biểu diễn bằng một mối quan hệ ngữ nghĩa giữa một hay nhiều thực thể thực hiện cùng một vai trò với nhau

Cải tiến hơn về việc phân nghĩa cho câu của nhóm tác giả Maofu Liu, bà phân 1 câu tiếng Việt ra nhưng thành phần như sau:

* **Thực thể**: là danh từ hoặc cụm danh từ biểu diễn cho chủ thể, có thể đây là tên người, tên tổ chức
* **Trigger sự kiện**: từ hoặc cụm từ trong câu, chỉ ra sự xuất hiện của sự kiện
* và mang kiểu của sự kiện.
* **Tham số (của sự kiện):** các thực thể hoặc các sự kiện khác góp phần mô tả
* sự kiện, cũng là một phần của sự biểu diễn sự kiện, và thường được phân
* loại theo vai trò ngữ nghĩa.
* **Vai trò tham số:** mỗi một tham số biểu diễn một sự tác động hoặc tham gia
* vào sự kiện của một thực thể.

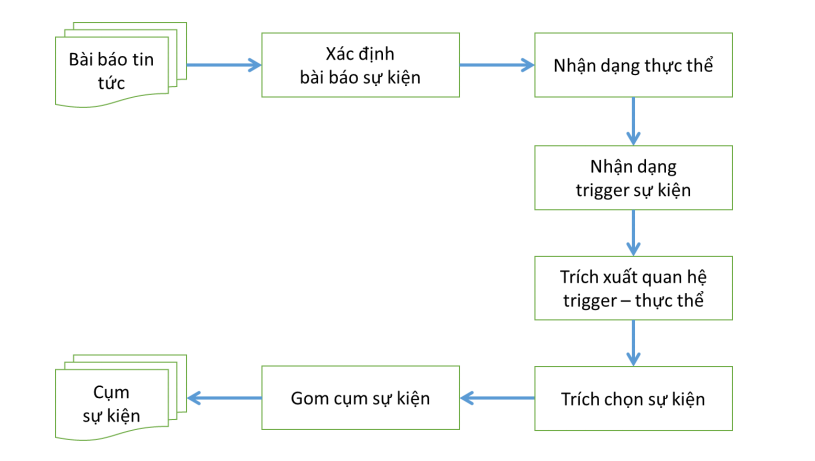
Bà cũng đưa ra 1 ví dụ như sau:



##### Hình 3: Ví dụ về sự kiện và các thành phần trong câu

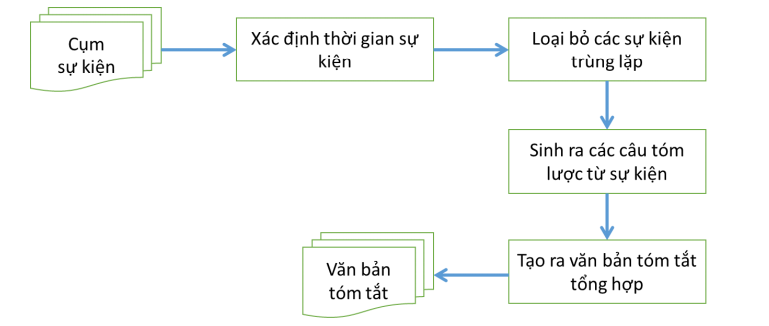
Trong ví dụ trên thì ta thấy có hai sự kiện xuất hiện là “Tom Cruise đáp trực thăng đến Etihad” và “Tom Cruise coi derby Manchester” . Các cụm từ bôi đỏ là các thực thể và các cụm từ bôi xanh là các Trigger. Mỗi một sự kiện có hai quan hệ liên quan đến Trigger mỗi một quan hệ thể hiện mối quan hệ giữa trigger và thực thể, ở đây thực thể là tham số của sự kiện và nhãn của quan hệ sẽ là vai trò của tham số

Sau đây, là các bước tiến hành của mô hình trích xuất sự kiện:



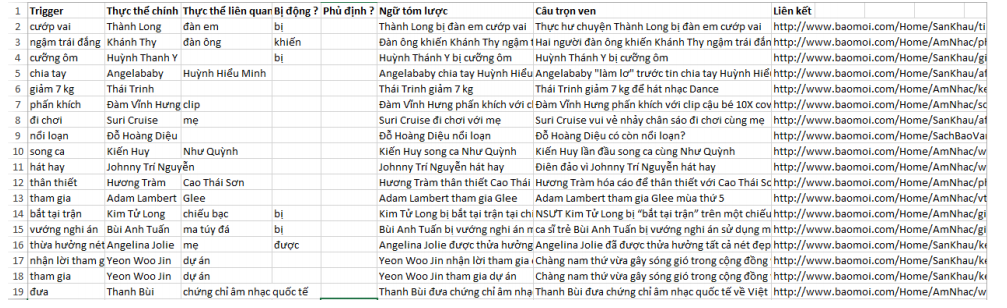
##### Hình 4: Mô hình trích chọn sự kiện

Sau đó, chúng ta tiến hành tóm tắt các sự kiện bằng mô hình như sau:



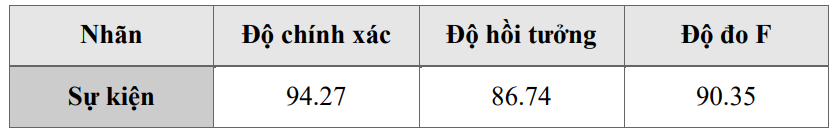
##### Hình 5: Mô hình tóm tắt sự kiện

Còn về dữ liệu cho mô hình thì bà đã lấy từ các mục Văn hóa và Giải trí từ website Baomoi.com và các bước xây dựng bộ dữ liệu chi tiết vui long tham khảo mục 4.1 của tài liệu số [9]. Cuối cùng, chúng ta sẽ nhận được tập dữ liệu như sau:



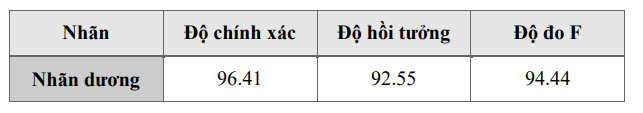
##### Hình 6: Bộ dữ liệu cho mô hình

Tiếp theo là về đánh giá mô hình, tác giả sử dụng kiểm thử 10-folds cross validation (với tỉ lệ 2 lớp 1:1)và độ do micro F trên nhãn Sự kiện cho mô hình phát hiện sự kiện, cho ra kết quả như sau:



##### Hình 7: Đánh giá mô hình phát hiện sự kiện

Áp dụng tương tự trên nhãn dương cho mô hình trích chọn sự kiện, cho ra kết quả như sau:



##### Hình 8: Đánh giá mô hình trích chọn sự kiện

### 2.1.2 Hướng tóm lược – Abstraction

Phương pháp này nhằm tạo ra các văn bản tóm lược có thể không bao gồm các từ hay cụm từ trong văn bản gốc. Khi thực hiện phương pháp này, chúng ta phải làm máy cố gắng hiểu và đánh giá văn bản bằng cách sử dụng những các kỹ thuật xử lý ngôn ngữ tự nhiên để tạo ra 1 văn bản tóm tắt càng ngắn càng tốt nhưng vẫn có thể truyền đạt được những thông tin quan trọng nhất từ văn bản gốc.

Nói một cách dễ hiểu hơn là hướng tóm lược đóng vai trò như 1 cây viết bình thường, thứ mà sẽ viết ra những thứ dựa trên văn bản gốc nhưng chưa có sẵn trong đó.



#### Hình 9: Viết thường = Hướng tóm lược

Ví dụ như sau:

**Văn bản gốc**: **Peter** và **Elizabeth** bắt taxi để đi **dự** **buổi** **tiệc** tối nay ở thành phố Duo Flora. Trong khi đang dự tiệc, **Elizabeth** đột nhiên bất tỉnh và phải **nhập** **viện** khẩn cấp.

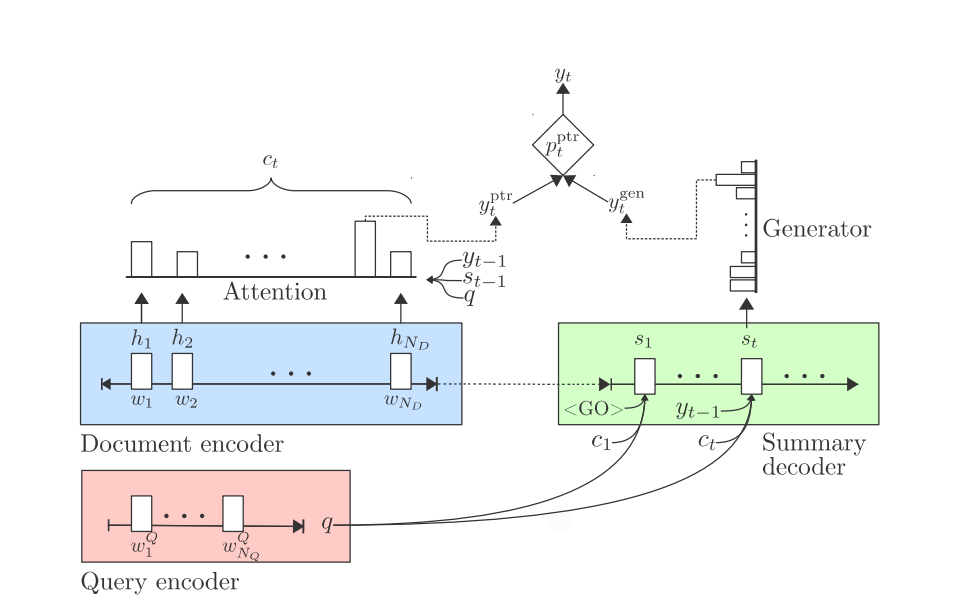
**Văn bản tóm tắt**: Elizabeth phải nhập viện sau khi nhập tiệc với Peter

#### **2.1.2.1 Phương pháp minh họa – Trích chọn hướng tóm lược trên cơ sở truy vấn:**

Như trích dẫn bên dưới, mục [10], tạm dịch là “Tóm tắt hướng tóm lược dựa trên truy vấn” của nhóm tác giả Johan Hasselqvist, Niklas Helmertz, Mikael Kagebac.

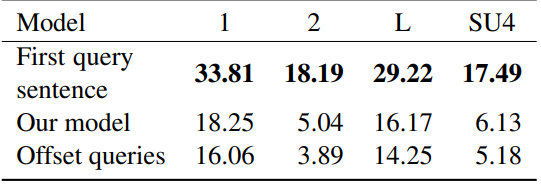
Dựa trên các đặc tả của nhóm tác giả Johan và Trần Mai Vũ [11], tóm tắt Query-based với mục đích chính là tìm ra một đoạn tóm tắt cho toàn bộ văn bản mà nội dung của đoạn văn bản sẽ bao quát toàn bộ nội dung của văn bản đó. Tóm tắt trên cơ sở truy vấn thì nội dung của văn bản tóm tắt sẽ dựa trên truy vấn của người dùng hay chương trình đưa vào, loại tóm tắt này thường được sử dụng trong quá trình tóm tắt các kết quả trả về từ máy tìm kiếm

Sau đây là mô hình giải quyết bài toán của nhóm tác giả:



##### Hình 10: Mô hình giải quyết bài toán

Phương pháp đánh giá mà nhóm tác giả chọn là ROUGE - Recall Oriented Understudy of Gisting Evaluation và đây cũng sẽ là phương pháp đánh giá mô hình mà chúng em sẽ chọn để đánh giá cho mô hình học máy của mình. Sau đây là kết quả mà nhóm tác giả đưa ra:



##### Hình 11: Điểm ROUGE sau khi đánh giá mô hình

Ngoài ra, tác giả có đề cập tới việc làm giúp tăng điểm đánh ROUGE lên cao hơn nếu như có ít từ được chọn hơn khi truy vấn nhưng điều này có mặt trái là làm cho câu không được hoàn thiện.

## 2.2 Phương pháp đề xuất giải quyết bài toán

Nhìn một cách tổng quan thì con người luôn linh hoạt hơn trong việc tóm tắt một cái gì đó và di nhiên không thể đi khớp được với phương pháp tóm tắt bằng cách trích chọn như đã trình bày trên mà là hướng tóm lược nhưng đa số các nghiên cứu được thực hiện trong thời gian gần đây đều ứng dụng hướng trích chọn là chủ yếu vì chính hướng trích chọn này cho ra kết quả khả quan và tối ưu hơn mặc dù ta phải hi sinh tính linh hoạt trong tư duy của con người hiện hữu trong mô hình này.

Dù vậy, mô hình tóm tắt bằng hướng tóm lược – abstractive sẽ được áp dụng cho bài làm của chúng em lần này. Với lượng kiến thức còn non trẻ cộng thêm tính căng thẳng của phương pháp tóm tắt bằng tóm lược khi chúng ta phải đương đầu với các vấn đề như cách thể hiện ngữ nghĩa, suy luận và sinh ngôn ngữ tự nhiên và để thực hiện được tốt, chúng ta cần nhiều thời gian hơn để nghiên cứu cũng như triển khai và đánh giá thì chất lượng kết quả trả ra có thể sẽ không được cao như các mô hinh khác đã hiện hữu trên thị trường. Nhưng tóm tắt bằng hướng tóm lược luôn được mong đợi là cách tóm tắt tối ưu và giống con người nhất nên việc thử sức với phương pháp này là một trải nghiệm thú vị.

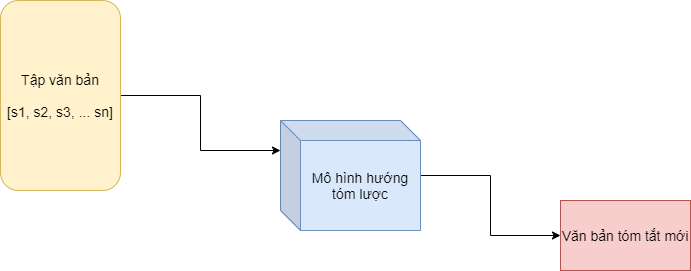
## 2.3 Phương pháp giải quyết bài toán

### 2.3.1 Mô hình tổng quát

Đầu tiên chúng ta sẽ cần dữ liệu đầu vào là rất nhiều văn bản tiếng Việt. Trong bài làm này, chúng em sẽ lấy dataset trực tiếp từ nguồn được đề cập chi tiết ở mục 3.1 bên dưới.

Sau đó sẽ là bước tiền xử lý dữ liệu đầu vào. Như bao mô hình học máy khác, tiền xử lý cho dữ liệu lần này cũng sẽ được thực hiện như vậy. Chi tiết xem ở mục 3.2 Xử lý dữ liệu bên dưới

Ý tưởng non trẻ nhất cho bài toán này là chúng ta sẽ có input là 1 tập câu được cắt từ văn bản tiếng Việt bình thường sau đó nó sẽ được đi qua 1 mô hình tóm tắt hướng tóm lược sau đó sẽ cho ra output là 1 đoạn văn bản tóm tắt mới hướng tóm lược cho tập câu input vừa rồi. Sau đây là mô hình tổng quan cho mô hình:

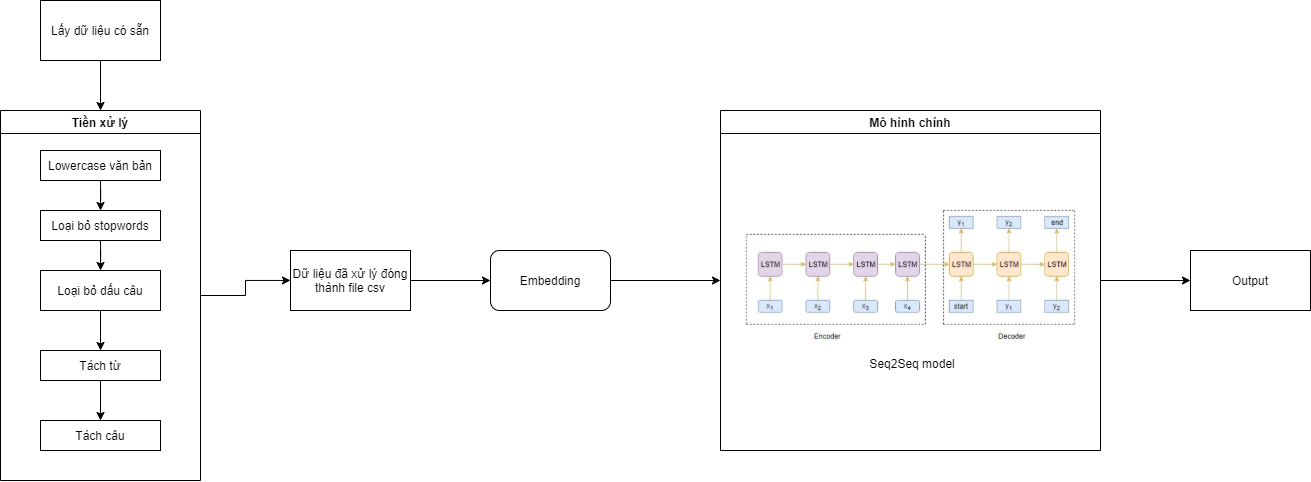


*Hình 11: Mô hình tổng quan cho bài toán*

Còn cho mô hình hướng tóm lược sẽ được ứng dụng mô hình học sâu như:

Seq2Seq ,RNN, LSTM. Chi tiết ứng dụng như thế nào sẽ được nêu rõ mục 2.3.2 Đặc trưng của mô hình đề xuất bên dưới

Tổng kết lại thì chúng ta có được hệ thống các bước đi như sau:



##### Hình 12: Mô hình tổng quát

### 2.3.2 Đặc trưng của mô hình đề xuất

### 2.3.2.1. Cào dữ liệu

Dữ liệu được chúng em lấy trực tiếp từ bộ Open Corpus của tác giả được đặc tả củ thể ở mục 3.1 Dữ liệu

### 2.3.2.2 Trích xuất đặc trưng

Có rất nhiều cách để trích xuất đặc trưng cho mô hình học sâu như là Embedding Layer – Keras, Word2Vec, GloVe, … nhưng lần này chúng em chọn trích xuất đặc trưng sử dụng Word2Vec

Mục tiêu của word embedding là mã hoá từ ngữ tiếng Việt thành những vector đặc trưng mang lại một ý nghĩa ở mức độ nào đó cho từ và nó là phương pháp biểu diễn dữ liệu cho mô hình tóm tắt văn bản tự động

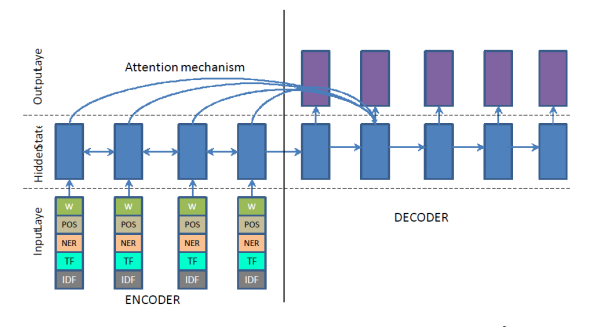
Đối với tiếng Việt, trước tiên, trong tiền xử lý các từ được xử lý chuyển về từ đơn, từ ghép, và in thường rồi loại bỏ các dấu câu ngoại trừ dấu chấm để phân biệt các câu với nhau. Sau đó, mỗi từ trong câu sẽ được sắp xếp theo thứ tự từ điển và chuyển về dạng one-hot vector nhưng đối với khối lượng lớn từ cùng với dấu từ phức tạp không gian vector sẽ bị phá vỡ dễ dàng

2.3.2.4 Phương pháp huấn luyện

Seq2Seq

Đầu tiên, Seq2Seq - Sequence-to-Sequence là một kỹ thuật điển hình áp dụng cho các mô hình tóm tắt văn bản đi theo hướng này. Mặc dù Seq2Seq đã được áp dụng thành công cho nhiều bài toán trong xử lý ngôn ngữ tự nhiên, chẳng hạn như dịch máy, nhưng với bài toán tóm tắt văn bản vẫn còn nhiều hạn chế. Thực tế là mô hình này có thể đạt được điểm ROUGE cao trên các bản tóm tắt với đầu vào nhỏ, nhưng thường không có khả năng tóm tắt khi đầu vào lớn.

Cũng như các mô hình ANN thuần túy, chúng ta có thể xây dựng 1 mô hình Seq2Seq nhiều nhiều – Many to many.



##### Hình 13: Mô hình Seq2Seq cộng cơ chế Attention

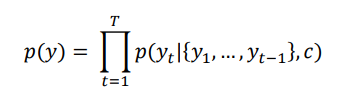
Có 2 components chính trong mô hình này là encoder – lớp mã hóa và decoder – lớp giải mã

Theo [12], đầu tiên 1 lớp encoder sẽ đọc 1 chuỗi vector *x = (x1,…,xTx)* thành 1 vector a có dạng sau

*c = q({h1,…,hTx})*

Trong đó ht là trạng thái ẩn tại thời điểm t, ht ∈ ℝ n và c là vector được sinh ra từ một chuỗi các trạng thái ẩn. f và q là các hàm phi tuyến.

Pha decoder, được huấn luyện để dự đoán từ tiếp theo yt’ cho ngữ cảnh c và tất cả các từ dự đoán đằng trước {y1,…,yt’-1}. Hiểu theo cách khác decoder định nghĩa một xác xuất trên chuyển dịch y bằng việc phân tích xác suất liên kết thành thứ tự các điều kiện:



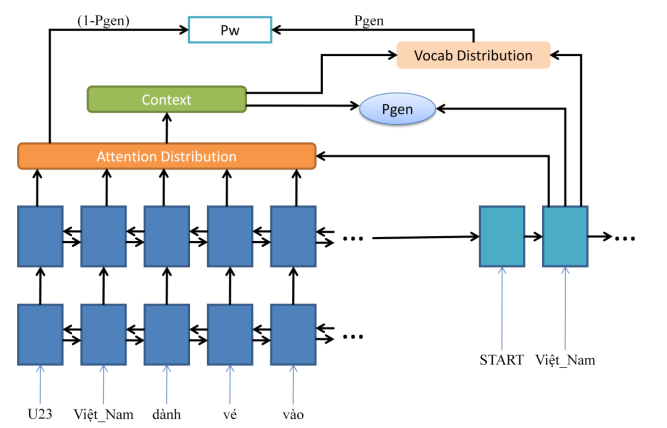
Trong đó y = (y1,…yTy). Với một mạng RNN, mỗi xác suất có điều kiện được mô hình bởi:



Trong đó g là hàm phi tuyến, yt là đầu ra và st là trạng thái ẩn của mạng RNN

Nói 1 cách đơn giản hơn, mô hình chuỗi sang chuỗi có thể được xây dựng bằng kết hợp của hai mạng RNN, một là một mạng chuỗi sang véc tơ (sequence-to-vector) thƣờng được gọi là bộ mã

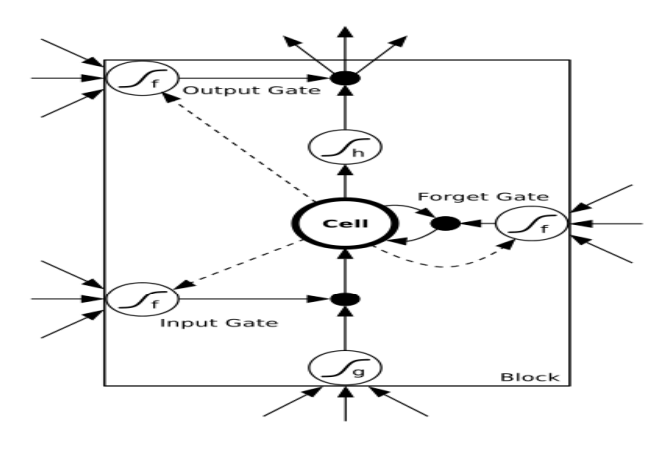
hóa (encoder), theo sau là một mạng véc tơ sang chuỗi (vector-to-sequence) thường được gọi là bộ giải mã (decoder).



##### Hình 14: Ví dụ ứng dụng

LSTM

Như đã được biết từ trước, Kiến trúc mạng LSTM bao gồm một tập các mạng con được kết nối hồi quy, còn gọi là các khối nhớ. Các khối có thể được liên tưởng như là phiên bản khác của các chip nhớ trong máy tính số. Mỗi khối nhớ chứa một hoặc nhiều ô nhớ tự liên kết và ba đơn vị: đầu vào, đầu ra và cổng quên cung cấp khả năng liên tục viết, đọc và hoạt động khởi động cho các ô nhớ.



##### Hình 15: 1 khối nhớ LSTM tiêu biểu

# PHẦN III

# THỰC NGHIỆM

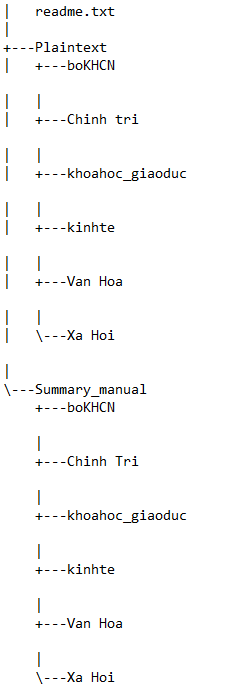
## 3.1 Dữ liệu

### 3.1.1 Nguồn gốc dữ liệu

Dữ liệu – dataset được chúng em lấy trực tiếp từ tác giả [4] Lê Thanh Hương và đã có trích dẫn ở mục Tài liệu tham khảo bên dưới.

### 3.1.2 Mô tả dữ liệu

Dữ liệu mới khi tải về sẽ được đóng gọi dạng file .zip và có cấu trúc thư mục như sau:

**

#### Hình 16: Cây thư mục – Folder tree tập dữ liệu

* Folder “Plaintext” là nơi chứa các bài báo nguyên văn được phân chia theo danh mục các chủ đề
* Folder “Summary\_Manual” là nơi chứa các bản tóm tắt thủ công cho từng bài báo tương ứng trong folder “Plaintext” và cũng được phân chia theo danh mục các chủ đề
* File “readme” đặc tả chi tiết dữ liệu từ phía tác giả, nội dung đầy đủ như sau:

*Tập ngữ liệu tóm tắt đơn văn bản:*

*Tập ngữ liệu liệu chứa 200 văn bản và các bản tóm tắt tương ứng, được tổ chức như sau. Tập ngữ liệu gồm có hai thư mục lớn:*

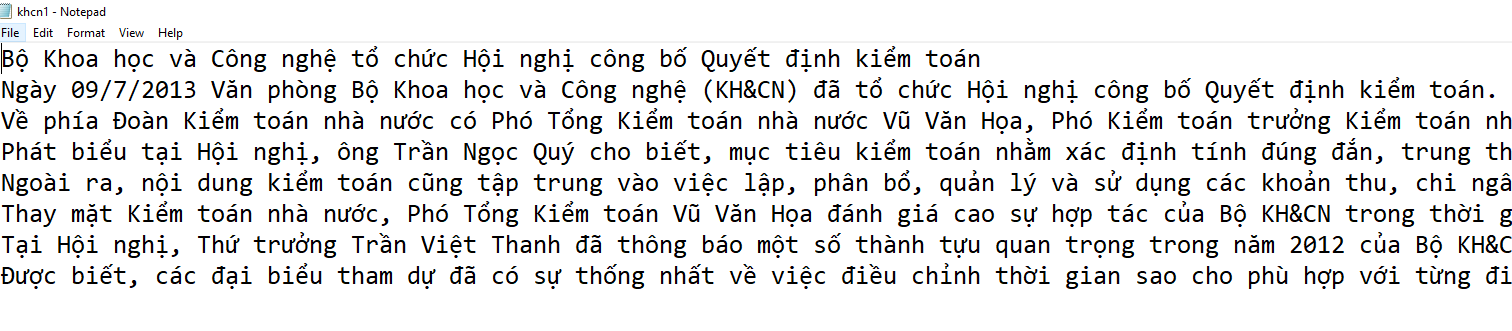
*+ Thư mục Plaintext chứa các văn bản gốc. Các văn bản này là các bản tin được lấy trên các báo điện tử như Dantri, VnExpress, và trên trang web của Bộ Khoa học Công nghệ. Các bản tin này được loại bỏ đoạn tóm tắt in đậm và các liên kết đến các bản tin khác ở đầu bản tin cũng như các hình ảnh trong bản tin. Các bản tin trên các báo Dantri và VnExpress được lấy ở các thể loại Chính trị, Khoa học giáo dục, Kinh tế, Văn hóa, Xã hội. Trong thư mục Plaintext có các thư mục con tương ứng với các thể loại đó.*

*+ Thư mục Summary\_manual chứa các bản tóm tắt. Mỗi bản tóm tắt có cùng tên với văn bản gốc trong thư mục Plaintext. Thư mục này cũng được chia thành các mục thư con tương ứng với các thư mục ở Plaintext.*

*Các bản tóm tắt được tạo ra một cách thủ công. Mỗi văn bản gốc có một bản tóm tắt tương ứng. Các bản tóm tắt có độ dài xấp xỉ 100 từ.*

## 3.2 Xử lý dữ liệu

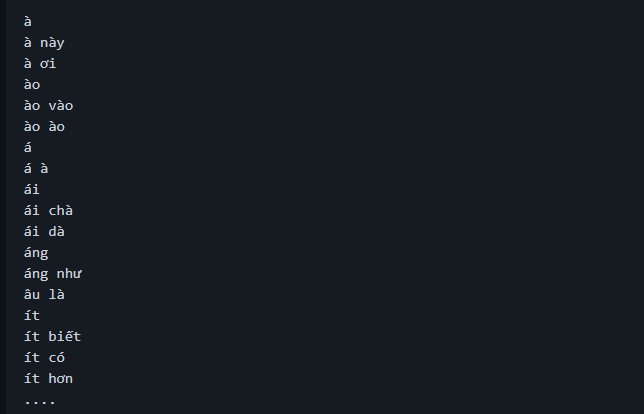
Dể dàng thấy được, văn bản đầu vô của chúng em chưa được chuẩn hóa.



#### Hình 17: Trích 1 đoạn nhỏ trong văn bản gốc – plaintext

Suy ra, chúng ta phải tiền xử lý nó và các bước tiền xử lý bao gồm các bước sau:

- Loại bỏ các từ dừng trong văn bản tiếng Việt, được chúng em dùng 1 tập từ dừng có sẵn để dò và loại từ động



## 3.3 Công nghệ sử dụng

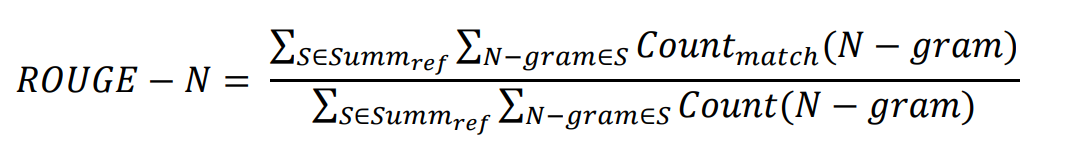
* Ngôn ngữ sử dụng: Python 3.9.4
* Trình soạn thảo: Google Colab
* Word Tokenizer: PyVi
* Word2Vec train: genism
* Ngrams: nltk
* Đọc và ghi file: Pandas
* Framwork Machine Learning: Keras

Specs – Cấu hình máy tính

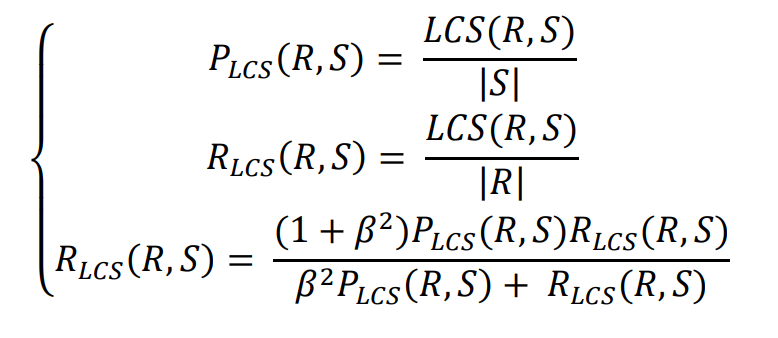
* Hiện tại: GPU và CPU, bộ nhớ được hỗ trợ bởi Google Colab
* Trong tương lai: Chi tiết ở mục *4.2 Hướng phát triển*

## 3.4 Cách đánh giá

Như đã đề cập ở trên, chúng em sẽ sử dụng độ đo ROUGE. Nó bao gồm năm độ đo, để xác định chất lượng bản tóm tắt bởi máy so với bản tóm tắt bởi con người, đó là: ROUGE-N, ROUGE-L, ROUGE-W, ROUGE-S và ROUGE-SU. Sự đo lường thực hiện bởi số lượng đơn vị trùng lặp như N-grams, chuỗi các từ, cặp các từ giữa văn bản tóm tắt ứng cử và văn bản tóm tắt dẫn xuất. ROUGE-N ước lượng độ phủ N-grams giữa văn bản tóm tắt ứng cử và văn bản tóm tắt dẫn xuất.



Trong đó N là chiều dài của N-grams, Countmatch(N-gram) là số lượng lớn nhất N-grams cùng xuất hiện giữa hai bản tóm tắt tương ứng, Count(N-gram) là số lượng N-grams trong trong văn bản tóm tắt dẫn xuất.



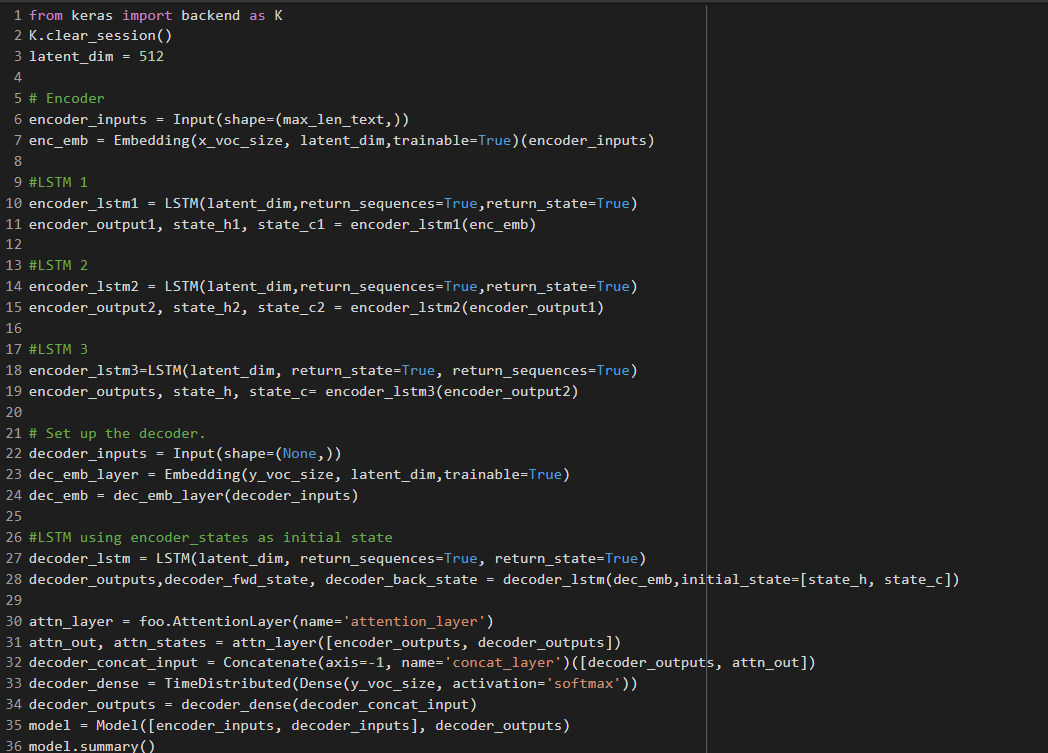
Trong đó: |R| và|S| tương ứng là chiều dài văn bản dẫn xuất R và văn bản ứng viên S. LCS(R,S) là LCS giữa R và S. PLCS(R,S) là độ chính xác của LCS(R,S) và RLCS(R,S) là độ phủ của LCS(R,S). β là PLCS(R,S) / RLCS(R,S).

## 3.5 Kết quả đạt được

* + Tổng cộng chúng em có 2 lớp encoder và decoder, encoder sử dụng 3 block LSTM và 1 lớp Dense – Fully Connected để sẵn sàng cho lớp decoder.
  + Lượng epoch chạy qua là 101
  + Chạy qua 1 phân tầng Attention

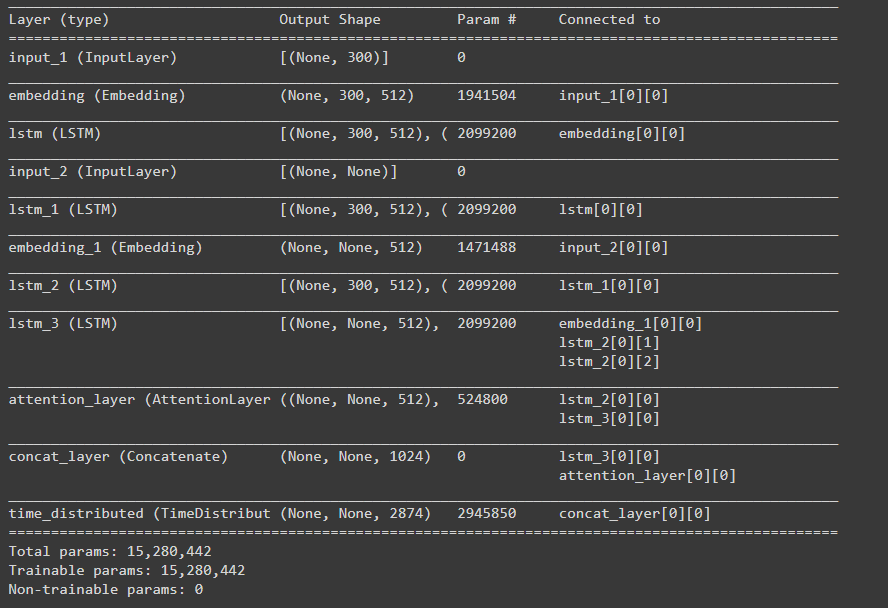
Trình bày code của các module chính

Trình bày các kết quả đạt được theo các độ đo ở trên và trên các bộ dữ liệu hay các phương pháp học máy, học sâu khác nhau.

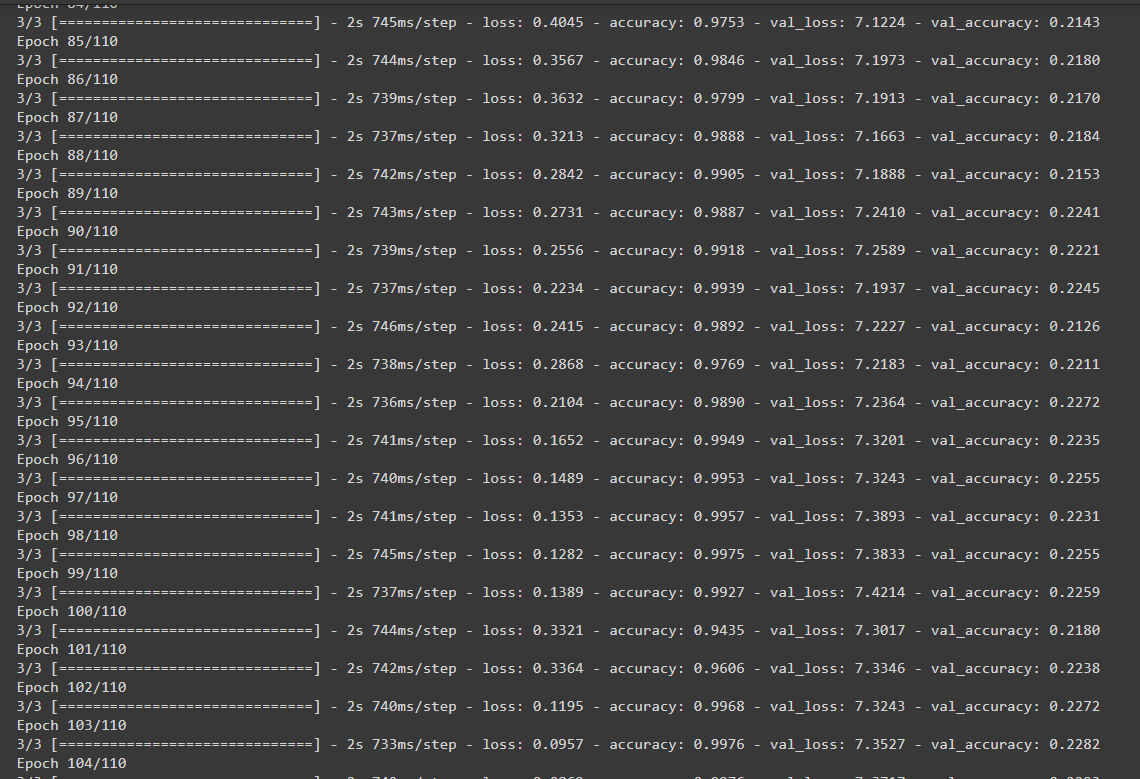


*Hình 18: Trích code model chính*

Cho ra kết quả như sau:



Độ do Accuracy được gia tăng khi chạy qua các epoch



# PHẦN IV

# KẾT LUẬN

## 4.1 Kết quả đạt được

Trình bày tóm tắt các kết quả đạt được của bài toán trên,

Ứng dụng độ đo ROUGE trong model chúng em nhận được bảng sau:

|  |  |  |  |
| --- | --- | --- | --- |
| ROUGE | 1 | 2 | L |
| Precision | 42.5 | 12.2 | 25.2 |
| Recall | 43.6 | 12.3 | 24.5 |
| F-score | 40.2 | 12.02 | 23.5 |

#### Bảng 1: Bảng đánh giá điểm số ROUGE mô hình

Dataset còn hạn chế và trích xuất chưa rõ rang nên độ đo ROUGE trả về là khá ổn định nhưng theo so sánh với các mô hình khác thì điểm số này chưa cao

## 4.2 Hướng phát triển

* Vì bộ dataset lần này còn khá nhỏ để đạt được kết quả mong đợi nên chúng ta sẽ thu thập nhiều dữ liệu hơn
* Thay thế LSTM stack bằng các Bi-Directional LSTM
* Dùng chiến thuật Beam Search cho phần decode sequence thay vì là chiến lược tham lam – argmax
* Kết nối hệ thống với UI như đã dự tính bằng NodeJS

# TÀI LIỆU THAM KHẢO

1. FloydHub - *“*[*A Gentle introduction to Text Summarization in Machine Learning*](https://blog.floydhub.com/gentle-introduction-to-text-summarization-in-machine-learning/)*”*
2. Nguyễn Viết Hạnh *- Luận văn thạc sĩ công nghệ thông tin về NGHIÊN CỨU TÓM TẮT VĂN BẢN TỰ ĐỘNG VÀ ỨNG DỤNG*
3. Nguyễn Mạnh Cường *– Luận văn thạc sĩ khoa học máy tính về TÓM TẮT VĂN BẢN TIẾNG VIỆT TỰ ĐỘNG DỰA TRÊN MÔ HÌNH ĐỒ THỊ*
4. Github Repo của nhóm làm việc – [*Vietnamese-Text-summarization*](https://github.com/NhutNguyen236/Vietnamese-Text-summarization)
5. Lê Thanh Hương - [*Vietnamese text summarization corpus (kho ngữ liệu tóm tắt văn bản tiếng Việt)*](https://users.soict.hust.edu.vn/huonglt/)
6. Nguyễn Minh Thành – [*Báo cáo về tóm tắt văn bản tự động*](https://sites.google.com/site/trangmonhocitc/text-summarization)
7. Hobson Lane, Cole Howard, Hannes Max Hapke *- Natural Language Processing in action*
8. Maofu Liu, Wenje Li, Mingli Wu, Quin Lu – [*Extractive Summarization based on Event term clustering*](https://www.aclweb.org/anthology/P07-2047.pdf)
9. Trần Mai Vũ – [*Hệ thống tóm tắt sự kiện*](https://eprints.uet.vnu.edu.vn/eprints/id/eprint/3716/1/Event%20Summarization%20System%20%28Technical_report%29.pdf)
10. Johan Hasselqvist, Niklas Helmertz, Mikael Kageback – [*Query Based abstractive Summarization using neural networks*](https://arxiv.org/pdf/1712.06100.pdf)
11. Trần Mai Vũ – [*Tóm tắt đa văn bản dựa trân trích xuất câu*](http://www.uet.vnu.edu.vn/~thuyhq/Student_Thesis/MSc10_Tran_Mai_Vu_Thesis.pdf)
12. Đoàn Xuân Dũng - [*TÓM TẮT VĂN BẢN SỬ DỤNG CÁC KỸ THUẬT TRONG DEEP LEARNING*](http://lib.uet.vnu.edu.vn/bitstream/123456789/956/3/Tomtat.pdf)

# TỰ ĐÁNH GIÁ

(Với nhóm có 2 thành viên)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Phần | Nội dung | Điểm chuẩn | Tự chấm | Ghi chú |
| I (1) | Phân tích yêu cầu của bài toán | 1 | 1 |  |
| II (3.25) | Phương pháp giải quyết bài toán | | | |
|  | 2.1 Các phương pháp giải quyết bài toán | 1 | 1 |  |
|  | 2.2 Phương pháp đề xuất giải quyết bài toán | 0.25 | 0.25 |  |
|  | 2.3 Phương pháp giải quyết bài toán | 2 | 1.75 |  |
| III (4.75) | Thực nghiệm | | | |
|  | 3.1 Dữ liệu | 0.5 | 0.5 |  |
|  | 3.2 Xử lý dữ liệu | 1 | 1 |  |
|  | 3.3 Công nghệ sử dụng | 0.25 | 0.25 |  |
|  | 3.4 Cách đánh giá | 0.5 | 0.5 |  |
|  | **3.5 Kết quả đạt được** | 2.5 | 1.75 |  |
| **IV (0.25)** | **Kết luận** | 0.25 | 0.25 |  |
|  | Điểm nhóm | 0.5 | 0.5 |  |
|  | Báo cáo theo mẫu | 0.25 | 0.25 |  |
| Tổng điểm đạt được | | 9 | |  |