TRƯỜNG ĐẠI HỌC SƯ PHẠM THÀNH PHỐ HỒ CHÍ MINH

Khoa Công Nghệ Thông Tin

KHAI THÁC VĂN BẢN

(TEXT MINING)

Đề tài:

TÓM TẮT VĂN BẢN (TEXT SUMMARIZATION)

GVHD: Nguyễn Hồng Bửu Long
SVTH: Nguyễn Văn Thịnh
Hồ Khả Việt Huấn
Lâm Phước Đạt

Thành phố Hồ Chí Minh 2020

Lời cảm ơn

Em xin chân thành cảm ơn Bộ môn Công nghệ thông tin – Trường đại học Sư Phạm thành phố Hồ Chí Minh đã tạo điều kiện và đưa môn khai thác văn bản vào giảng dạy.

Em xin chân thành cảm ơn thầy Nguyễn Hồng Bửu Long đã tận tình hướng dẫn, chỉ bảo chúng em trong suốt quá trình học tập và thực hiện đề tài.

Em cũng xin chân thành cảm ơn quý Thầy Cô trong các bộ môn đã tận tình giảng dạy trang bị cho em những kiến thức cần thiết trong suốt quá trình học tập tại trường.

Môn học khai thác văn bản là môn học vô cùng thú vị, vô cùng bổ ích và có tính thực tế cao. Mặc dù đã rất có gắng hoàn thành đồ án với tất cả nỗ lực của tất cả cả các thành viên trong nhớm, nhưng do kiến thức bản thân có hạn nên chắc chắn đồ án không tránh khỏi những sai sót và hạn chế, kính mong sự thông cảm, chỉ bảo của quí Thầy Cô và các bạn.

Sinh viên

Nguyễn Văn Thịnh

Hồ Khả Việt Huấn

Lâm Phước Đạt

MỤC LỤC

I. Giới thiệu về khai thác văn bản (Text Mining)

1) Khái niệm.

Với sự phát triển nhanh chóng của công nghệ và sự phát triển của các mạng xã hội. Các doanh nghiệp hiện nay đang đối mặt với "cơn lũ" dữ liệu về mọi mặt: phản hồi của khách hàng, thông tin đối thủ cạnh tranh, emails của khách hàng, thông tin họp báo, hồ sơ pháp lý, các văn bản về sản phẩm và kĩ thuật. Việc khai thác được những dữ liệu này là điểm mấu chốt để các doanh nghiệp có thể triển khai nhanh chóng các quyết định của mình so với đối thủ cạnh tranh.

Vấn đề ở đây là gì? Có quá nhiều thông tin để xử lý cùng lúc (hơn 85% dữ liệu trên thế giới không có cấu trúc), và kích thước dữ liệu ngày càng tăng. Đối với nhiều doanh nghiệp, điều này là bất khả thi để điều động nhân sự đọc tất cả mọi thứ được cho là quan trọng (các khách hàng đang nói gì về sản phẩm, những đối thủ cạnh tranh của chúng ta đang làm gì). Vì thế khai thác văn bản (*Text mining*) ra đời để giải quyết các vấn đề trên.

Khai thác văn bản, tương tự như phân tích văn bản (Text Analytics), là quá trình lấy thông tin chất lượng cao từ văn bản . Nó liên quan đến việc máy tính phát hiện ra thông tin mới, chưa biết trước đây, bằng cách tự động trích xuất thông tin từ các nguồn tài liệu viết khác nhau. Tài nguyên viết có thể bao gồm trang web , sách , email , bài đánh giá và bài báo. Thông tin chất lượng cao thường thu được bằng cách đưa ra các mẫu và xu hướng bằng các phương tiện như học mẫu thống kê.

Khai thác văn bản bao gồm việc truy xuất thông tin , phân tích từ vựng để nghiên cứu sự phân bố tần số từ, nhận dạng mẫu , gắn thẻ / chú thích , trích xuất thông tin , các kỹ thuật khai thác dữ liệu bao gồm phân tích mối quan hệ và sự liên kết, trực quan hóa và phân tích dự đoán . Về cơ bản, mục tiêu bao trùm là biến văn bản thành dữ liệu để phân tích, thông qua ứng dụng xử lý ngôn ngữ tự nhiên (NLP), các loại thuật toán và phương pháp phân tích khác nhau. Một giai đoạn quan trọng của quá trình này là giải thích thông tin thu thập được.

NLP là một nhánh của Trí tuệ nhân tạo, tập trung vào việc nghiên cứu sự tương tác giữa máy tính và ngôn ngữ tự nhiên của con người. Mục tiêu của lĩnh vực này là giúp máy tính hiểu và thực hiện hiệu quả những nhiệm vụ liên quan đến ngôn ngữ của con người như: tương tác giữa người và máy, cải thiện hiệu quả giao tiếp giữa con người với con người, hoặc đơn giản là nâng cao hiệu quả xử lý văn bản và lời nói.

2) Các ứng dụng của khai thác văn bản

a) Sentiment Analysis

Sentiment Analysis (Phân tích quan điểm) là quá trình phân tích, đánh giá quan điểm của một người về một đối tượng nào đó (quan điểm mang tính tiêu cực, tích cực hay bình thường,..). Quá trình này có thể được thực hiện bằng việc sử dụng các tập luật (rule-based), sử dụng Machine Learning (đặc biệt là Deep Learning) hoặc phương pháp Hybrid (kết hợp hai phương pháp trên).

Một trong các cách đơn giản nhất ta có thể dùng là tính điểm tiêu cực/tích cực của từng từ trong câu rồi cộng tổng điểm của các từ để lấy kết quả cuối cùng. Tuy nhiên cách này thường đưa ra độ chính xác thấp và không hiệu quả. Các phương pháp hiện đại thường không chỉ đánh giá tính

tiêu cực/tích cực của từng từ mà còn xét cả mối quan hệ của các từ trong câu cũng như toàn bộ cấu trúc câu. Bằng việc sử dụng Deep Learning và xây dựng Sementic Tree, tính hiệu quả và độ chính xác trong việc đánh giá quan điểm đã được cải thiện rất nhiều so với các phương pháp truyền thống.

Sentiment Analysis được ứng dụng nhiều trong các sản phẩm thực tế, đặc biệt là trong hoạt động quảng bá kinh doanh. Việc phân tích các đánh giá của người dùng về một sản phẩm xem họ đánh giá tiêu cực, tích cực hoặc đánh giá các mặt hạn chế của sản phẩm sẽ giúp công ty nâng cao chất lượng sản phẩm và tăng cường hình ảnh của công ty. Một ví dụ khác có thể kể đến là việc phân tích quan điểm của người dân về một chính sách, quy định hay dự luật mà nhà nước chuẩn bị ban hành có thể giúp các nhà hoạch định chính sách biết được chính sách nào sẽ mang lại hiệu quả cao và được người dân ủng hộ.

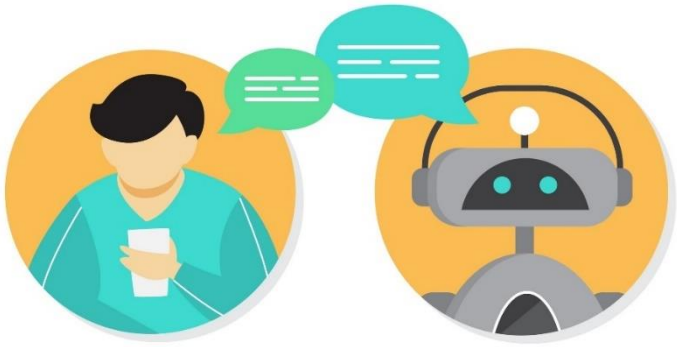
b) Question Answering (QA)

QA là một hệ thống được xây dựng để có thể trả lời các câu hỏi mà con người đặt ra trong một lĩnh vực nhất định. Một QA đơn giản có thể dùng để thay con người trả lời các câu hỏi lặp đi lặp lại của người dùng như: “Sự kiện X diễn ra khi nào?”, “Vinaphone MAX70 là gì?”, “iPhone X giá bao nhiêu?”, …

Việc trả lời câu hỏi phụ thuộc rất nhiều vào một kho dữ liệu tìm kiếm tốt - vì không có tài liệu chứa câu trả lời, hệ thống trả lời câu hỏi có thể làm được rất ít. Do đó, tập dữ liệu có kích thước càng lớn thì khả năng trả lời tốt hơn.

Hệ thống lấy một câu hỏi ngôn ngữ tự nhiên làm đầu vào thay vì một tập hợp các từ khóa, ví dụ: "Bác Hồ sinh năm bao nhiêu?" Câu sau đó được chuyển thành một truy vấn thông qua hình thức logic của nó. Việc có đầu vào dưới dạng câu hỏi ngôn ngữ tự nhiên làm cho hệ thống thân

thiện với người dùng hơn, nhưng khó triển khai hơn, vì có nhiều loại câu hỏi khác nhau và hệ thống sẽ phải xác định câu hỏi chính xác để đưa ra câu trả lời hợp lý.



Hệ thống trả lời cầu hỏi sẽ giúp tiết kiệm thời gian, công sức cũng như chi phí để trả lời các câu hỏi của khách hàng. Nó ứng dụng trong nhiều lĩnh vực như bán hàng trực tuyến, nó có thể trả lời số lượng khách hàng không giới hạn và gần như là trả lời tức thì khi khách hàng có câu hỏi, và đương nhiên cũng có thể gia tăng doanh số đáng kể cho công ty. Hệ thống giải đáp các thắc mắc tuyển sinh, gần như các học sinh khi chuẩn bị chọn ngành, chọn trường cho mình đều có những câu hỏi giống nhau, nên hệ thống có thể trả lời thay thê cho con người, giúp tiết kiệm thời gian cho các học sinh cũng như là các nhân viên tư vấn…

c) Các ứng dụng khác

Và xử lý ngôn ngữ tự nhiên cũng còn nhiều ứng dụng khác như:

- **Truy xuất thông tin** (Information Retrieval – IR) có nhiệm vụ tìm các tài liệu dưới dạng không có cấu trúc (thường là văn bản) đáp ứng nhu cầu về thông tin từ những nguồn tổng hợp lớn.
- **Trích chọn thông tin** (Information Extraction) nhận diện một số loại thực thể được xác định trước, mối quan hệ giữa các thực thể và các sự kiện trong văn bản ngôn ngữ tự nhiên.
- **Tóm tắt văn bản tự động** là bài toán thu gọn văn bản đầu vào để cho ra một bản tóm tắt ngắn gọn với những nội dung quan trọng nhất của văn bản gốc.

II. Tóm tắt văn bản (Text Summarization)

1) Giới thiệu đề tài

a) Lý do chọn đề tài

Trong kỷ nguyên dữ liệu lớn, đã có sự bùng nổ về lượng dữ liệu văn bản từ nhiều nguồn khác nhau. Khối lượng văn bản này là một nguồn thông tin và kiến thức vô giá cần được tóm tắt một cách hiệu quả để trở nên hữu ích. Tính sẵn có ngày càng tăng của tài liệu đã đòi hỏi phải có nghiên cứu toàn diện trong lĩnh vực NLP để tóm tắt văn bản tự động. Tóm tắt văn bản tự động là nhiệm

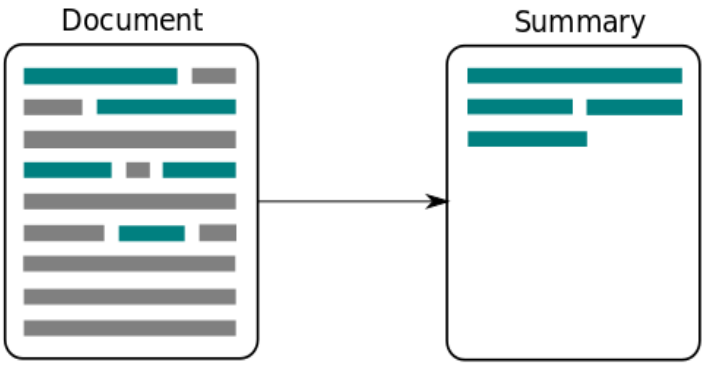
vụ tạo ra một bản tóm tắt ngắn gọn và trôi chảy mà không cần bất kỳ sự trợ giúp nào của con người trong khi vẫn giữ được ý nghĩa của tài liệu văn bản gốc.

Thay vì đọc tất cả các bài viết, đọc một bản tóm tắt sẽ cung cấp cho bạn một cái nhìn tổng quan về câu chuyện và tiết kiệm thời gian của bạn. Mặt khác, nó có thể sử dụng để chỉ định các chủ đề quan trọng nhất của tài liệu. Ví dụ, khi tóm tắt blog, có các cuộc thảo luận hoặc nhận xét sau khi bài đăng blog là nguồn thông tin tốt để xác định phần nào của blog là quan trọng và thú vị.

Có nhiều lý do giải thích tại sao việc Tóm tắt văn bản tự động giúp chúng ta rất nhiều như là Tóm tắt giảm thiểu thời gian đọc. Khi đọc hay nghiên cứu tài liệu, tóm tắt làm cho quá trình lựa chọn tài liệu trở nên dễ dàng hơn, nhanh hơn và chính xác hơn. Việc tự động tóm tắt văn bản cải thiện hiệu quả của việc lập chỉ mục. Thuật toán tóm tắt tự động ít sai lệch hơn so với việc tóm tắt của con người. Tóm tắt thông tin cá nhân rất hữu ích trong các hệ thống trả lời câu hỏi tự động vì chúng cung cấp thông tin cá nhân. Sử dụng các hệ thống tóm tắt tự động hoặc bán tự động cho phép các dịch vụ thương mại tăng số lượng tài liệu văn bản mà chúng có thể xử lý.

b) Khái niệm.

Tóm tắt là nhiệm vụ cô đọng một phần văn bản thành một phiên bản ngắn hơn, giảm kích thước của văn bản ban đầu trong khi đồng thời bảo tồn các yếu tố thông tin chính và ý nghĩa của nội dung.



Nó rất khó khăn, bởi vì khi chúng ta là con người tóm tắt một đoạn văn bản, chúng ta thường đọc nó hoàn toàn để phát triển sự hiểu biết của mình, và sau đó viết một bản tóm tắt nêu bật những điểm chính của nó. Vì máy tính thiếu kiến thức của con người và khả năng ngôn ngữ, nó làm cho việc tóm tắt văn bản tự động trở thành một nhiệm vụ rất khó khăn và không hề nhỏ. Nhiều mô hình khác nhau dựa trên máy học đã được đề xuất cho nhiệm vụ này.

Có hai cách tiếp cận khác nhau để tóm tắt tự động: **Trích xuất văn bản (Extractive)** và **Tóm tắt văn bản (Abstractive).**

c) Các ứng dụng của hệ thống tóm tắt văn bản tự động

+ **Tóm tắt tin tức**: Một hệ thống có thể thu thập hàng trăm hàng nghàn bài báo từ nhiều nguồn uy tín khác nhau. Sau đó phân cụm, phân loại và tóm tắt chúng để người dùng dễ dàng tìm ra tin tức mà họ quan tâm nhất, muốn đọc nhất.

+ **Tóm tắt sách**: Theo zingnews.vn có tới 45 nghìn tỷ trang sách được in ra mỗi ngày, với số lượng sách khổng lồ như thế làm cho việc chọn lựa những quyển sách về chủ đề mình mong muốn trở nên khó khăn hơn rất nhiều, vì vậy một hệ thống tóm tắt sách có thể thu thập, tóm tắt rồi đề xuất những cuốn sách về chủ đề người đọc quan tâm, và đọc những tóm tắt của hệ thống giúp cho sự lựa chọn của người đọc trở nên chính xác hơn.

+ **Tóm tắt các văn bản hành chính:** Mặc dù các quốc gia đang chuyển đổi từ tài liệu pháp lý bằng giấy sang tài liệu pháp lý trực tuyến, nhưng số tài liệu pháp lý vẫn còn chiếm chủ yếu và số lượng tài liệu này cực kỳ lớn, gây khó khăn cho việc tìm kiếm, đọc hiểu,… Vì thế một hệ thống tóm tắt văn bản pháp lý tự động sẽ giúp cho các công viên chức, các công ty tìm kiếm văn bản một cách dễ dàng trong một kho dữ tài liệu vô cùng lơn qua nhiều năm.

III. Phương pháp tiếp cận: trích xuất văn bản (Extractive)

1) Giới thiệu:

Tóm tắt trích xuất là chọn các câu trực tiếp từ tài liệu dựa trên chức năng cho điểm để tạo thành một tóm tắt mạch lạc.

Phương pháp này hoạt động bằng cách xác định các phần quan trọng của văn bản cắt ra và ghép các phần nội dung lại với nhau để tạo ra một phiên bản cô đọng. Vì vậy, chúng chỉ phụ thuộc vào việc trích xuất các câu từ văn bản gốc. Hầu hết các nghiên cứu tóm tắt ngày nay đều tập trung vào tóm tắt chiết tách, một khi nó dễ dàng hơn và tạo ra các tóm tắt ngữ pháp một cách tự nhiên yêu cầu tương đối ít phân tích ngôn ngữ. Hơn nữa, tóm tắt chiết xuất chứa các câu quan trọng nhất của đầu vào, có thể là một tài liệu hoặc nhiều tài liệu.

Cách tiếp cận trích xuất nhanh hơn và đơn giản hơn so với cách tiếp cận tóm tắt. Cách tiếp cận này dẫn đến độ chính xác cao hơn vì trích xuất trực tiếp các câu để độc giả đọc tóm tắt với các thuật ngữ chính xác tồn tại trong văn bản gốc.

Đương nhiên bên cạnh đó nó cũng có nhiều mặt hạn chế ví dụ như: các câu trích xuất có thể dài hơn bình thường, thí ngữ nghĩa và sự liên kết trong các câu tóm tắt, tóm tắt của bạn có thể rất dài vì các thông tin quan trọng rải đều ở các câu…

2) Các phương pháp trích xuất văn bản

- Phương pháp dựa trên thống kê: Các phương pháp này trích xuất các câu và từ quan trọng từ văn bản nguồn dựa trên phân tích thống kê của một tập hợp các tính năng. Câu "quan trọng nhất" được định nghĩa là "có vị trí thuận lợi nhất", "thường xuyên nhất", v.v. Các bước cho điểm câu của trình tóm tắt chiết xuất dựa trên thống kê bao gồm chọn và tính toán một số đặc điểm thống kê, sau đó ấn định trọng số cho họ ấn định điểm cuối cùng cho mỗi câu trong tài liệu được xác định bằng cách sử dụng phương trình trọng lượng của đối tượng địa lý (nghĩa là tất cả điểm số của các đối tượng địa lý đã chọn được tính toán và tổng hợp để thu được điểm của mỗi câu).
- Phương pháp dựa trên máy học: Các phương pháp này chuyển đổi bài toán tóm tắt thành bài toán phân loại có giám sát ở cấp độ câu. Hệ thống học theo các ví dụ để phân loại từng câu của

tài liệu kiểm tra dưới dạng lớp "tóm tắt" hoặc "không tóm tắt" bằng cách sử dụng một bộ tài liệu đào tạo (tức là một bộ sưu tập các tài liệu và các bản tóm tắt tương ứng do con người tạo ra). Đối với việc trích xuất dựa trên máy học, các bước cho điểm câu bao gồm trích xuất các câu từ tài liệu được xử lý trước cung cấp các tính năng đã trích xuất cho mạng nơ-ron tạo ra giá trị duy nhất dưới dạng điểm đầu ra.

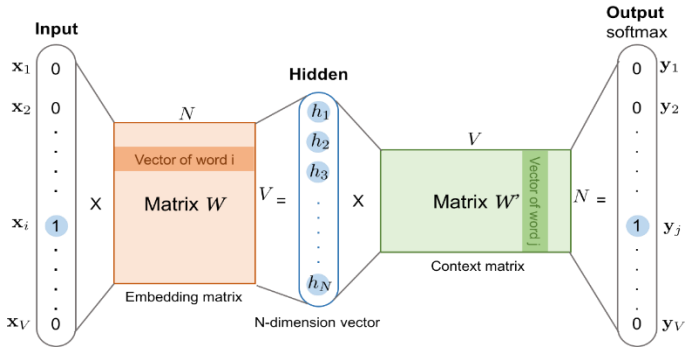
- Phương pháp dựa trên học tập sâu: Biểu diễn các câu bằng phương pháp nhúng (Embedding) thành các matrix (tập hợp các vector của từ) sau đó dùng CNN để đánh giá điểm của từng câu. Với việc chọn ra ngưỡng phù hợp để xác định xem câu đó là 'tóm tắt' hay 'không tóm tắt'. Phương pháp này cho kết quả tốt nhất trong các phương pháp và thời giảm đáng kể thời gian chạy của mô hình.

IV. Kiến thức liên quan

1) Embedding

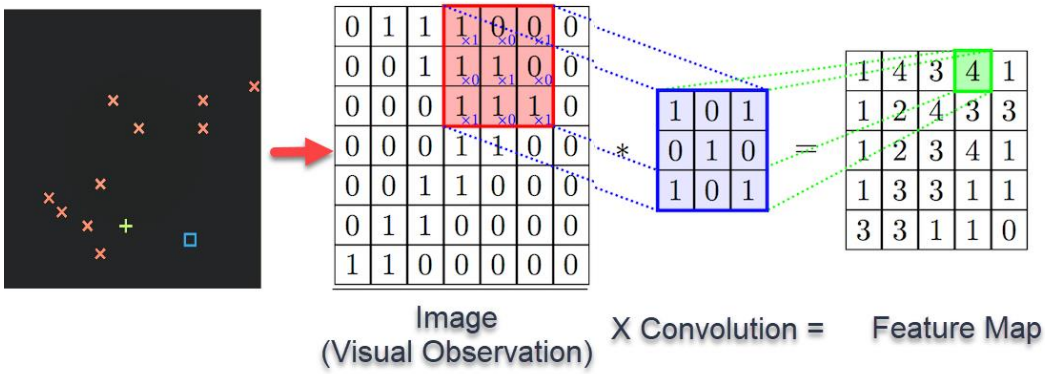
Các phương pháp máy học đều luôn yêu cầu đầu vào là một tập các giá trị số được thể hiện dưới dạng vector. Tuy nhiên, với các bài toán xử lý ngôn ngữ tự nhiên, dữ liệu đầu vào luôn là các chuỗi ký tự. Để áp dụng được các mô hình máy học, chúng ta cần phải chuyển đổi các chuỗi ký tự này thành dạng số mà vẫn giữ được các yếu tố về hình thái (hoa thường, nguyên âm, hợp âm,…), ngữ nghĩa và ngữ pháp.

Với sự ra đời của mô hình word2vec được Mikolov (Mikolov, Chen, Corrado, & Dean, 2013) giới thiệu để biểu diễn từ dưới dạng vector đã mang lại một bước tiếng lớn cho cộng đồng xử lý ngôn ngữ tự nhiên khi có thể dung nó để sử dụng cho các kiến trúc Neural Network, chúng ta sử dụng ma trận Word Embedding để ánh xạ các từ thành các vector số thực. Ma trận word embedding có số cột tương đương với số từ nằm trong bộ từ điển từ vựng, số dòng là số chiều của vector đại diện cho từ. Như vậy, khi chuyển đổi một từ thành vector số thực, chúng ta chỉ cần truy cập vào ma trận word embedding và lấy ra hàng tương ứng. Các giá trị trong ma trận đạt được bằng cách áp dụng các thuật toán như word2vec (CBOW, Skip-gram)[3] , GloVe (Pennington, Socher and Manning)[2] hoặc FastText (Bojanowski and Grave and Joulin and Mikolov)[1] trên tập dữ liệu văn bản không có nhãn như tin tức online. Ngoài ra, chúng ta cũng có thể khởi tạo các giá trị này ngẫu nhiên và được cập nhật chung trong quá trình huấn luyện mô hình gán nhãn.

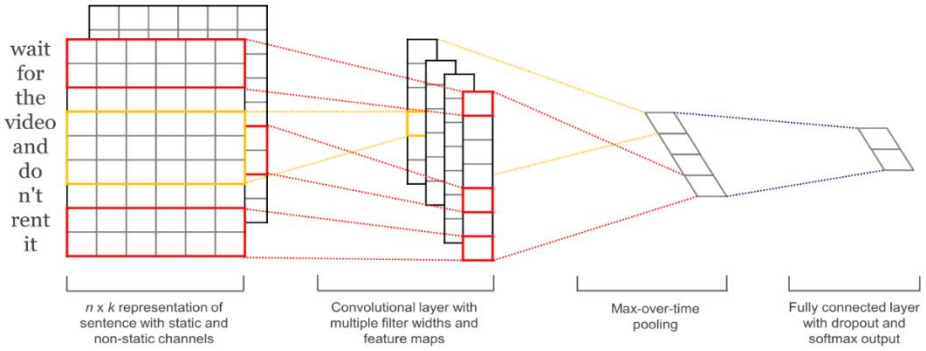


Mô hình embedding

2) Mạng tích chập (Convolution neural network)



Mạng tích chập là một mô hình nổi tiếng trong xử lý ảnh, nó hoạt động bằng cách cho các kernel trượt theo chiều dài và chiều rộng của tấm ảnh. Với cách hoạt động theo từng cụm kernel như vậy thì qua một lần tính tích chập thì ta có thể học được một vùng điểm ảnh. Qua vài lớp tích chập đầu, mô hình có thể học được các đường nét trong tấm ảnh, càng vào các layer sau thì các đường nét lúc đầu được hợp lại và tạo ra một hình dạng cụ thể khi ở lớp cuối. Từ đó ta có thể sử dụng chúng để dự đoán. Với đặt tính học theo từng vùng lân cận khiến để rút trích ra những đặt điểm quan trọng của từng vùng rồi tập hợp lại về sau và việc khởi tạo những kernel với kích cỡ 3x3 hay 5x5 làm parameter ít đi giúp giảm được dung lượng mô hình, cải thiện tốc độ tính toán của mô hình làm cho các mạng tích chập trở nên nổi tiếng và mạnh mẻ.



Bên cạnh sử dụng mạng tích chập trong việc xử lý ảnh thì ta cũng có thể áp dụng nó vào việc xử lý ngôn ngữ tự nhiên như Yoon Kim[4] hay Zhang 2016 [5] (Zhang, Meng, Pratama) từng đề xuất. Với đặc điểm khác biệt đầu giữa xử lý ảnh và xử lý ngôn ngữ tự nhiên làm cho qua trình tích chập của ta cũng khác đi đôi chút. Trong xử lý ngôn ngữ, ta chỉ khởi tạo kernel và trượt nó trên dọc từ đầu câu đến cuối câu. Nhờ vậy mà mô hình tích chập có thể học được những từ cạnh nhau, mang sự kết nối giữa hai từ liền kệ để có thể học ra những cụm từ có ý nghĩa chứ khôn còn riêng lẻ nữa.

Ngoài tích chập ra thì nhiều hướng phát triển khác vẫn được nhiều người hướng đến như sử dụng pretrained của bert trong các bài báo của Yang Liu 2019(Liu, Lapata)[6] và Zhang 2019 (Zhang, Cai, Xu, Wang)[7]

V. Quá trình tiến hành

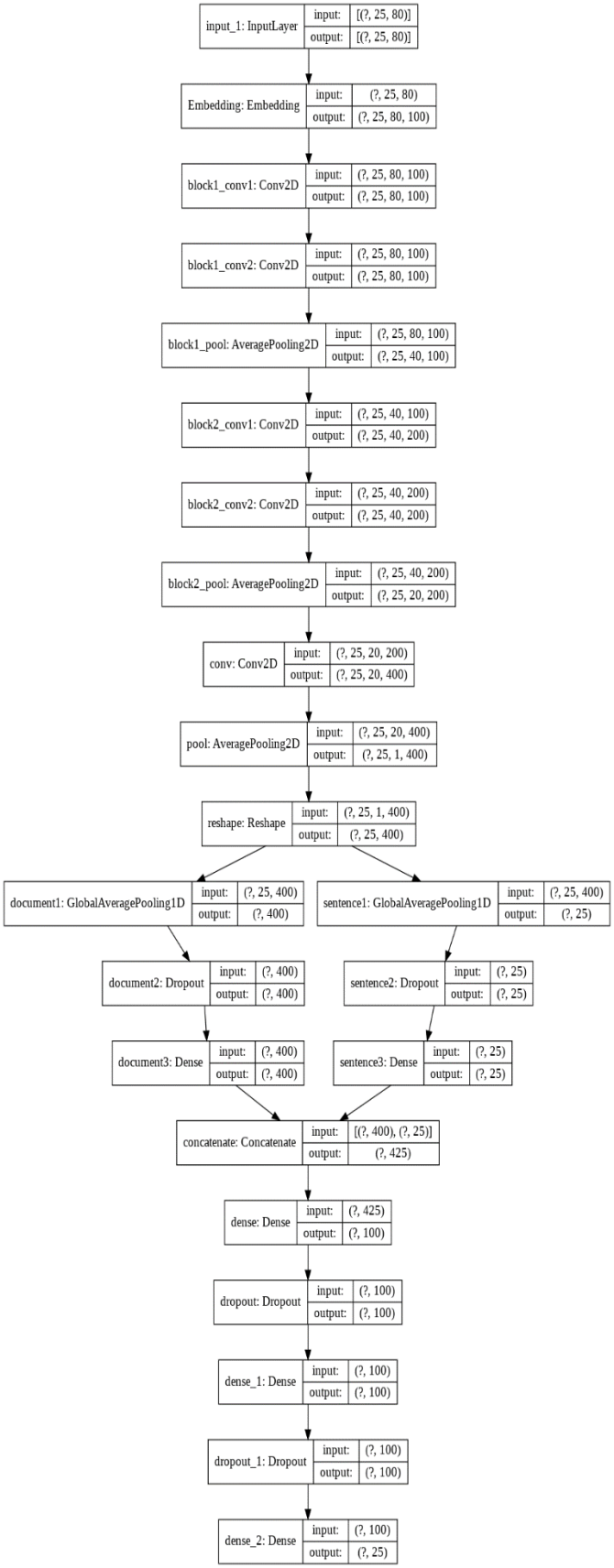
1) Tiền xử lý:

Với cộng việc là đánh giá điểm số của từng câu trong đoạn văn so với cả đoạn văn của nó thì ta có thể bỏ qua cấu trúc và thứ tự các từ trong câu. Thay vào đó, ta sẽ cố gắng để tìm ra các cụm từ mang ý nghĩa chính và rút trích thông tin ra để so với thông tin của cả đoạn văn. Do việc chú tâm vào nội dung và ý nghĩa nên ta có thể chuyển những động từ ở nhiều dạng khác nhau về dạng nguyên mẫu để có thể giảm thiểu tập từ, tăng tốc độ khi chuyển từ xâu sang số nhưng vẫn không mất đi ý nghĩa chính trong câu văn. Bên cạnh công việc xử lý các dạng động từ thì ta cũng tiến hành thêm việc bỏ những kí tự đặt biệt, chuyển từ kí tự viết hoa về kí tự viết thường.

Sau khi công việc trên thì ta tiếp tục phân từng câu trong đoạn và thực hiện công việc mã hóa những kí tự trong đoạn văn về dạng số với tập từ điển là 30000 từ, và giới hạn một câu chỉ có 80 từ và một đoạn văn sẽ có tối đa 25 câu.

2) Mô hình:

Lấy ý tưởng chính là mô hình mạng tích chập, nhóm xây dựng mô hình với nhiều lớp tích chập chồng lên nhau. Đầu vào của mô hình có dạng (None, 25, 80) tương ứng với 25 câu và mỗi câu 80 từ được cho qua lớp embedding. Lớp embedding này được nhóm chọn pre-trained GloVe của Pennington [2] 2014 (Pennington, Socher, Manning) với số chiều của mỗi vector là 100 nên sao khi qua lớp này, mỗi sample sẽ có dạng là (None, 25, 80, 100). Tiếp đến, ta cho qua 2 block, mỗi block bao gồm 2 lớp tích chập và 1 lớp pooling. Sau khi cho qua



2 block này thì ta có được một tập tập feature có dạng là (None, 25, 400). Như vậy, từ 80 từ trong câu, ta đã học và kết hợp từng cụp từ lại và cho ra 400 feature mới mang ý nghĩa tổng hợp giữa các từ trong câu.

Sau khi tách feature thì ta sẽ xây dựng mô hình đi theo 2 hướng:

- Hướng 1: Ta cho phép mô hình học được mối liên hệ giữa từng câu trong đoạn để có được một ý nghĩa tổng quát của cả đoạn văn.
- Hướng 2: Ta vẫn tiếp tục cho mô hình học về mối liên hệ giữa từng từ trong mỗi câu.

Bước tiếp theo ta sẽ gộp 2 hướng trên lại và cho qua vài lớp fully-connected để có thể cho ra kết quả tính điểm số của từng câu so với cả đoạn văn. Đầu ra sẽ có dạng là (None, 25) tương ứng với điểm của 25 câu trong đầu vào.

3) Kết quả:

Với mô hình dự đoán mức độ quan trọng của từng câu trong đoạn văn so với toàn bộ đoạn văn nên đầu ra sẽ là một con số đánh giá nằm trong khoản từ 0 đến 1 và tổng điểm tất cả câu trong đoạn điều bằng 1.Bài toán được đưa về dạng regression nên metric đánh giá nhóm dùng sẽ là mse và loss cũng là mse.

Mới loss là mse và metric cũng là mse thì kết quả của nhóm hiện tại trên các tập là:

Set	Metric
Train	0.063
Valid	0.066
Test	0.065

4) Hướng phát triển

Với hướng tiếp cận này thì bài toán vẫn còn nhiều vấn đề gặp phải khi mà sử dụng điểm đánh giá nằm trong khoản từ 0-1 và kèm theo đó là hàm loss cùng metric là mse. Việc này dẫn đến hàm loss nhỏ và thật sự khó có thể đánh giá và cập nhật trong mỗi epoch. Bên cạnh có một sốt mặt hạn chết khi sử dụng CNN làm cho mô hình trở nên chưa tốt. Ngoài ra, phương pháp extractive này cũng có mặt hạn chết khi chúng ta lượt bỏ bớt câu nên sẽ mất đi một số câu chứa một số nội dung đặt biệt nhưng chỉ ít quan trọng hơn chút so với những câu được chọn.

Chính vì vậy, trong tương lai nhóm sẽ tìm hiểu thêm về một số mô hình mới như bert để có thể xử lý và khắc phục những lỗi và mô hình, Bên cạnh đó nhóm cũng sẽ tìm hiểu và tiếp cận bài toán theo hướng abstractive để có thể tạo ra đoạn văn tóm tắt mang nội dung đầy đủ và tối ưu hơn so với phương pháp extractive.

References

[1] Enriching Word Vectors with Subword Information (Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov)

[2] GloVe: Global Vectors for Word Representation (Jeffrey Pennington, Richard Socher, Christopher D. Manning)

[3] Efficient Estimation of Word Representations in Vector Space (Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean)

[4] Convolutional Neural Networks for Sentence Classification (Yoon Kim)

[5] Extractive document summarization based on convolutional neural networks (Yong Zhang; Joo Er Meng; Mahardhika Pratama)

[6] Text Summarization with Pretrained Encoders (Yang Liu and Mirella Lapata)

[7] Pretraining-Based Natural Language Generation for Text Summarization (Haoyu Zhang , Jingjing Cai , Jianjun Xu , Ji Wang)