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**

**CÁC CHỮ VIẾT TẮT**

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

**DANH MỤC BẢNG**

**DANH MỤC HÌNH**

CHƯƠNG 1: GIỚI THIỆU

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

2.1 BERT

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

2.3 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.

**Thuật toán BPE**

BPE (Byte Pair Encoding) là một kỹ thuật nén từ cơ bản giúp chúng ta index được toàn bộ các từ kể cả trường hợp từ mở (không xuất hiện trong từ điển) nhờ mã hóa các từ bằng chuỗi các từ phụ (subwords). Nguyên lý hoạt động của BPE dựa trên phân tích trực quan rằng hầu hết các từ đều có thể phân tích thành các thành phần con.

Chẳng hạn như từ: ***low***, ***lower***, ***lowest*** đều là hợp thành bởi low và những đuôi phụ ***er***, ***est***. Những đuôi này rất thường xuyên xuất hiện ở các từ. Như vậy khi biểu diễn từ lower chúng ta có thể mã hóa chúng thành hai thành phần từ phụ (subwords) tách biệt là **low** và **er**. Theo cách biểu diễn này sẽ không phát sinh thêm một index mới cho từ lower và đồng thời tìm được mối liên hệ giữa ***lower***, ***lowest*** và ***low*** nhờ có chung thành phần từ phụ là ***low***.

Phương pháp BPE sẽ thống kê tần suất xuất hiện của các từ phụ cùng nhau và tìm cách gộp chúng lại nếu tần suất xuất hiện của chúng là lớn nhất. Cứ tiếp tục quá trình gộp từ phụ cho tới khi không tồn tại các subword để gộp nữa, ta sẽ thu được tập subwords cho toàn bộ bộ văn bản mà mọi từ đều có thể biểu diễn được thông qua subwords.

2.4 Metric

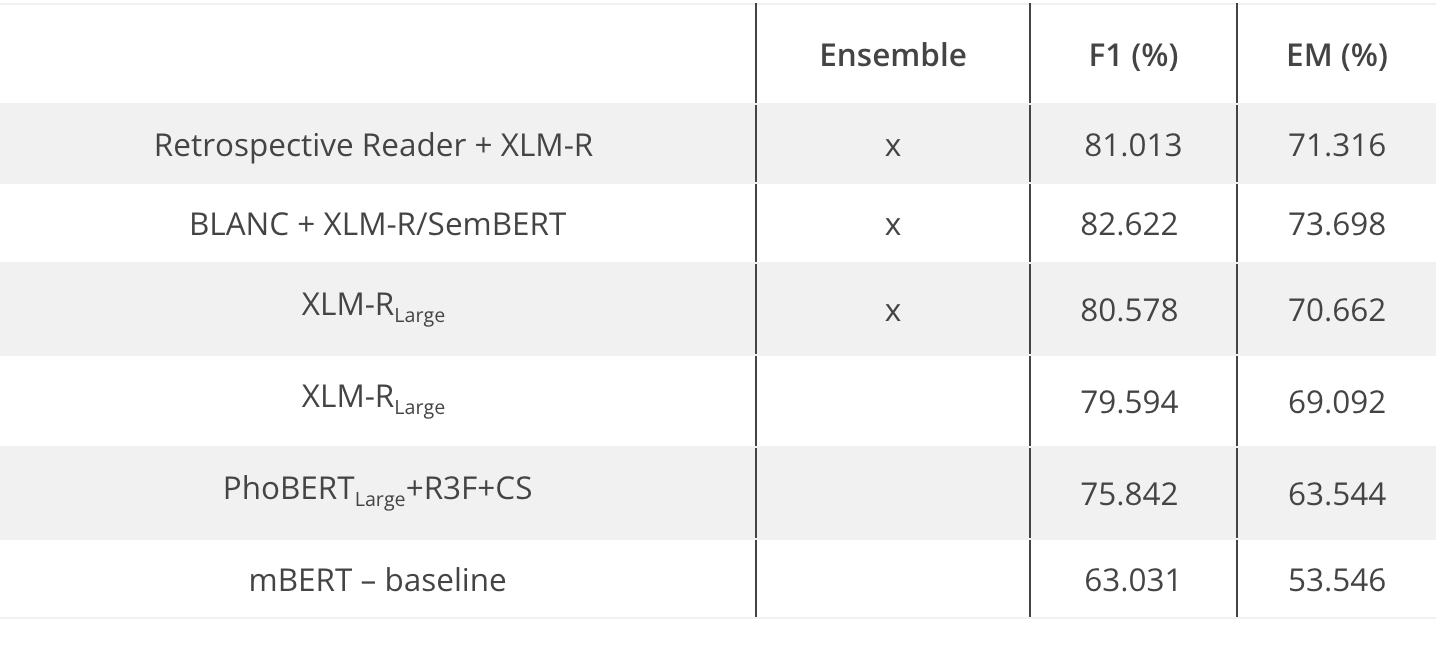
2.4.1 F1 –score

**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:

CHƯƠNG 4: THỰC NGHIỆM

CHƯƠNG 5 KẾT LUẬN VÀ ĐÁNH GIÁ

**TÀI LIỆU THAM KHẢO**

**Tiếng Anh**

**Tiếng việt**