**TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI**

**BÁO CÁO TỔNG KẾT**

**ĐỀ TÀI NGHIÊN CỨU KHOA HỌC CỦA SINH VIÊN**

**NĂM 2021**

**XÂY DỰNG HỆ THỐNG TÓM TẮT NHẬN XÉT**

**CỦA NGƯỜI DÙNG TRÊN CÁC TRANG THƯƠNG MẠI ĐIỆN TỬ**

**DỰA TRÊN MÔ HÌNH HỌC SÂU**

**Sinh viên thực hiện:**

**Lê Minh Hiếu Lớp: CNTT 2 Khoa: Công Nghệ Thông Tin**

**Người hướng dẫn:** TS. Nguyễn Hiếu Cường

**TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI**

**BÁO CÁO TỔNG KẾT**

**ĐỀ TÀI NGHIÊN CỨU KHOA HỌC CỦA SINH VIÊN**

**NĂM 2021**

**XÂY DỰNG HỆ THỐNG TÓM TẮT NHẬN XÉT**

**CỦA NGƯỜI DÙNG TRÊN CÁC TRANG THƯƠNG MẠI ĐIỆN TỬ**

**DỰA TRÊN MÔ HÌNH HỌC SÂU**

**Sinh viên thực hiện:**

Lê Minh Hiếu Nam, Nữ: Nam Dân tộc: Kinh

Lớp: CNTT 2 Khoa: Công Nghệ Thông Tin Năm thứ: 3/4

Ngành học: Công Nghệ Thông Tin

**Người hướng dẫn:** TS. Nguyễn Hiếu Cường

**MỤC LỤC**

[**1.** **Mở đầu** 7](#_Toc78547709)

[**1.1.** **Giới thiệu đề tài** 7](#_Toc78547710)

[**1.2.** **Các hướng tiếp cận tóm tắt văn bản** 9](#_Toc78547711)

[**1.3.** **Một số phương pháp tóm tắt văn bản** 10](#_Toc78547712)

[**1.4.** **Lý do chọn đề tài** 11](#_Toc78547713)

[**1.5.** **Mục tiêu của đề tài** 12](#_Toc78547714)

[**2.** **Kết quả nghiên cứu và phân tích kết quả** 13](#_Toc78547715)

[**2.1.** **Cơ sở lý thuyết** 13](#_Toc78547716)

[2.1.1. Mạng nơ-ron nhân tạo 13](#_Toc78547717)

[2.1.2. Mạng nơ-ron hồi quy RNN 14](#_Toc78547718)

[2.1.3. Mạng LSTM 15](#_Toc78547719)

[**2.2. Mô hình đề xuất** 17](#_Toc78547720)

[2.2.1. Mô hình Sequence to Sequence 17](#_Toc78547721)

[2.2.2. Cơ chế Attention 18](#_Toc78547722)

[2.2.3. Xây dựng mô hình 19](#_Toc78547723)

[**2.3.** **Kết quả** 20](#_Toc78547724)

[2.3.1. Quá trình huấn luyện 20](#_Toc78547725)

[2.3.2. Phân tích kết quả 22](#_Toc78547726)

[2.3.3. Triển khai model lên hệ thống sử dụng Flask API 22](#_Toc78547727)

[2.3.4. Tích hợp tính năng dịch văn bản, nhận diện giọng nói 23](#_Toc78547728)

[**3.** **Kết luận và kiến nghị** 24](#_Toc78547729)

[**3.1. Kết luận** 24](#_Toc78547730)

[**3.2. Kiến nghị** 24](#_Toc78547731)

[**Tài liệu tham khảo** 25](#_Toc78547732)

**DANH MỤC BẢNG BIỂU**

|  |  |  |
| --- | --- | --- |
| **STT** | **Tên danh mục** | **Trang** |
| 1 | Hình 1.1.1: Tổng quan về ngành trí tuệ nhân tạo | 8 |
| 2 | Hình 1.1.2: Tóm tắt văn bản | 8 |
| 3 | Hình 1.3.1. Phương pháp tóm tắt trích chọn | 10 |
| 4 | Hình 1.3.2. Phương pháp tóm tắt tóm lược ý | 11 |
| 5 | Hình 1.4.1: Ứng dụng của tóm tắt văn bản | 11 |
| 6 | Hình 1.5.1: Mục tiêu của đề tài | 12 |
| 7 | Hình 2.1.1: Mạng lưới thần kinh | 13 |
| 8 | Hình 2.1.2: Mạng nơ-ron hồi quy RNN | 14 |
| 9 | Hình 2.1.3. Mô hình LSTM | 16 |
| 10 | Hình 2.2.1: Mô hình Sequence to Sequence | 17 |
| 11 | Hình 2.2.2: Các thành phần trong Attention | 18 |
| 12 | Hình 2.2.3: Mô hình tóm tắt nhận xét của người dùng | 19 |
| 13 | Hình 2.3.1.1: Dữ liệu Amazon Fine Food | 20 |
| 14 | Hình 2.3.1.2: Thư viện NLTK sử dụng trong lĩnh vực NLP | 21 |
| 15 | Hình 2.3.1.3: Mô tả quá trình cập nhật loss quá trình huấn luyện | 21 |
| 16 | Hình 2.3.2.1: Phân tích kết quả của mô hình | 22 |
| 17 | Hình 2.3.2.2: Đánh giá độ chính xác của mô hình | 22 |
| 18 | Hình 2.3.3: Triển khai mô hình lên hệ thống | 23 |
| 19 | Hình 2.3.4: Tính năng nhận diện giọng nói | 23 |

**TRƯỜNG ĐẠI HỌC GIAO THÔNG VẬN TẢI**

**THÔNG TIN KẾT QUẢ NGHIÊN CỨU CỦA ĐỀ TÀI**

**1. Thông tin chung:**

**-** Tên đề tài: Xây dựng hệ thống tóm tắt nhận xét của người dùng trên các trang thương mại điện tử dựa trên mô hình học sâu.

- Sinh viên thực hiện: Lê Minh Hiếu

- Lớp: CNTT 2 Khoa: Công Nghệ Thông Tin

Năm thứ: 3 Số năm đào tạo: 4

- Người hướng dẫn: TS. Nguyễn Hiếu Cường

**2. Mục tiêu đề tài:**

Phát triển một công nghệ mới tự động tóm tắt các nhận xét, đánh giá của người dùng trên các trang thương mại điện tử, giúp con người giảm thiểu thời gian đọc thông tin mà vẫn có thể nắm bắt được những nội dung chính của văn bản.

**3. Tính mới và sáng tạo:**

Ứng dụng các kỹ thuật vào mô hình học sâu trong lĩnh vực xử lý ngôn ngữ tự nhiên giúp hệ thống phân tích, xử lý, và hiểu ngôn ngữ con người. Ngoài ra, hệ thống có thể thu được nhận xét, đánh giá của người dùng thông qua chức năng nhận diện giọng nói.

**4. Kết quả nghiên cứu:**

Xây dựng được mô hình tóm tắt những nhận xét, đánh giá của người dùng trên các trang thương mại điện tử thu được kết quả sát với văn bản tóm tắt do con người thực hiện.

**5. Đóng góp về mặt kinh tế - xã hội,** **giáo dục và đào tạo, an ninh, quốc phòng và khả năng áp dụng của đề tài:**

Hệ thống có thể được ứng dụng trực tiếp vào các trang thương mại điện tử giúp người dùng thay vì đọc toàn bộ nội dung của các nhận xét, chỉ cần đọc đoạn tóm tắt, ngoài ra có thể ứng dụng vào việc tóm tắt biên bản cuộc họp, đặt tiêu đề bài báo, …

**6.** **Công bố khoa học của sinh viên từ kết quả nghiên cứu của đề tài** *(ghi rõ họ tên tác giả, nhan đề và các yếu tố về xuất bản nếu có)* hoặc nhận xét, đánh giá của cơ sở đã áp dụng các kết quả nghiên cứu *(nếu có)*:

Ngày tháng năm

|  |  |
| --- | --- |
|  | **Sinh viên chịu trách nhiệm chính**  **thực hiện đề tài**  *(ký, họ và tên)* |

**Nhận xét của người hướng dẫn về những đóng góp khoa học của sinh viên thực hiện đề tài** *(phần này do người hướng dẫn ghi):*

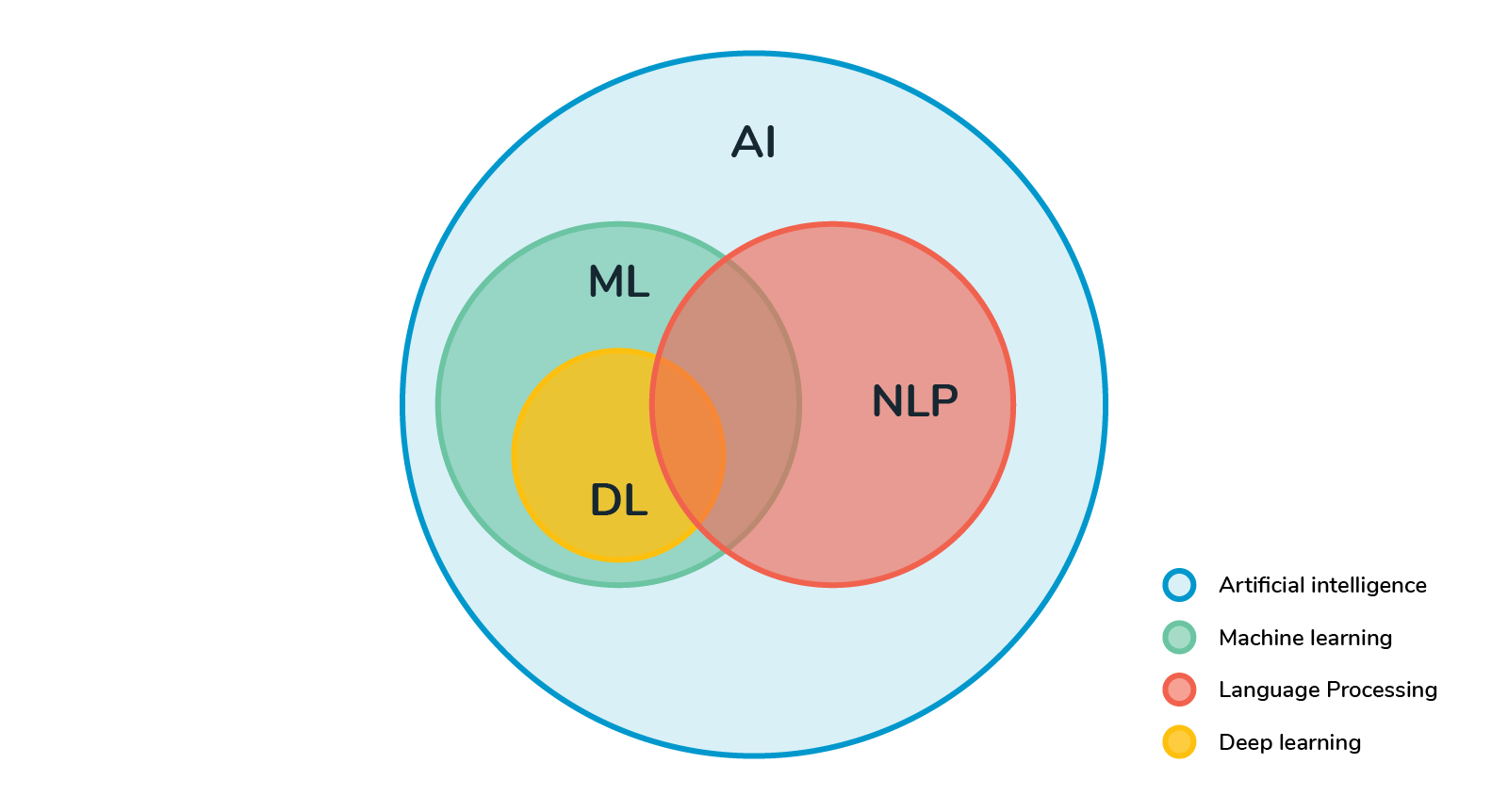
Ngày tháng năm

|  |  |
| --- | --- |
|  | **Người hướng dẫn**  *(ký, họ và tên)* |

1. **Mở đầu**
   1. **Giới thiệu đề tài**

Những năm gần đây, AI - Artificial Intelligence (Trí Tuệ Nhân Tạo), và cụ thể hơn là Machine Learning (Máy Học) nổi lên như một minh chứng của cuộc cách mạng công nghiệp lần thứ tư (1 - động cơ hơi nước, 2 - năng lượng điện, 3 - công nghệ thông tin). AI hiện diện trong mọi lĩnh vực của đời sống con người, từ kinh tế, giáo dục, y khoa cho đến những công việc nhà, giải trí hay thậm chí là trong quân sự. Những ứng dụng nổi bật trong việc phát triển AI đến từ nhiều lĩnh vực để giải quyết nhiều vấn đề khác nhau. AI đã giúp máy tính thực thi những việc tưởng chừng như không thể vào 15 năm trước: phân loại cả ngàn vật thể khác nhau trong các bức ảnh, tự tạo chú thích cho ảnh, bắt chước giọng nói và chữ viết của con người, giao tiếp với con người, hay thậm chí cả sáng tác văn, phim, ảnh, âm nhạc.

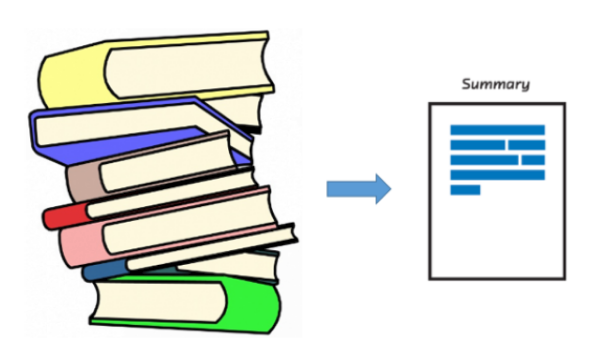
Một trong những mong muốn mãnh liệt, xuất hiện từ rất sớm của các nhà khoa học máy tính nói chung và trí tuệ nhân tạo nói riêng là xây dựng thành công các hệ thống, chương trình máy tính có khả năng giao tiếp với con người thông qua ngôn ngữ tự nhiên (natural language), tức thứ ngôn ngữ con người sử dụng hàng ngày thay vì các ngôn ngữ lập trình (programming language) hay ngôn ngữ máy (computer language) bậc thấp. Xử lý ngôn ngữ tự nhiên (natural language processing), một nhánh nghiên cứu của trí tuệ nhân tạo, trong đó phát triển các thuật toán, xây dựng các chương trình máy tính có khả năng phân tích, xử lý, và hiểu ngôn ngữ của con người. Xử lý ngôn ngữ tự nhiên có thể được chia ra thành hai nhánh lớn, không hoàn toàn độc lập, bao gồm xử lý tiếng nói (speech processing) và xử lý văn bản (text processing). Xử lý tiếng nói tập trung nghiên cứu, phát triển các thuật toán, chương trình máy tính xử lý ngôn ngữ của con người ở dạng tiếng nói (dữ liệu âm thanh). Các ứng dụng quan trọng của xử lý tiếng nói bao gồm nhận dạng tiếng nói và tổng hợp tiếng nói. Nếu như nhận dạng tiếng nói là chuyển ngôn ngữ từ dạng tiếng nói sang dạng văn bản thì ngược lại, tổng hợp tiếng nói chuyển ngôn ngữ từ dạng văn bản thành tiếng nói. Xử lý văn bản tập trung vào phân tích dữ liệu văn bản. Các ứng dụng quan trọng của xử lý văn bản bao gồm tìm kiếm và truy xuất thông tin, dịch máy, tóm tắt văn bản tự động, hay kiểm lỗi chính tả tự động. Xử lý văn bản đôi khi được chia tiếp thành hai nhánh nhỏ hơn bao gồm hiểu văn bản và sinh văn bản. Lĩnh vực này được xem là một quy trình phức tạp bởi công nghệ tích hợp các quy trình giúp máy tính hiểu, giải thích và mô phỏng ngôn ngữ của con người bằng cách học hỏi từ nhiều lĩnh vực khác nhau, bao gồm ngôn ngữ học máy tính và ngôn ngữ học tính toán. Với nỗ lực rút ngắn khoảng cách giữa cách giao tiếp của con người và sự hiểu biết của máy, công nghệ này ngày càng phát triển với nhiều ứng dụng thực tiễn trong đời sống như Robot, Chatbot hay Callbot giúp tự động hoá quá trình giao tiếp, tương tác, quảng bá của doanh nghiệp với khách hàng hay dịch máy, tóm tắt văn bản tự động, kiểm lỗi chính tả tự động, …



*Hình 1.1.1: Tổng quan về ngành trí tuệ nhân tạo*

Ngày nay, con người đang bước vào kỷ nguyên của cách mạng công nghiệp 4.0 với sự phát triển mạnh mẽ của công nghệ thông tin và mạng máy tính, thông tin được sinh ra liên tục mỗi ngày trên Internet, lượng thông tin văn bản khổng lồ này đã và đang mang lại lợi ích không nhỏ cho con người, tuy nhiên, nó cũng khiến chúng ta khó khăn trong việc tìm kiếm và tổng hợp thông tin. Để tăng hiệu quả cũng như dễ dàng hơn trong việc tiếp nhận thông tin của người dùng, 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 là tóm tắt văn bản tự động.

Tóm tắt văn bản là quá trình rút trích những thông tin quan trọng nhất từ một văn bản để tạo ra phiên bản ngắn gọn, xúc tích mang lại đầy đủ lượng thông tin của văn bản gốc kèm theo đó là tính đúng đắn về ngữ pháp và chính tả. Bản tóm tắt phải giữ được những thông tin quan trọng của toàn bộ văn bản chính và bỏ qua các thông tin dư thừa. Bên cạnh đó, bản tóm tắt cần phải có bố cục chặt chẽ, tính đến các thông số như độ dài câu, phong cách viết và cú pháp của văn bản. Đây được xem là một bài toán thuộc lĩnh vực khái phá dữ liệu văn bản; việc áp dụng tóm tắt văn bản sẽ giúp người dùng tiết kiệm thời gian đọc, cải thiện tìm kiếm cũng như tăng hiệu quả đánh chỉ mục cho máy tìm kiếm.



*Hình 1.1.2: Tóm tắt văn bản*

* 1. **Các hướng tiếp cận tóm tắt văn bản**

Trong nhiều năm qua, từ nhu cầu thực tế, các nhà khoa học thuộc ngành Xử lý Ngôn ngữ Tự nhiên (NLP), những nhóm nghiên cứu cũng như các công ty lớn trên thế giới đã mong muốn phát triển một chương trình tóm tắt văn bản thực sự hiệu quả. Do đó, họ đã có rất nhiều cách tiếp cận bài toán khác nhau, tóm tắt văn bản có thể được phân loại dựa trên đầu vào (đơn hay đa văn bản), mục đích (tổng quát, theo lĩnh vực cụ thể, hay dựa trên truy vấn) và loại đầu ra (trích chọn hay tóm lược).

Đầu tiên là tóm tắt dựa vào mục đích hay chức năng như: tóm tắt theo chỉ thị (Indicative) là tóm tắt nhằm cung cấp một chức năng tham khảo để chọn tài liệu đọc chi tiết hơn (ứng dụng trong tóm tắt kết quả tìm kiếm). Ví dụ: Trong tóm tắt tin tức, tóm tắt đưa ra chi tiết chính của từng sự kiện. Tóm tắt thông tin là tóm tắt bao gồm tất cả các thông tin nổi bật có trong văn bản nguồn tại nhiều mức độ chi tiết khác nhau. Tóm tắt đánh giá (Evaluation) để đánh giá vấn đề chính của văn bản nguồn, thể hiện quan điểm của tác giả đối với công việc của họ.

Hướng tiếp cận thứ hai là tóm tắt theo nội dung văn bản gồm có tóm tắt chung và tóm tắt hướng truy vấn. Tóm tắt chung (Generalize) 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 gốc. Tóm tắt hướng truy vấn (Query-based) thì kết quả dựa vào câu truy vấn của người dùng hay chương trình đưa vào. Tóm tắt này thường được sử dụng trong quá trình tìm kiếm thông tin.

Hướng tiếp theo là tóm tắt dựa trên miền dữ liệu như cụ thể về một lĩnh vực nào đó. Tóm tắt trên một miền dữ liệu (domain) là tóm tắt nhắm vào một miền nội dung cụ thể (như tin tức thời sự, tin tức tài chính, …). Tóm tắt trên một thể loại (Genre): tóm tắt nhắm vào một thể loại văn bản nào đó, như báo chí, email, web, bài báo, … Tóm tắt độc lập (Independent): tóm tắt cho nhiều thể loại và trên nhiều miền dữ liệu.

Bên cạnh đó, tóm tắt cũng có thể dựa vào mức độ chi tiết. Tóm tắt tổng quan (overview) để miêu tả tổng quan tất cả các nội dung nổi bật trong văn bản nguồn. Tóm tắt tập trung sự kiện là tóm tắt miêu tả một sự kiện cụ thể nào đó trong văn bản nguồn.

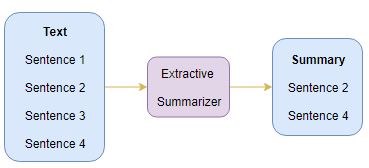
Một hướng tiếp cận khác là phụ thuộc vào số lượng các văn bản, kỹ thuật tóm tắt có thể chia làm hai lớp: đơn văn bản và đa văn bản. Tóm tắt đơn văn bản chỉ đơn giản là rút gọn một văn bản thành một sự trình bày ngắn gọn. Tóm tắt đa văn bản có thể xem như một sự mở rộng của tóm tắt đơn văn bản và thường dùng với thông tin chứa trong các cụm văn bản, để người dùng có thể hiểu được cụm văn bản đó. Tóm tắt đa văn bản phức tạp hơn tóm tắt đơn văn bản vì phải làm việc trên số lượng văn bản nhiều hơn.

Ngoài ra, còn một số hướng tiếp cận như tóm tắt theo đơn ngôn ngữ, đa ngôn ngữ, xuyên ngôn ngữ (cross-language), …

* 1. **Một số phương pháp tóm tắt văn bản**

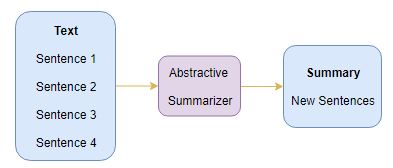
Tóm tắt là một việc khá tốn thời gian, cũng như chất xám. Cùng một bài viết, mỗi người sẽ có một cách tóm tắt khác nhau, có cách hay, có cách dở, có cách dễ hiểu, cũng có cách khó hiểu, vì thế mà cùng với sự phát triển của AI, tóm tắt văn bản là một trong những công việc mà các nhà nghiên cứu luôn muốn giải đáp. Nhưng đây không phải là một việc đơn giản, việc tóm tắt được đánh giá là một công việc có độ khó cao cũng như nhiều thách thức trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP). Xét về phương pháp thực hiện, tóm tắt văn bản có hai phương pháp chính gồm tóm tắt theo kiểu trích chọn (Extractive Summarizer) và tóm tắt theo kiểu tóm lược ý (Abstractive Summarizer).

Phương pháp tóm tắt trích chọn là công việc chọn ra một tập con những từ đã có, những lời nói hoặc những câu của văn bản gốc và kết quả là một tóm tắt bao gồm toàn bộ các phần quan trọng được trích ra từ văn bản đầu vào. 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 văn bản gốc. Các hệ thống tóm tắt văn bản theo hướng trích chọn thường gồm các tác vụ: xây dựng một đại diện trung gian (intermediate representation) của văn bản đầu vào thể hiện các đặc điểm chính của văn bản, tính điểm (xếp hạng) các câu dựa trên đại diện trung gian đã xây dựng, chọn các câu đưa vào tóm tắt. Đầu tiên, mỗi hệ thống tóm tắt văn bản sẽ tạo ra một số đại diện trung gian của văn bản mà nó thực hiện tóm tắt và tìm các nội dung nổi bật dựa trên đại diện trung gian này. Khi các đại diện trung gian đã được tạo ra, một điểm số thể hiện mức độ quan trọng sẽ được gán cho mỗi câu. Cuối cùng hệ thống tóm tắt có thể áp dụng các thuật toán tham lam để lựa chọn các câu quan trọng nhất từ văn bản gốc.

****

*Hình 1.3.1. Phương pháp tóm tắt trích chọn*

Trong khi đó, các phương pháp tóm tắt không sử dụng trích xuất để tạo ra tóm tắt có thể xem như là một phương pháp tiếp cận tóm tắt theo tóm lược. Phương pháp tóm tắt tóm lược xây dựng một biểu diễn ngữ nghĩa bên trong và sau đó sử dụng kỹ thuật xử lý ngôn ngữ để tạo ra bản tóm tắt gần gũi hơn so với những gì con người có thể tạo ra. Bản tóm tắt như vậy có thể không gồm các từ hay các cụm từ trong văn bản gốc giúp nó có tính kết dính, dễ đọc và đúng ngữ pháp. Tóm tắt tóm lược được chia theo cách tiếp cận cấu trúc, theo cách tiếp cận ngữ nghĩa và gần đây là theo cách tiếp cận học sâu. Hệ thống tóm tắt sẽ cố gắng hiểu và đánh giá văn bản sử dụng các kỹ thuật xử lý ngôn ngữ tự nhiên tiên tiến để tạo ra một văn bản ngắn hơn, truyền đạt được những thông tin quan trọng nhất từ văn bản gốc.

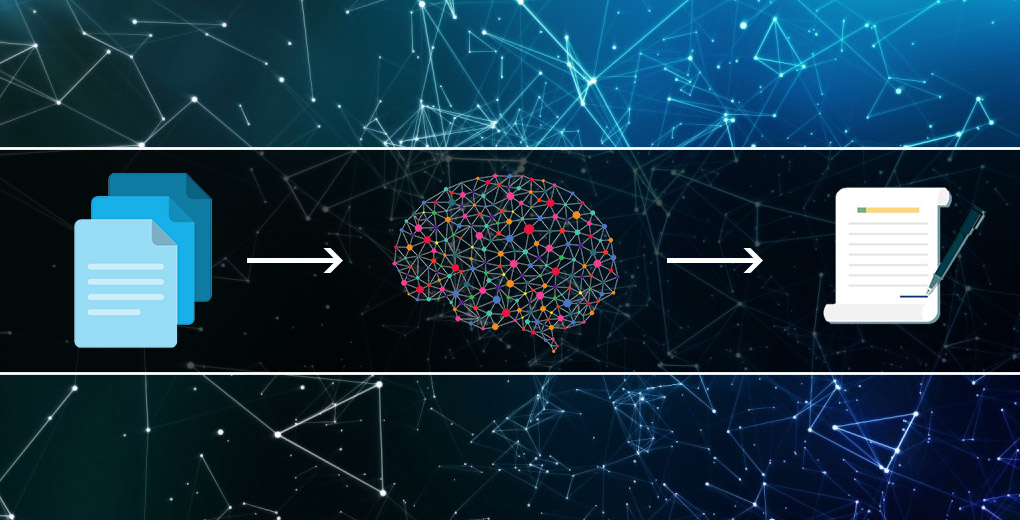
****

*Hình 1.3.2. Phương pháp tóm tắt tóm lược ý*

* 1. **Lý do chọn đề tài**

Tóm tắt nội dung văn bản là một trong những bài toán thú vị và phức tạp nhất hiện nay thuộc lĩnh vực khai phá dữ liệu văn bản. Với cuộc sống ngày càng bận rộn như ngày nay thì mọi người thường thích đọc các bài tóm tắt với nội dung ngắn gọn hơn là đọc toàn bộ cả bài. Bên cạnh đó, việc đọc một bản tóm tắt sẽ cung cấp cho người đọc một cái nhìn tổng quan về câu chuyện và nó có thể sử dụng để chỉ định các chủ đề quan trọng nhất của tài liệu. Xuất phát từ việc con người ngày càng mất nhiều thời gian đọc email, báo điện tử và mạng xã hội, việc ứng dụng các thuật toán sử dụng machine learning để tự động tóm tắt các văn bản dài một cách gãy gọn và chính xác ngày càng trở nên cần thiết và có vai trò to lớn đối trong bất kỳ lĩnh vực nào. Tự động tóm tắt sẽ là một trong những công nghệ quan trọng có thể giúp con người trong rất nhiều lĩnh vực khác nhau như: tóm tắt tin tức, tóm tắt bệnh án hỗ trợ bác sĩ trong điều trị, tóm tắt kết quả tìm kiếm trong các search engine, tóm tắt bài báo khoa học, tóm tắt nội dung hội nghị, cuộc họp, ...

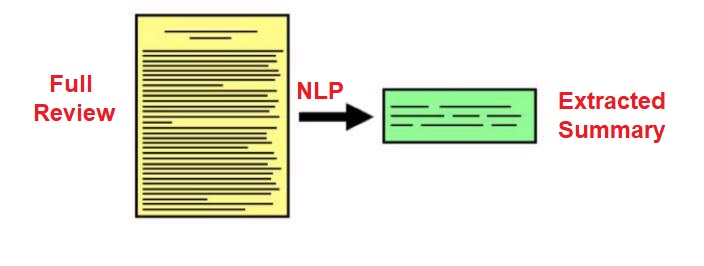
Từ nhu cầu thực tế như thế, có thể thấy bài toán “Xây dựng hệ thống tóm tắt tự động nhận xét, đánh giá của người dùng trên các trang thương mại điện tử” là một đề tài có tính ứng dụng vào thực tiễn cao.



*Hình 1.4.1: Ứng dụng của tóm tắt văn bản*

* 1. **Mục tiêu của đề tài**

Mục tiêu của đề tài là nghiên cứu sử dụng học sâu cho bài toán tóm tắt văn bản tự động, sử dụng mô hình Sequence to sequence with Attention nhằm tạo kết quả đầu ra là chuỗi các từ. Cuối cùng kết quả thu được là các câu tóm tắt phản ánh đúng nội dung văn bản. Từ đó xây dựng hệ thống tự động tóm tắt những câu nhận xét, đánh giá của người dùng, tạo ra bản tóm tắt giống như cách con người thực hiện, giúp người sử dụng giảm thiểu lượng lớn thời gian đọc và có cái nhìn tổng quan về sản phẩm trên các trang thương mại điện tử.



***Hình 1.5.1: Mục tiêu của đề tài***

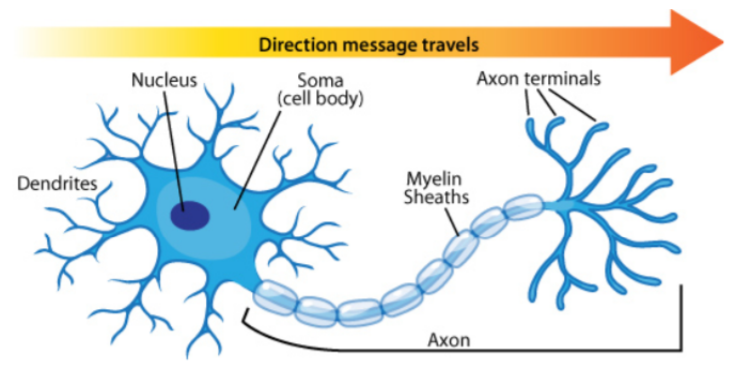
1. **Kết quả nghiên cứu và phân tích kết quả**
   1. **Cơ sở lý thuyết**

2.1.1. Mạng nơ-ron nhân tạo

Mạng nơ-ron nhân tạo (Neural Network - NN) là một mô hình lập trình được lấy cảm hứng từ mạng nơ-ron thần kinh. Kết hợp với các kĩ thuật học sâu (Deep Learning - DL), NN đang trở thành một công cụ rất mạnh mẽ mang lại hiệu quả tốt nhất cho nhiều bài toán khó như nhận dạng ảnh, giọng nói hay xử lý ngôn ngữ tự nhiên.

Mạng lưới là một mô phỏng xử lý thông tin, được nghiên cứu ra từ hệ thống thần kinh của con người, giống như bộ não để xử lý thông tin. Nó bao gồm số lượng lớn các mối gắn kết cấp cao để xử lý các thông tin trong mối liên hệ rõ ràng. Nó có khả năng học bởi kinh nghiệm từ huấn luyện, lưu những kinh nghiệm thành tri thức và áp dụng trong những dữ liệu mới trong tương lai.

Cấu trúc cơ bản của một mạng nơ-ron nhân tạo là một mạng lưới các tế bào nhỏ, hoặc nút, tham gia với nhau bởi các kết nối trọng số. Xét về mặt mô hình sinh học gốc, các nút đại diện cho tế bào nơ-ron, và các trọng số kết nối đại diện cho sức mạnh của các khớp nơ-ron giữa các tế bào nơ-ron. Các mạng kích hoạt bằng cách cung cấp một đầu vào cho một số hoặc tất cả các nút, và kích hoạt này sau đó lây lan khắp các mạng cùng các kết nối trọng số.



*Hình 2.1.1: Mạng lưới thần kinh*

2.1.2. Mạng nơ-ron hồi quy RNN

Deep learning có 2 mô hình lớn là Convolutional Neural Network (CNN) cho bài toán có input là ảnh và Recurrent neural network (RNN) cho bài toán dữ liệu dạng chuỗi (sequence). Trong mô hình mạng nơ-ron thông thường, chúng ta coi input data là các dữ liệu độc lập, không có mối liên hệ với nhau. Tuy nhiên, trong ngôn ngữ tự nhiên thì mối liên hệ giữa các từ và ngữ cảnh đóng một vai trò quan trọng, quyết định ý nghĩa của câu văn. Do đó việc áp dụng môt hình mạng Neuron thông thường vào các bài toán NLP thường không đạt kết quả mong muốn. Để khắc phục nhược điểm này, chúng ta sử dụng mô hình RNN (Recurrent Neural Network). RNN coi dữ liệu đầu vào là một chuỗi liên tục, nối tiếp nhau theo thứ tự thời gian.

Ý tưởng chính của RNN là sử dụng chuỗi các thông tin. Trong các mạng nơ-ron truyền thống tất cả các đầu vào và cả đầu ra là độc lập với nhau. RNN được gọi là hồi quy (Recurrent) bởi lẽ chúng thực hiện cùng một tác vụ cho tất cả các phần tử của một chuỗi với đầu ra phụ thuộc vào cả các phép tính trước đó. Nói cách khác, RNN có khả năng nhớ các thông tin được tính toán trước đó.



*Hình 2.1.2: Mạng nơ-ron hồi quy RNN*

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

* là đầu vào tại bước .
* là trạng thái ẩn tại bước . Nó chính là bộ nhớ của mạng. được tính toán dựa trên cả các trạng thái ẩn phía trước và đầu vào tại bước đó:

. Hàm thường là hàm phi tuyến tính như hàm tanh hay ReLu. Để làm phép toán cho phần tử ẩn đầu tiên ta cần khởi tạo thêm , thường giá trị khởi tạo được gắn bằng 0 hoặc giá trị ngẫu nhiên. Có thể hiểu là ban đầu chưa có dữ liệu gì để học thì memory rỗng.

* là đầu ra tại bước . Ví dụ, ta muốn dự đoán từ tiếp theo có thể xuất hiện trong câu thì chính là một vec-tơ xác xuất các từ trong danh sách từ vựng của ta:

RNN chỉ sử dụng một mạng Neural duy nhất (thường là 1 layer) để tính giá trị output của từng timestep. Do đó các outputs khi trở thành input sẽ được nhân với cùng một weights matrix (Hình 2.1.2). Đây cũng chính là lý do tại sao có từ Recurrent trong tên của RNN. Recurrent có nghĩa là mô hình sẽ thực hiện các phép tính toán giống hệt nhau cho từng phần tử của chuỗi dữ liệu đầu vào và kết quả output sẽ phụ thuộc vào kết quả của các tính toán ở phần trước. Về mặt lý thuyết thì RNN có thể xử lý và lưu trữ thông tin của một chuỗi dữ liệu với độ dài bất kỳ. Tuy nhiên trong thực tế thì RNN chỉ tỏ ra hiệu quả với chuỗi dữ liệu có độ dài không quá lớn (short-term memory hay còn gọi là long-term dependency problem).

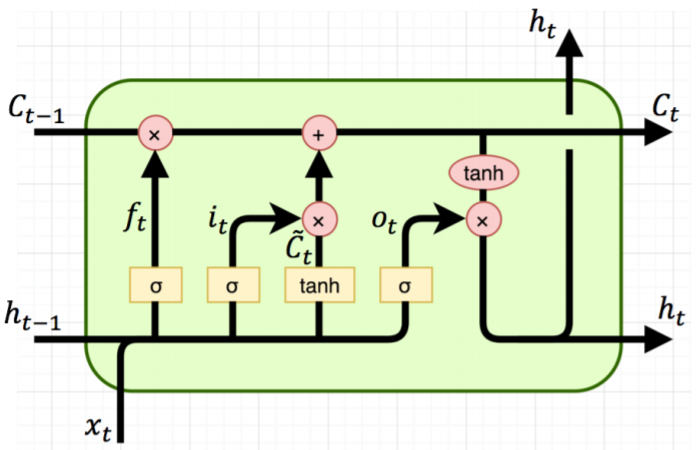
Nguyên nhân của vấn đề này là do vanishing gradient problem (gradient được sử dụng để cập nhật giá trị của weight matrix trong RNN và nó có giá trị nhỏ dần theo từng layer khi thực hiện back propagation). Khi gradient trở nên rất nhỏ (có giá trị gần bằng 0) thì giá trị của weight matrix sẽ không được cập nhật thêm và do đó mạng Neuron sẽ dừng việc learning tại layer này. Đây cũng chính là lý do khiến cho RNN không thể lưu trữ thông tin của các timesteps đầu tiên trong một chuỗi dữ liệu có độ dài lớn. Vấn đề này đặt ra trong nhiều năm và mạng LSTM (Hochreiter & Schmidhuber - 1997) và mới đây là mạng GRU (Cho - 2014) được đề xuất để giải quyết vấn đề này.

2.1.3. Mạng LSTM

Mạng bộ nhớ dài-ngắn (Long Short Term Memory networks), thường được gọi là LSTM - là một mạng cải tiến của RNN nhằm giải quyết vấn đề nhớ các bước dài của RNN, nó có khả năng học được các phụ thuộc xa. LSTM được giới thiệu bởi Hochreiter & Schmidhuber (1997), và sau đó đã được cải tiến và phổ biến bởi rất nhiều người trong ngành. Chúng hoạt động cực kì hiệu quả trên nhiều bài toán khác nhau nên dần đã trở nên phổ biến như hiện nay.

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.

Đầu vào gồm ba thành phần. là đầu vào tại bước hiện tại. là đầu ra từ một khối LSTM trước và là “nhớ” của khối trước, và đây cũng chính là điểm quan trọng nhất của LSTM. Đầu ra của nó gồm là kết quả của khối LSTM hiện tại và giống như một băng chuyền ở trên mô hình RNN, thông tin nào cần quan trọng và dùng ở sau sẽ được gửi vào và dùng khi cần. Như vậy, một khối đơn LSTM đưa ra quyết định dựa vào việc xem xét đầu vào hiện tại, kết quả và ô nhớ của khối trước và nó sinh ra một đầu ra mới cũng như là ô nhớ của nó.



*Hình 2.1.3. Mô hình LSTM*

Một mạng LSTM tương đương với mạng RNN trừ việc các đơn vị tổng hợp trong tầng ẩn được thay thế bằng các khối nhớ. Các khối LSTM cũng có thể được hòa trộn với các đơn vị tổng hợp mặc dù về cơ bản là không cần thiết. Tầng đầu ra có thể được sử dụng cho các mạng LSTM như cho mạng RNN chuẩn.

Các cổng nhân lên cho phép các ô nhớ LSTM được lưu trữ và truy cập thông tin trên một thời gian dài, vì thế giảm nhẹ vấn đề biến mất đạo hàm. Ví dụ ngay khi cổng đầu vào được đóng lại (có hàm kích hoạt gần 0), sự kích hoạt của ô sẽ không bị ghi đè bởi đầu vào đang đến trong mạng, do đó có thể cung cấp cho mạng sau này bằng cách mở cổng đầu ra. LSTM khá thành công trong một loạt các nhiệm vụ yêu cầu bộ nhớ phạm vi dài, và nó còn được áp dụng trong các vấn đề trong thế giới thực như là cấu trúc thứ cấp proteion, sinh âm nhạc, nhận dạng âm thanh, nhận dạng chữ viết.

**2.2. Mô hình đề xuất**

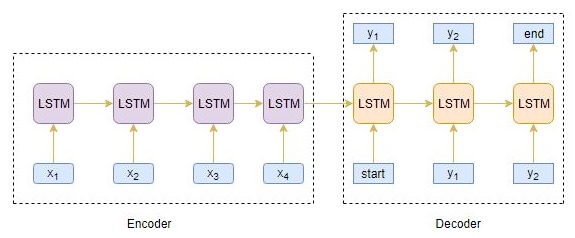
Trong những năm gần đây, cùng với sự phát triển mạnh mẽ của các mô hình huấn luyện end-to-end đã tạo ra hướng đi mới để giải quyết bài toán tóm tắt văn bản tự động. Tóm tắt văn bản tự động theo hướng tóm lược với mô hình Sequence-to-Sequence. Dựa vào ý tưởng sinh ra câu mới từ câu đã có của mô hình Sequence-to-Sequence, từ đó xây dựng mô hình rút trích những thông tin quan trọng của cả đoạn văn, thu gọn lại thành một câu tóm tắt mang đầy đủ thông tin của đoạn văn; bên cạnh đó, phối hợp với các phương pháp xử lý ngôn ngữ tự nhiên như tách từ, loại bỏ dấu câu… để tinh chỉnh và xử lý văn bản.

2.2.1. Mô hình Sequence to Sequence

Bài toán RNN được phân làm một số dạng như: one to one, one to many, many to one và many to many. Mô hình sequence to sequence (seq2seq) sinh ra để giải quyết bài toán many to many và rất thành công trong các bài toán như Speech recognition, Text summarization, Image captioning, …

Sequence to Sequence Model (Seq2seq) là một mô hình Deep Learning với mục đích tạo ra một output sequence từ một input sequence mà độ dài của 2 sequences này có thể khác nhau. Seq2seq được giới thiệu bởi nhóm nghiên cứu của Google vào năm 2014 trong bài báo Sequence to Sequence with Neural Networks.

Mô hình gồm 2 phần chính là Encoder và Decoder, cả hai thành phần này đều được hình thành từ các mạng Neural Networks. Tại bước encoder, đầu vào của mạng LSTM là văn bản gốc và output ra context vector, các vector này được tạo ra bằng cách mã hoá chuỗi từ với mô hình word embedding. Sau đó, khi decoder, cũng sử dụng một LSTM để sinh ra chuỗi từ mới dựa vào chuỗi đầu vào được tạo ra ở phần Encoder và các từ được sinh ra tại câu tóm tắt.

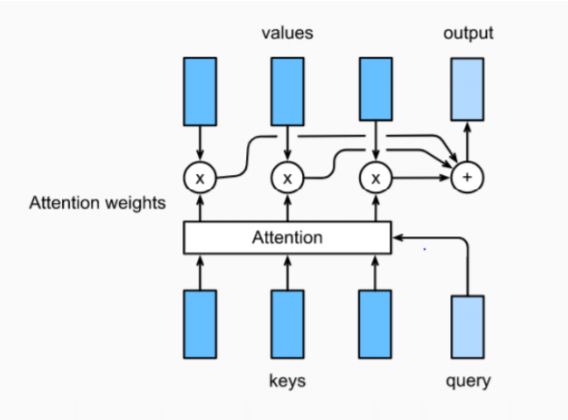


*Hình 2.2.1: Mô hình Sequence to Sequence*

2.2.2. Cơ chế Attention

Mô hình seq2seq encode cả văn bản gốc thành 1 context vector, rồi dùng context vector để sinh ra các từ trong câu tóm tắt tương ứng. Như vậy khi câu dài thì rất khó cho decoder chỉ dùng 1 context vector có thể sinh ra được câu output chuẩn. Thêm vào đó các mô hình RNN đều bị mất ít nhiều thông tin ở các node ở xa nên bản thân context vector cũng khó để học được thông tin ở các từ ở phần đầu của encoder. Cơ chế Attention cho phép bộ decoder tập trung vào một phần khác nhau từ đầu ra của encoder, không chỉ dùng context vector mà còn sử dụng hidden state ở từng từ trong input với trọng số ảnh hưởng tương ứng, nên việc dự đoán từ tiếp theo sẽ tốt hơn cũng như không sợ tình trạng từ ở xa bị mất thông tin ở context vector.

Nguyên tắc hoạt động chung của Attention là tại mỗi Decoding Step, Decoder sẽ chỉ tập chung vào phần liên quan trong input sequence thay vì toàn bộ input sequence. Như vậy, tại mỗi Decoding step, Decoder nhận 3 đầu vào là: Hidden state của decoding step trước, output của step trước và attention vector. Attention vector chứa attention weight của từng từ trong input sequence. Từ nào chứa nhiều thông tin cần thiết cho việc decoding thì sẽ có giá trị weight lớn hơn và tổng các weights của tất cả các từ trong input sequence phải bằng 1. Giá trị Attention weights này được học thông qua quá trình huấn luyện với việc sử dụng input sequence và hidden state của decoding step trước. Mỗi Decoding step có một giá trị Attention vector riêng, do đó với một input sequence có chiều dài ‘n’ và output sequence có chiều dài ‘m’, ta phải thực hiện việc tính toán ‘n \* m’ Attention weights.

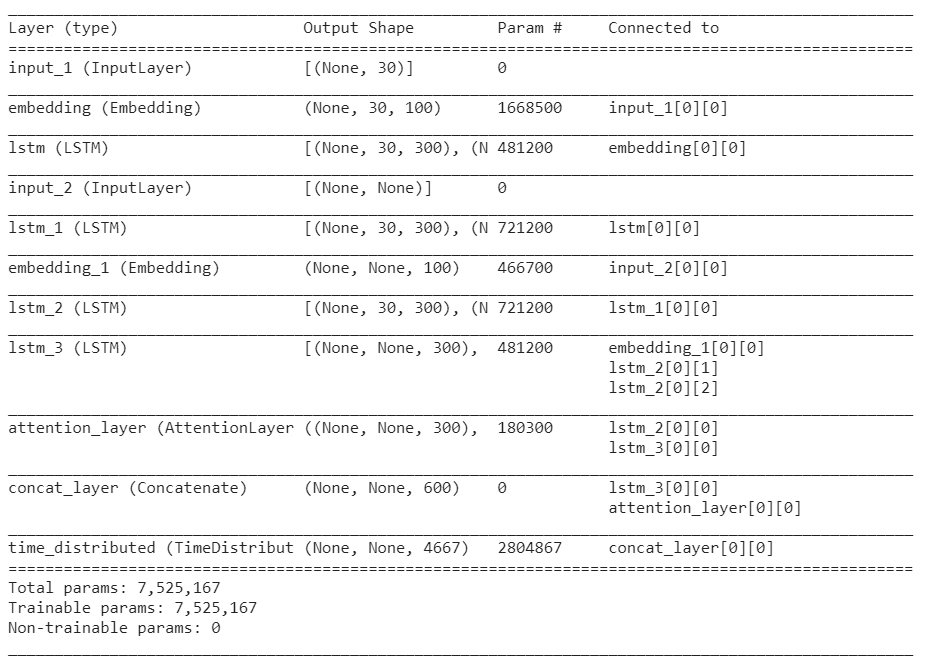
**

*Hình 2.2.2: Các thành phần trong Attention*

Ngoài ra các mô hình deep learning hay bị nói là hộp đen (black box) vì mô hình không giải thích được, attention phần nào giúp visualize được kết quả dự đoán, ví dụ từ nào ở ouptut ảnh hưởng nhiều bởi từ nào trong input. Do đó model học được quan hệ giữa các từ trong input và output để đưa ra kết quả dự đoán.

2.2.3. Xây dựng mô hình

Mục đích của việc tóm tắt văn bản là chuyển từ một văn bản gốc thành một văn bản có độ dài ngắn hơn và mang đầy đủ ý nghĩa. Theo mục đích đó, chúng ta xây dựng một RNN để rút trích đặc trưng từ chuỗi đầu vào và sinh ra chuỗi từ mới phụ thuộc vào chuỗi từ trước đó bằng một RNN thứ hai. Hai RNN này được điều chỉnh trọng số phù hợp để chuỗi đầu ra ngắn hơn và phụ thuộc xác suất chuỗi đầu vào. Quá trình tóm tắt văn bản tự động được huấn luyện end-to-end với mô hình “Sequence-tosequence with Attention” kết hợp với các khối LSTM cho cả bộ Encoder và bộ Decoder.

****

*Hình 2.2.3: Mô hình tóm tắt nhận xét của người dùng*

Các token của văn bản đầu vào được lần lượt đưa vào Encoder, sinh ra một chuỗi các trạng thái ẩn của bộ Encoder. Word embedding được khởi tạo ngẫu nhiên theo phân phối chuẩn và được học để điều chỉnh các hệ số trong quá trình huấn luyện. Decoder nhận các word embedding của các từ ở thời điểm trước: trong quá trình huấn luyện chính là các từ của văn bản tóm tắt tham chiếu và trong quá trình chạy thì các từ ở thời điểm trước chính là các từ được sinh bởi Decoder. Để Decoder có thể học cách tự sinh các từ cho văn bản tóm tắt, mô hình sử dụng cơ chế Attention, dựa trên phân phối xác suất của các từ trong văn bản gốc, giúp Decoder xác định được vị trí của từ sẽ được lựa chọn cho văn bản tóm tắt và tổng hợp lại thành câu tóm tắt hoàn chỉnh.

* 1. **Kết quả**

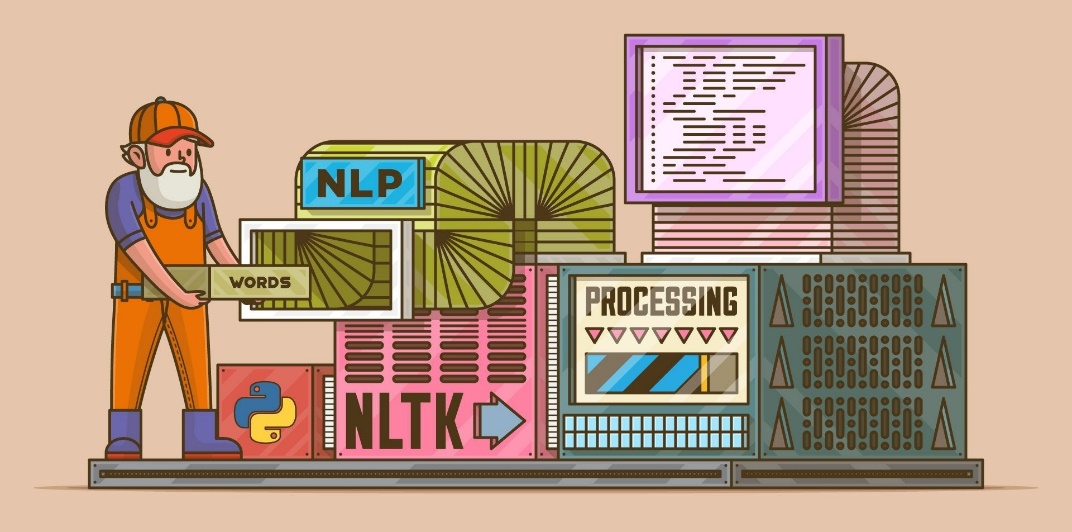
2.3.1. Quá trình huấn luyện

Bộ dữ liệu để huấn luyện mô hình gồm khoảng ~ 500.000 bài đánh giá, nhận xét của người dùng về các loại thực phẩm, đồ ăn từ Amazon Fine Food Reviews. Tệp dữ liệu bao gồm thông tin về sản phẩm và người dùng, cùng với xếp hạng đánh giá và nhận xét.



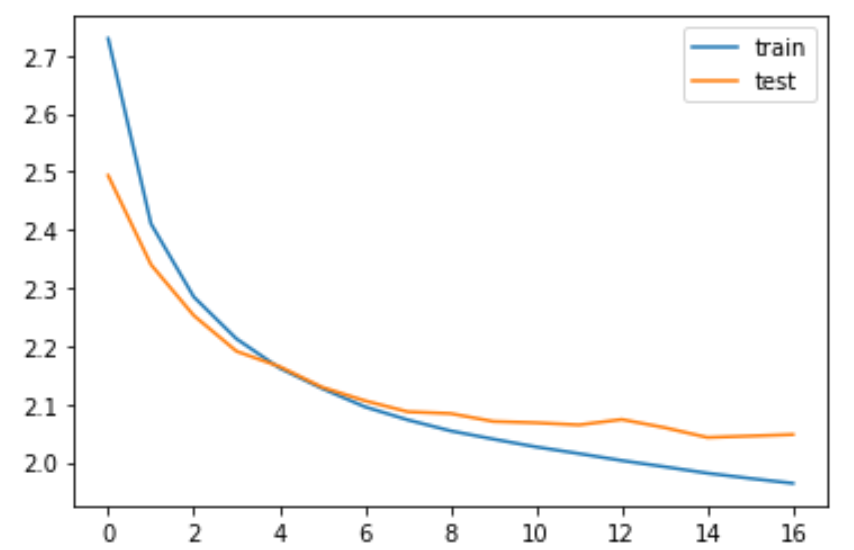
*Hình 2.3.1.1: Dữ liệu Amazon Fine Food*

Một model tốt ở các bài toán thực tế luôn luôn có hình bóng của một tập training data chất lượng. Chính vì vậy, việc thực hiện các bước tiền xử lý cơ bản là rất quan trọng trước khi xây dựng mô hình. Đầu tiên, loại bỏ tất cả các ký hiệu, ký tự đặc biệt, khoảng trắng thừa ra khỏi văn bản mà không ảnh hưởng đến nội dung của văn bản gốc, và chuyển các ký tự viết hoa thành viết thường, chuyển đổi các cụm từ viết tắt. Tiếp theo sử dụng thư viện NLTK (NLTK là một thư viện Python được sử dụng phổ biến trong NLP.) Dùng để xoá các từ trong “StopWords” (StopWords là những từ xuất hiện nhiều trong ngôn ngữ tự nhiên, tuy nhiên lại không mang nhiều ý nghĩa. Ở tiếng Việt StopWords là những từ như: để, này, kia..., còn trong tiếng Anh là những từ như: is, the, a, an…). Sau khi làm sạch văn bản, sẽ gắn các tag “START” và “END” ở đầu và cuối của bản tóm tắt trong tập dữ liệu.



*Hình 2.3.1.2: Thư viện NLTK sử dụng trong lĩnh vực NLP*

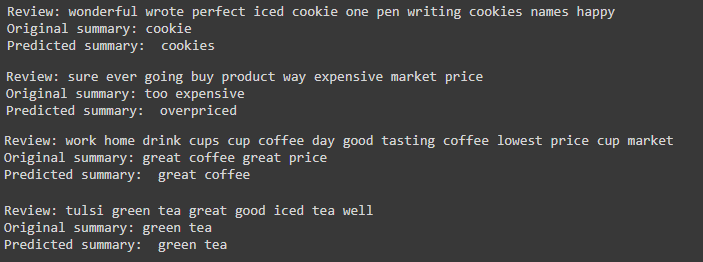
Trước khi tiến hành huấn luyện mô hình cần chia tập dữ liệu, sử dụng 80% tập dữ liệu làm dữ liệu đào tạo, 10% cho đánh giá và 10% cho tập test. Sau đó, huấn luyện mô hình sử dụng "Seq2seq with Attention". Dưới đây là biểu diễn của hàm lỗi khi huấn luyện mô hình Sequence-to-Sequence with Attention:



*Hình 2.3.1.3: Mô tả quá trình cập nhật loss trong quá trình huấn luyện*

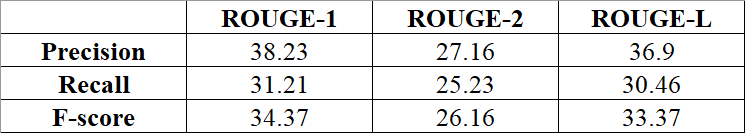
* + 1. **Phân tích kết quả**

Nhìn chung, mô hình đã cho ra kết quả khá sát với bản tóm tắt gốc.



*Hình 2.3.2.1: Phân tích kết quả của mô hình*

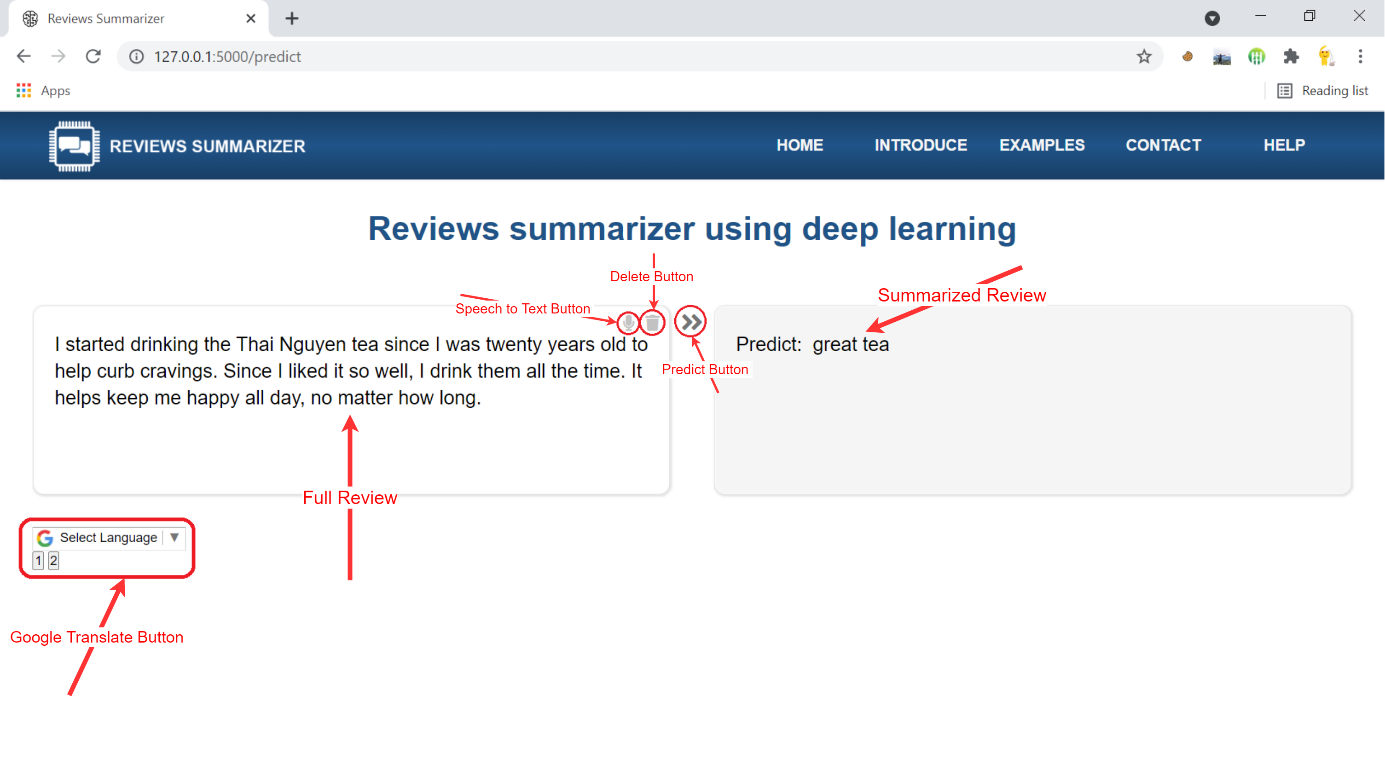
Để tính toán độ chính xác của việc tóm tắt văn bản, người ta thường sử dụng độ đo ROUGE. ROUGE là viết tắt của Recall-Oriented Understudy. Nó thực chất là một tập hợp các số liệu để đánh giá tự động cho việc tóm tắt các văn bản cũng như dịch máy. Nó hoạt động bằng cách so sánh một bản tóm tắt hoặc bản dịch được tạo tự động so với một tập hợp các bản tóm tắt tham chiếu.



*Hình 2.3.2.2: Đánh giá độ chính xác của mô hình*

* + 1. **Triển khai model lên hệ thống sử dụng Flask API**

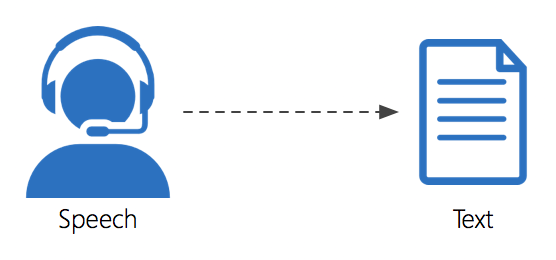
Flask là một framework web của Python, còn được coi là một microframework bởi nó không yêu cầu các công cụ hoặc thư viện cụ thể, được phân loại là Web Framework siêu nhỏ, nhẹ được phát triển bởi Armin Ronacher. Flask cũng có ưu điểm đó là tính bảo mật rất cao, code đẹp và dễ hiểu.

*****Hình 2.3.3: Triển khai mô hình lên hệ thống*

* + 1. **Tích hợp tính năng dịch văn bản, nhận diện giọng nói**

Một ứng dụng rất phổ biến trong lĩnh vực xử lý ngôn ngữ tự nhiên là nhận diện giọng nói chuyển thành văn bản. Ngày nay, có rất nhiều API resources có sẵn trên thị trường, giúp người dùng dễ dàng lựa chọn cái này hay cái khác như Google, Amazon, IBM, …

Web Speech API được định nghĩa là một JavaScript API cho phép chúng kết hợp nhận dạng và tổng hợp giọng nói vào trang web. Nó cho phép chúng ta sử dụng script để chuyển text-to-speech và sử dụng nhận dạng giọng nói làm đầu vào. JavaScript API cho phép trang web của chúng ta kiểm soát được việc kích hoạt, thời gian và xử lý kết quả. Sau khi cấp quyền cho microphone recognition service bắt đầu lắng nghe và chúng ta nói điều gì đó, tùy theo cài đặt thuộc tính của chúng ta việc ngừng nói để gọi đến phương thức stop() hay nút bấm để thông báo với recognition service sẽ dừng việc lắng nghe. Đối với Chrome, trình duyệt sẽ lấy âm thanh và gửi đến máy chủ của Google để thực hiện quá trình phiên âm.



*Hình 2.3.4: Tính năng nhận diện giọng nói*

1. **Kết luận và kiến nghị**

**3.1. Kết luận**

Bản báo cáo đã tiến hành nghiên cứu giải quyết bài toán tóm tắt tự động những đánh giá, nhận xét của người dùng dựa vào phương pháp tóm tắt tóm lược (abstractive summarization). Bài toán này được xác định là một bài toán có độ phức tạp cao và là nền tảng của nhiều ứng dụng thực tế. Phương pháp giải quyết của báo cáo tập trung vào xây dựng mô hình học sâu dựa trên các nghiên cứu về các mô hình mạng LSTM, các mô hình chuỗi sang chuỗi (sequence-to-sequence), các kỹ thuật vec tơ hóa từ và văn bản, bản báo cáo đã xây dựng một kiến trúc mô hình học sâu sử dụng LSTM cho bài toán tóm tắt tự động với các tham số được tối ưu hóa cho việc huấn luyện và thử nghiệm trên máy tính cá nhân.

Bên cạnh đó, báo cáo nghiên cứu cũng đã thử nghiệm mô hình đã xây dựng với dữ liệu tiếng Anh, bộ dữ liệu gồm dữ liệu thô và dữ liệu đã được xử lý về dạng nhị phân và đánh giá mô hình bằng phương pháp ROUGE. Hơn nữa, báo cáo cũng đã triển khai mô hình lên nền tảng web và áp dụng thêm các API của Google giúp hệ thống ứng dụng vào nhiều lĩnh vực. Quá trình thực nghiệm với dữ liệu về các đánh giá, nhận xét trên các trang thương mại điện tử đã đạt được kết quả khả quan, hệ thống đã có thể tạo một bản tóm tắt dễ đọc dựa trên ngữ cảnh có trong văn bản. Từ đó, cho thấy tính đúng đắn của việc lựa chọn cũng như kết hợp các phương pháp, đồng thời hứa hẹn nhiều tiềm năng phát triển hoàn thiện.

**3.2. Kiến nghị**

Việc ứng dụng Deep Learning để tóm tắt văn bản tự động vẫn chưa đạt được mức độ giống hoàn toàn với con người. Để tăng độ chính xác cho mô hình, cần tăng kích thước tập dữ liệu đào tạo vì khả năng tổng quát hóa của mô hình học sâu được cải thiện khi kích thước tập dữ liệu lớn hơn. Đánh giá hiệu suất của mô hình dựa trên hệ thống độ đo BLEU (Bilingual Evaluation Understudy), và xây dựng hệ thống tóm tắt văn bản bằng tiếng Việt. Qua nghiên cứu khoa học này, em hi vọng có thể đưa ra một gợi ý mang tính định hướng cho việc tóm tắt văn bản tự động, từ đó sẽ giúp nâng cao chất lượng cuộc sống.

Em xin chân thành cảm ơn thầy Nguyễn Hiếu Cường đã hướng dẫn em trong quá trình nghiên cứu đề tài này và mong sẽ nhận được những sự góp ý của thầy cô để có thể hoàn thiện hơn nghiên cứu khoa học của nhóm.

**Tài liệu tham khảo**

[1]. Aravind Pai, Text Summarization using Deep Learning in Python.

[2]. Amrzaki, Text Summurization Abstractive Methods.

[3]. Ayako Mikami, Long Short-Term Memory - Recurrent Neural Network Architectures for Generating Music and Japanese Lyrics.

[4]. Sachin Joglekar, Understanding the new Google Translate.

[5]. Lâm Quang Tường, Phạm Thế Phi, Đỗ Đức Hào. “Tóm tắt văn bản tiếng việt tự động với mô hình Sequence-to-Sequence”.

[6]. Đoàn Xuân Dũng, “Tóm tắt văn bản sử dụng các kỹ thuật trong Deep Learning”.

[7]. Nguyễn Thanh Tuấn, “Deep Learning cơ bản”.