TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**VÕ THIỆN TRUNG - 51703211**

**NGUYỄN HOÀNG TÚ - 51703215**

**KHỞI TẠO NGỮ LIỆU VÀ XÂY DỰNG ỨNG DỤNG HỎI ĐÁP TỰ ĐỘNG TRONG LĨNH VỰC DU LỊCH**

**DỰ ÁN CÔNG NGHỆ THÔNG TIN 2**

**HỆ THỐNG THÔNG TIN**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2022**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**VÕ THIỆN TRUNG - 51703211**

**NGUYỄN HOÀNG TÚ - 51703215**

**KHỞI TẠO NGỮ LIỆU VÀ XÂY DỰNG ỨNG DỤNG HỎI ĐÁP TỰ ĐỘNG TRONG LĨNH VỰC DU LỊCH**

**DỰ ÁN CÔNG NGHỆ THÔNG TIN 2**

**HỆ THỐNG THÔNG TIN**

Người hướng dẫn

**TRẦN THANH PHƯỚC**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2022**

**LỜI CẢM ƠN**

Để hoàn thành bộ môn dự án CNTT 2 lần này chúng em xin gửi lời cảm ơn tới Thầy Trần Thanh Phước đã tận tình hướng dẫn giúp đỡ, đóng góp ý kiến và chỉ bảo nhiệt tình

Đồng thời, chúng em xin chân thành bày tỏ lòng biết ơn đến toàn thể quý thầy cô giáo trong trường đã tận tình hỗ trợ trong suốt quá trình làm bài. Nhờ vào các kiến thức, đặc biệt là kiến thức chuyên môn, được truyền dạy lại sẽ là cơ sở lý thuyết vững vàng và tạo điều kiện giúp đỡ tôi trong suốt quá trình học tập và làm việc hiện tại cũng như sau này.

Chúng em đã cố gắng hoàn thành đồ án một cách tốt nhất nhưng sẽ có những thiếu sót mà chúng em không nhận thấy. Chúng em mong nhận được những phản hồi và ý kiến của các thầy cô để bài đồ án sau chúng em có thể hoàn thành tốt hơn, chúng em xin chân thành cảm ơn.

**ĐỒ ÁN ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Tôi xin cam đoan đây là sản phẩm đồ án của riêng chúng tôi và được sự hướng dẫn của Thầy Trần Thanh Phước. 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 tháng năm*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

*Võ Thiện Trung Nguyễn Hoàng Tú*

**PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN**

**Phần xác nhận của GV hướng dẫn**

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

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

**Phần đánh giá của GV chấm bài**

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

TP. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

**TÓM TẮT**

**MỤC LỤC**

[DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT 7](#_Toc121771580)

[CÁC CHỮ VIẾT TẮT 7](#_Toc121771581)

[DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ 8](#_Toc121771582)

[DANH MỤC BẢNG 8](#_Toc121771583)

[DANH MỤC HÌNH 8](#_Toc121771584)

[CHƯƠNG 1: GIỚI THIỆU 9](#_Toc121771585)

[1.1 Giới thiệu về đề tài 9](#_Toc121771586)

[1.2 Phân tích yêu cầu bài toán 9](#_Toc121771587)

[CHƯƠNG 2: CƠ SỞ LÝ THUYẾT 10](#_Toc121771588)

[2.1 Encoder và Decoder trong BERT 10](#_Toc121771589)

[2.2 Giới thiệu về BERT 11](#_Toc121771590)

[2.2.1 Các kiến trúc model BERT 11](#_Toc121771591)

[2.2.2 Cách huấn luyện BERT 12](#_Toc121771592)

[2.2.3 Embeddings BERT 12](#_Toc121771593)

[2.3 RoBERTa, DistillBERT, PhoBERT 14](#_Toc121771594)

[2.3.1 RoBERTa 14](#_Toc121771595)

[2.3.2 DistillBERT 14](#_Toc121771596)

[2.3.3 PhoBERT 14](#_Toc121771597)

[2.4 Mã hóa BPE (Byte Pair Encoding) 16](#_Toc121771598)

[2.5 Metric 17](#_Toc121771599)

[2.4.1 F1 –score 17](#_Toc121771600)

[2.4.2 Exact Match 17](#_Toc121771601)

[CHƯƠNG 3: XÂY DỰNG MÔ HÌNH 18](#_Toc121771602)

[3.1 SQuAD Dataset 18](#_Toc121771603)

[CHƯƠNG 4: THỰC NGHIỆM 19](#_Toc121771604)

[4.1 Dữ liệu 19](#_Toc121771605)

[4.2 Tiến hành thực nghiệm 19](#_Toc121771606)

[4.2.1 Quá trình huấn luyện 19](#_Toc121771607)

[4.2.2 Phương pháp đánh giá 19](#_Toc121771608)

[CHƯƠNG 5: KẾT LUẬN VÀ ĐÁNH GIÁ 20](#_Toc121771609)

[TÀI LIỆU THAM KHẢO 21](#_Toc121771610)

# DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT

# CÁC CHỮ VIẾT TẮT

# DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ

# DANH MỤC BẢNG

# DANH MỤC HÌNH

CHƯƠNG 1: GIỚI THIỆU

* 1. Giới thiệu về đề tài

Hỏi đáp (question answering) là một bài toán khó giải của Xử lý ngôn ngữ tự nhiên. Trong những năm gần đây, cộng đồng Học máy đã áp dụng nhiều kỹ thuật mới để tăng độ chính xác của các hệ thống hỏi đáp tự động. Tuy nhiên, hầu hết các công trình tập trung vào ngôn ngữ Anh vì phần lớn cộng đồng nghiên cứu sử dụng ngôn ngữ này.

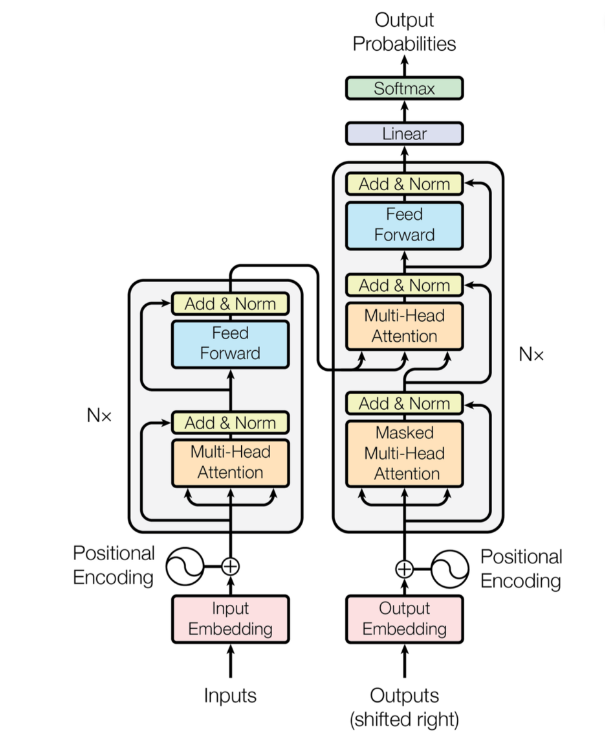
* 1. Phân tích yêu cầu bài toán

Hỏi đáp (Question and Answering) Là thuật toán hỏi và đáp. Đầu vào là một cặp câu (pair sequence) bao gồm: câu hỏi (question) có chức năng hỏi và đoạn văn bản (paragraph) chứa thông tin trả lời cho câu hỏi. Một bộ dữ liệu chuẩn nằm trong GLUE dataset được sử dụng để đánh giá nhiệm vụ hỏi và đáp là SQuAD - Stanford Question Answering Dataset

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

2.1 Encoder và Decoder trong BERT

Trước khi hiểu về BERT cần tìm hiểu về kỹ thuật transformer. Mô hình hoàn toàn không sử dụng các kiến trúc mạng hồi quy của RNN mà chỉ sử dụng các các tầng attention để nhúng các từ trong câu. Kiến trúc cụ thể của mô hình như sau:



Cấu trúc mạng Encoder – Decoder bao gồm 2 khối chính: Mã hoá (Encode) và Giải mã (Decode). Hai khối được kết nối với nhau bởi vector bối cảnh (context vector). Trong đó, khối mã hoá (Encoder) có nhiệm vụ xử lý dữ liệu đầu vào là một chuỗi các từ hoặc token. Tất cả các thông tin sẽ được xử lý, biến đổi thành một vector có độ dài nhất định (vector ngữ cảnh) và chuyển sang khối Giải mã. Khối Giải mã (Decoder) sẽ có chức năng đọc vector ngữ cảnh và dự đoán kết quả đầu ra sẽ là một chuỗi từ hoặc ký tự

Hai khối chính của mô hình:

* Encoder: Bao gồm 6 tầng liên tiếp nhau. Mỗi một tầng sẽ bao gồm một tầng con (sub-layer) là Multi-Head Attention kết hợp với tầng kết nối đầy đủ (fully-connected layer) như mô tả ở nhánh encoder bên trái của hình vẽ. Kết thúc quá trình encoder ta thu được một véc tơ đầu ra nhúng cho mỗi từ
* Decoder: Kiến trúc cũng bao gồm các tầng liên tiếp nhau. Mỗi một tầng của Decoder cũng có các tầng con gần tương tự như tầng của Encoder nhưng bổ sung thêm tầng con đầu tiên là Masked Multi-Head Attention có tác dụng loại bỏ các từ trong tương lai khỏi quá trình chú ý (attention)

2.2 Giới thiệu về BERT

[BERT](https://arxiv.org/pdf/1810.04805.pdf) là viết tắt của cụm từ Bidirectional Encoder Representation from Transformer có nghĩa là mô hình biểu diễn từ theo 2 chiều ứng dụng kỹ thuật Transformer. BERT được thiết kế để huấn luyện trước các biểu diễn từ (pre-train word embedding). Điểm đặc biệt ở BERT đó là nó có thể điều hòa cân bằng bối cảnh theo cả 2 chiều trái và phải. Một điểm đặc biệt ở BERT mà các model embedding trước đây chưa từng có đó là kết quả huấn luyện có thể fine-tuning được. Chúng ta sẽ thêm vào kiến trúc model một output layer để tùy biến theo tác vụ huấn luyện, do đó BERT được áp dụng vào nhiều bài toán như: Hỏi đáp tự động, phân loại, tóm tắt văn bản…

Cơ chế attention của Transformer sẽ truyền toàn bộ các từ trong câu văn đồng thời vào mô hình một lúc mà không cần quan tâm đến chiều của câu. Do đó Transformer được xem như là huấn luyện hai chiều (bidirectional) mặc dù trên thực tế chính xác hơn chúng ta có thể nói rằng đó là huấn luyện không chiều (non-directional). Đặc điểm này cho phép mô hình học được bối cảnh của từ dựa trên toàn bộ các từ xung quanh nó bao gồm cả từ bên trái và từ bên phải.

2.2.1 Các kiến trúc model BERT

Hiện tại có nhiều phiên bản khác nhau của model BERT. Các phiên bản đều dựa trên việc thay đổi kiến trúc của Transformer tập trung ở 3 tham số:

*L*: số lượng các block sub-layers trong transformer

*H*: kích thước của embedding véc tơ (hay còn gọi là hidden size)

*A*: Số lượng head trong multi-head layer, mỗi một head sẽ thực hiện một self-attention.

Tên gọi của 2 kiến trúc bao gồm:

* **BERTBASE** (L = 12, H = 768, A =12): Tổng tham số : 110 triệu
* **BERTLARGE**(L = 24, H = 1024, A =16): Tổng tham số : 340 triệu

Như vậy ở kiến trúc BERT Large tăng gấp đôi số layer, tăng kích thước hidden size của embedding véc tơ gấp 1.33 lần và tăng số lượng head trong multi-head layer gấp 1.33 lần.

2.2.2 Cách huấn luyện BERT

BERT được huấn luyện đồng thời 2 nhiệm vụ gọi là Masked LM (để dự đoán từ thiếu trong câu) và Next Sentence Prediction (NSP – dự đoán câu tiếp theo câu hiện tại). Hai nhiệm vụ này được huấn luyện đồng thời và loss tổng sẽ là kết hợp loss của 2 nhiệm vụ và mô hình sẽ cố gắng minimize loss tổng này.

Chi tiết 2 nhiệm vụ này như sau:

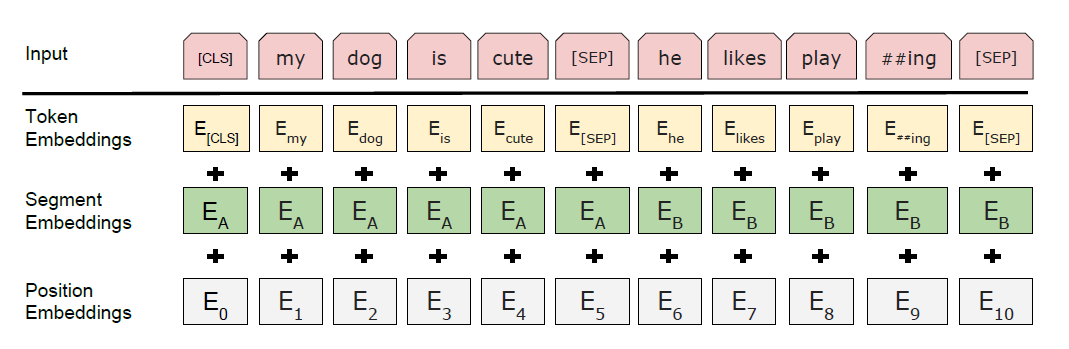
Masked Language Model: Với nhiệm vụ này, huấn luyện sẽ thực hiện che đi tầm 15% số từ trong câu và đưa vào mô hình Và sẽ huấn luyện để mô hình predict ra các từ bị che đó dựa vào các từ còn lại. Cụ thể là:

* Thêm một lớp classification lên trên encoder đầu ra
* Đưa các véc tơ trong encoder ouput về véc tơ bằng với vocab size, sau đó softmax để chọn ra từ tương ứng tại mỗi vị trí trong câu.
* Loss sẽ được tính tại vị trí masked và bỏ qua các vị trí khác (để đánh giá xem mô hình dự đoán từ mask đúng/sai ntn mà, các từ khác đâu có liên quan)

Next Sentence Prediction (NSP): Với nhiệm vụ này thì mô hình sẽ được feed cho một cặp câu và nhiệm vụ của nó là đầu ra ra giá trị 1 nếu câu thứ hai đúng là câu đi sau câu thứ nhất và 0 nếu không phải. Trong quá trinh huấn luyện, ta chọn 50% mẫu là Positive (đầu ra là 1) và 50% còn lại là Negative được ghép linh tinh (đầu ra là 0).

2.2.3 Embeddings BERT

Các Embeddings không gì ngoài các vector gói gọn ngữ nghĩa của từ, các từ tương tự sẽ có số gần giống nhau trong vertor của chúng . Các embeddings đầu vào được tạo thành từ ba phần riêng biệt. Sơ đồ dưới đây sẽ cho thấy cách các Embeddings kết hợp với nhau để tạo ra đầu vào cho BERT



Lý do chuyển đổi các từ thành Embeddings là để làm cho mô hình hoạt động dễ dàng hơn. Khi các từ được chuyển đổi thành dạng Embeddings mô hình có thể hiểu tầm quan trọng về ngữ nghĩa của một từ ở dạng số và có thể thực hiện các phép toán trên từ đó.Trong hình trên, có thể thấy ngoài từ Embeddings, còn có một số mã khác như CLS và SEP. Đây là những mã đặc biệt truyền đạt ý nghĩa cụ thể cho Encoder. Chi tiết như sau:

* **CLS:** là từ viết tắt của classification. Nó được thêm vào để thể hiện sự phân loại ở cấp độ câu. Nó thường được thêm vào đầu một chuỗi để nó đại diện cho toàn bộ câu.
* **SEP:** được sử dụng trong quá trình dự đoán câu tiếp theo. Đó là dấu phân cách giúp mô hình biết đâu là câu tiếp theo. Trong trường hợp của một câu duy nhất, nó chỉ được thêm vào cuối.
* **MASK:** Mã này được sử dụng trong giai đoạn MLM. Nó được sử dụng để đại diện cho từ ẩn cho mô hình.

Ngữ nghĩa của từ (token embeddings): Thông qua các embedding véc tơ cho từng từ. Các véc tơ được khởi tạo từ pretrain model.

Loại câu (segment embeddings): Gồm hai véc tơ là EA nếu từ thuộc câu thứ nhất và EB nếu từ thuộc câu thứ hai.

Vị trí của từ trong câu (position embedding): là các véc tơ E0,…E10. Tương tự như positional embedding trong transformer.

Véc tơ input sẽ bằng tổng của cả ba thành phần embedding theo token, segment và position

2.3 RoBERTa, DistillBERT, PhoBERT

2.3.1 RoBERTa

RoBERTa (A Robustly Optimized BERT) được giới thiệu bởi Facebook là một phiên bản được huấn luyện, kế thừa các kiến trúc và thuật toán của model BERT trên framework pytorch với một phương pháp huấn luyện tốt hơn với dữ liệu được tăng gấp 10 lần.

Đây là một project hỗ trợ việc huấn luyện lại các model BERT trên những bộ dữ liệu mới cho các nguôn ngữ khác ngoài một số ngôn ngữ phổ biến. Kể từ khi ra đời, đã có rất nhiều các mô hình pretrain cho những ngôn ngữ khác nhau được huấn luyện trên RoBERTa.

Để tăng cường quá trình huấn luyện, RoBERTa không sử dụng cơ chế dự đoán câu kế tiếp (NSP) từ BERT mà sử dụng kỹ thuật mặt nạ động (dynamic masking), theo đó các [token](https://trituenhantao.io/tu-dien-thuat-ngu/token/) mặt nạ sẽ bị thay đổi trong quá trình huấn luyện. Sử dụng kích thước [batch](https://trituenhantao.io/tu-dien-thuat-ngu/batch/) lớn hơn cho thấy hiệu quả tốt hơn khi huấn luyện.

2.3.2 DistillBERT

DistilBERT học một phiên bản xấp xỉ của BERT, giữ lại 97% hiệu quả dự đoán nhưng chỉ sử dụng một nửa tham số. DistilBERT sử dụng kỹ thuật gọi là distillation, giúp xấp xỉ BERT như một giáo viên của DistilBERT. Ý tưởng ở đây là khi một mạng lớn đã được huấn luyện, phân bố xác xuất đầu ra của nó có thể được xấp xỉ bởi một mạng nhỏ hơn. Hàm loss được sử dụng trong xấp xỉ hậu nghiệm trong thống kê Bayes là Kulback Leiber divergence cũng được sử dụng khi huấn luyện DistilBERT.

2.3.3 PhoBERT

Mặc dù BERT là một nghiên cứu mới mang đầy tính đột phá, một bước nhảy vọt thực sự của Google trong lĩnh vực xử lý ngôn ngữ tự nhiên. Sự ra đời của mô hình huấn luyện trước BERT đã mang lại những cải tiến đáng kể cho rất nhiều bài toán như Question Answering, Sentiment Analysis, ... Tuy nhiên, huấn luyện mô hình BERT cho Tiếng Việt lại không hề đơn giản do đó rất khó để có thể áp dụng BERT cho các nhiệm vụ Tiếng Việt dù cho Google cũng có huấn luyện trước cho nhiều ngôn ngữ (pre-trained multilingual) bao gồm cả tiếng Việt nhưng chưa cho kết quả thực hiện tốt nhất.

PhoBERT đã được ra đời là một mô hình BERT được huấn luyện trước cho tiếng Việt và đạt được nhiều kết quả tốt nhất cho nhiều nhiệm vụ trong xử lý ngôn ngữ tiếng Việt. Tác giả lấy tên Pho vì đây là món ăn phổ biến của Việt Nam. PhoBERT dễ sử dụng, nó được xây dựng để sử dụng trong các thư viện như FAIRSeq của Facebook hay Transformers của Hugging Face nên giờ đây BERT lại càng phổ biến ngay cả với ngôn ngữ tiếng Việt hay tiếng Anh.

Đây là một mô hình huấn luyện trước được huấn luyện cho đơn ngôn ngữ (monolingual language), tức là chỉ huấn luyện dành riêng cho tiếng Việt. Tương tự như BERT, PhoBERT cũng có 2 phiên bản là PhoBERTbase với 12 khối transformers và PhoBERTlarge với 24 khối transformers.

PhoBERT có hai phiên bản: PhoBERT base và PhoBERT large

Là mô hình ngôn ngữ đơn ngữ có quy mô lớn được huấn luyện huấn luyện dành riêng cho Tiếng Việt. Thực nghiệm kết quả cho thấy PhoBERT luôn hoạt động tốt hơn so với mô hình đa ngôn ngữ được huấn luyện trước tốt nhất gần đây XLM-R và cải thiện tính năng hiện đại trong nhiều nhiệm vụ NLP dành riêng cho tiếng Việt bao gồm các nhiệm vụ như: phân tích giọng nói, Phân tích cú pháp phụ thuộc, Nhận dạng thực thể được đặt tên và Suy luận ngôn ngữ tự nhiên…



Tên hình :Khác biệt giữa BERT,PhoBERT , XLM-RoBERTa

2.4 Mã hóa BPE (Byte Pair Encoding)

Toknenize là quá trình mã hóa các văn bản thành các index dạng số mang thông tin của văn bản để cho máy tính có thể huấn luyện được. Khi đó mỗi một từ hoặc ký tự sẽ được đại diện bởi một index.

Trong NLP có một số kiểu tokenize như sau:

* **Tokenize theo word level**: Chúng ta phân tách câu thành các token được ngăn cách bởi khoảng trắng hoặc dấu câu. Khi đó mỗi token là một từ đơn âm tiết. Ví dụ: GloVe, word2vec
* **Tokenize theo multi-word level**: sử dụng thêm từ điển bao gồm cả từ đa âm tiết và đơn âm để tokenize câu. Ví dụ: VnCoreNLP, pyvivn, underthesea
* **Tokenize theo character level**:  Từ được token dựa trên level ký tự sẽ có tác dụng giảm kích thước từ điển mà vẫn biểu diễn được các trường hợp từ nằm ngoài từ điển. Đây là phương pháp được áp dụng trong mô hình fasttext.
* **Phương pháp mới BPE (SOTA) :** Nhược điểm của phương pháp tokenize theo character level đó là các token không có ý nghĩa nếu đứng độc lập. Do đó đối với các bài toán sentiment analysis, áp dụng tokenize theo character level sẽ mang lại kết quả kém hơn. Token theo word level cũng tồn tại hạn chế đó là không giải quyết được các trường hợp từ ngằm ngoài từ điển.

Một phương pháp mới đã được đề xuất trong bài báo [Neural Machine Translation of Rare Words with Subword Units](https://arxiv.org/pdf/1508.07909.pdf) vào năm 2016, có khả năng tách từ theo level nhỏ hơn từ và lớn hơn ký tự được gọi là subword. Phương pháp đó chính là BPE (byte pair encoding). Theo phương pháp mới này, hẫu hết các từ đều có thể biểu diễn bởi subword và chúng ta sẽ hạn chế được một số lượng đáng kể các token <unk> đại diện cho từ chưa từng xuất hiện trước đó. Rất nhanh chóng, Phương pháp mới đã được áp dụng ở hầu hết các phương pháp NLP hiện đại từ các lớp model BERT cho tới các biến thể của nó như OpenAI GPT, RoBERTa, DistilBERT, XLMNet. Kết quả áp dụng tokenize theo phương pháp mới đã cải thiện được độ chính xác trên nhiều tác vụ dịch máy, phân loại văn bản, dự báo câu tiếp theo, hỏi đáp, dự báo mối quan hệ văn bản.

2.5 Metric

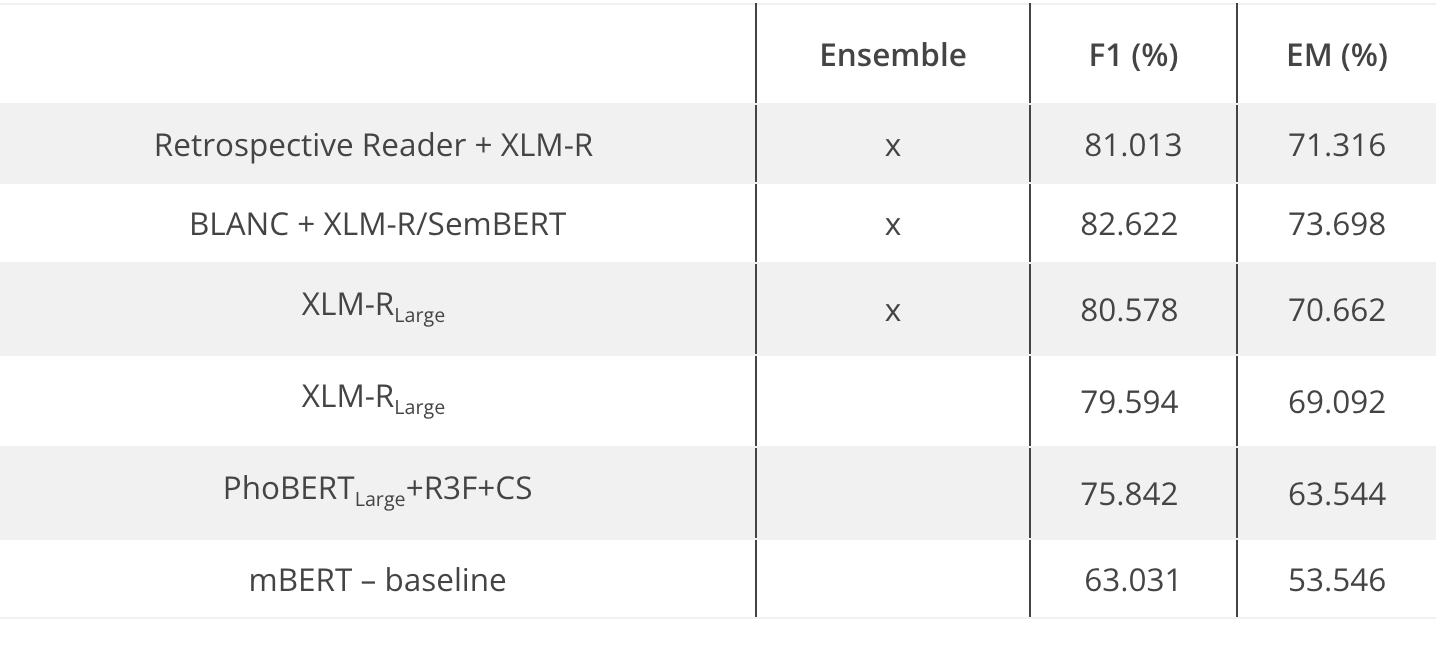
2.4.1 F1 –score

F1 score Chỉ số F1 là chỉ số kết hợp cả precision và recall, tính bằng harmonic mean. Chỉ số F1 có khoảng giá trị (0, 1], điểm càng cao cho thấy model có độ chính xác cao.

F1 = 2 \* Precision \* Recall / (Precision + Recall)

2.4.2 Exact Match

Exact match (EM) Để đánh giá độ chính xác của model MRC, câu trả lời mẫu và câu trả lời dự đoán sẽ được so sánh xem có giống hệt nhau hay không. Nếu giống nhau, EM là 1. Nếu không giống nhau, EM là 0. Giá trị EM cuối cùng là trung bình cộng của tất cả các giá trị EM của từng cặp câu hỏi-trả lời.



. Tên hình: Kết quả một số phương pháp trên tập kiểm tra công khai của VLSP2021 – ViMRC challenge

CHƯƠNG 3: XÂY DỰNG MÔ HÌNH

3.1 SQuAD Dataset

CHƯƠNG 4: THỰC NGHIỆM

Trong đồ án này, chúng tôi đề xuất sử dụng phương pháp PhoBERT để ứng dụng cho bài toán xây dựng hệ thống hỏi đáp tự động trên miền dữ liệu tiếng Việt

4.1 Dữ liệu

Dữ liệu để huấn luyện và đánh giá mô hình sẽ là những bài viết về du lịch trên các trang thông tin du lịch điện tử. Chúng tôi thu thập và sử dụng nguồn dữ liệu được công khai từ cẩm nang du lịch của Vnexpress (https://vnexpress.net/du-lich/cam-nang) và traveloka (https://www.traveloka.com/). Bộ dữ liệu thu được gồm các đoạn văn bản về lĩnh vực du lịch - ẩm thực của các khu vực như: Đà Lạt, Đà Nẵng, Phú Quốc, Vũng Tàu…

…………….còn tiếp

4.2 Tiến hành thực nghiệm

4.2.1 Quá trình huấn luyện

4.2.2 Phương pháp đánh giá

CHƯƠNG 5: KẾT LUẬN VÀ ĐÁNH GIÁ

# TÀI LIỆU THAM KHẢO

**Tiếng Anh**

**Tiếng việt**