TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**CUỐI KỲ**

**XỬ LÝ NGÔN NGỮ TỰ NHIÊN**

*Người hướng dẫn*: **PGS.TS. LÊ ANH CƯỜNG**

*Người thực hiện*: **HOÀNG ĐÌNH QUÝ VŨ – 521H0517**

**LÊ HUỲNH HUYỀN TRANG - 520C0156**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**

A red and blue logo

Description automatically generated

**CUỐI KỲ**

**XỬ LÝ NGÔN NGỮ TỰ NHIÊN**

*Người hướng dẫn*: **PGS.TS. LÊ ANH CƯỜNG**

*Người thực hiện*: **HOÀNG ĐÌNH QUÝ VŨ – 521H0517**

**LÊ HUỲNH HUYỀN TRANG - 520C0156**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024**

LỜI CẢM ƠN

Đây là phần tác giả **tự viết** ngắn gọn, thể hiện sự biết ơn của mình đối với những người đã giúp mình hoàn thành Luận văn/Luận án. Tuyệt đối không sao chép theo mẫu những “lời cảm ơn” đã có.

**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 **PGS.TS. Lê Anh Cường**. 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)*

*Hoàng Đình Quý Vũ*

*Lê Huỳnh Huyền Trang*

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 KÝ HIỆU**

*f Tần số của dòng điện và điện áp (Hz)*

*p Mật độ điện tích khối (C/m3)*

**CÁC CHỮ VIẾT TẮT**

CSTD Công suất tác dụng

MF Máy phát điện

BER Tỷ lệ bít lỗi

DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ

**DANH MỤC HÌNH**

[Hình 2.1: Kiến trúc FTP 1](#_Toc387689394)

**DANH MỤC BẢNG**

[Bảng 3.1 Ví dụ cho chèn bảng 1](#_Toc387689363)

CHƯƠNG 1 – Kiến trúc và huấn luyện BERT

CHƯƠNG 2 – SO SÁNH BERT VỚI CÁC BIẾN THỂ

* 1. Giới thiệu

Trong chương này, chúng ta sẽ so sánh BERT với hai biến thể nổi bật là RoBERTa và PhoBERT, đồng thời trình bày chi tiết kiến trúc, ưu điểm và nhược điểm của một trong những mô hình biến thể này.

Tóm tắt lại Bert:

* Sử dụng mô hình Transformer Encoder với 12 hoặc 24 layers.
* Training dựa trên hai nhiệm vụ chính: Masked Language Model (MLM) và Next Sentence Prediction (NSP).
* Sử dụng token embedding, segment embedding, và positional embedding.
  1. RoBERTa

Mô hình RoBERTa (A Robustly Optimized BERT) là phiên bản cải tiến của BERT. Kiến trúc RoBerta được đào tạo theo Kiến trúc của BERT (L = 24, h = 1024, A = 355M) và khác ở số tham số.

A white background with black text

Description automatically generated

So với BERT, về quy mô mô hình, khả năng tính toán và dữ liệu, đã có những cải tiến sau:

* Thông số mô hình lớn hơn (dựa trên thời gian huấn luyện được cung cấp trong bài báo, mô hình sử dụng 1024 GPU V100 để huấn luyện trong thời gian 1 ngày).
* Kích thước Batch size lớn hơn. ReBERTa sử dụng kích thước batch size lớn hơn trong quá trình đào tạo, đã thử với kích thước batch từ 256 đến 8000. Ppl là độ phức tạp(perlexity)

A black and white text

Description automatically generated

* A screenshot of a computer

  Description automatically generatedNhiều training data hơn (bao gồm 160GB plain text CC-NEWS. Trước đó BERT sử dụng 16GB dataset BookCorpus và Wikipedia tiếng Anh để training).

Ngoài ra, RoBERTa còn có những cải tiến về phương pháp training như sau:

* Loại bỏ dự đoán câu tiếp theo (Next Sentence Prediction – NSP).

A table with numbers and a number of objects

Description automatically generated with medium confidence

* Mặt nạ động (Dynamic Masking): BERT dựa trên masking ngẫu nhiên (randomly masking) và dự đoán token (predicting tokens). Khi triển khai nguyên bản BERT, trong quá trình tiền xử lý dữ liệu, kết quả sẽ thu được một static mask. Trong khi đó RoBERTa sử dụng dynamic mask, mỗi khi một sequence được đưa vào mô hình thì một mẫu mặt nạ (masking pattern) mới sẽ được tạo ra. Bằng cách này, trong quá trình nhập liên tục một lượng lớn dữ liệu, mô hình sẽ dần thích ứng với các chiến lược masking khác nhau, và học được cách biểu diễn ngôn ngữ khác nhau.
* A table with numbers and text

  Description automatically generated
* Mã hóa văn bản (text encoding): Byte-Pair Encoding (BPE) là sự kết hợp giữa các biểu diễn cấp độ ký tự và cấp độ từ, đồng thời hỗ trợ xử lý nhiều từ thông dụng trong kho dữ liệu ngôn ngữ tự nhiên. Hiện thực nguyên bản BERT sử dụng từ vựng BPE cấp độ ký tự với kích thước 30K, được học sau khi xử lý trước đầu vào bằng quy tắc mã hóa heuristic (heuristic tokenization rule). Các nhà nghiên cứu của Facebook đã không áp dụng phương pháp này, nhưng đã cân nhắc sử dụng từ vựng BPE cấp byte lớn hơn để đào tạo BERT. Từ vựng này 50K đơn vị subword, mà không cần bất kỳ xử lý trước nào khác về mã hóa (tokenization) đầu vào nào.

Dưới đây tóm tắt hai phương pháp triển khai BPE:

* Dựa trên char-level: phương pháp BERT nguyên thủy, có dược bằng cách lấy văn bản đầu vào theo phương pháp heuristically (heuristically stemming).
* Dựa trên bytes-level: sự khác biệt với char-level là bytes-level sử dụng byte thay vì ký tự unicode làm đơn vị cơ bản của từ, việc này cải thiện mã hóa mà không cần sử dụng label UNKOWN.
* Khi BPE sử dụng bytes-level, kích thước từ vựng được tăng lên từ 30,000 (char-level của BERT nguyên thủy) lên 50,000. Điều này bổ sung thêm 15 triệu tham số cho BERT-base và 20 triệu tham số cho BERT-large.
* Các nghiên cứu trước đây đã chỉ ra rằng phương pháp tiếp cận này có thể dẫn đến sự giảm hiệu suất trong việc thực hiện các nhiệm vụ phụ thuộc. Tuy nhiên, tác giả của bài báo này tin rằng những lợi ích của việc học mã hóa này sẽ vượt trội hơn sự giảm hiệu suất, và tác giả sẽ tiếp tục so sánh các phương pháp mã hóa khác nhau trong các nhiệm vụ sau này.

Kết quả dưới đây thể kết quả cải tiến hơn so với BERT-large:

A screenshot of a computer

Description automatically generated

Kết quả GLUE:

A screenshot of a report

Description automatically generated

Kết quả SquaD:

A table with numbers and text

Description automatically generated

Kết quả RACE:

A black and white text on a white background

Description automatically generated

2.3 PhoBERT

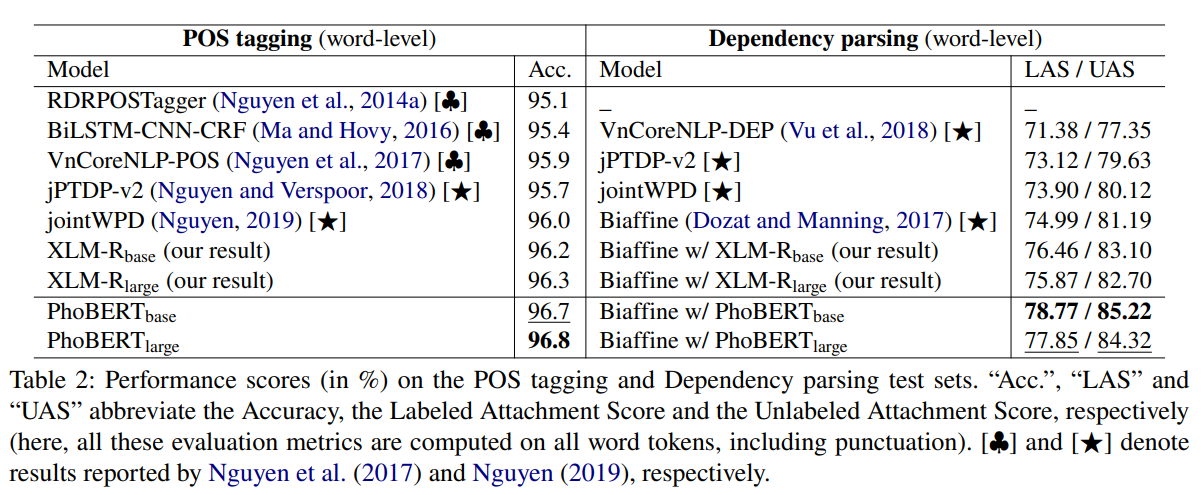
PhoBERT: Các mô hình ngôn ngữ được huấn luyện trước cho tiếng Việt Các mô hình PhoBERT được huấn luyện trước là những mô hình ngôn ngữ tiên tiến nhất cho tiếng Việt (Pho, tức là "Phở", là một món ăn phổ biến ở Việt Nam):

Hai phiên bản PhoBERT "base" và "large" là những mô hình ngôn ngữ đơn ngữ quy mô lớn đầu tiên được công khai và huấn luyện trước cho tiếng Việt. Phương pháp huấn luyện trước của PhoBERT dựa trên RoBERTa, tối ưu hóa quy trình huấn luyện trước của BERT để đạt được hiệu suất mạnh mẽ hơn.

PhoBERT vượt trội hơn so với các phương pháp đơn ngữ và đa ngữ trước đó, đạt được các thành tích tiên tiến nhất mới trên bốn tác vụ NLP tiếng Việt bao gồm: Gán nhãn từ loại (**Part-of-speech tagging - POS)**, Phân tích phụ thuộc (**Dependency parsing - Dep**), Nhận diện thực thể có tên (**Named-entity recognition – NER**) và Suy luận ngôn ngữ tự nhiên **(Natural language inference - NLI**). Kiến trúc tổng quát và kết quả thực nghiệm của PhoBERT có thể được tìm thấy trong bài báo: “**PhoBERT: Pre-trained language models for Vietnamese**” là của nhóm nghiên cứu của Nguyễn Tuấn Anh và Nguyễn Quốc Đạt tại Viện Nghiên cứu Dữ liệu lớn VinAI Research, Việt Nam.

Thông qua bài báo trên, có những điểm chú ý sau:

* Đây là một pre-trained được huấn luyện monolingual language, tức là chỉ huấn luyện dành riêng cho tiếng Việt. Việc huấn luyện dựa trên kiến trúc và **cách tiếp cận giống** **RoBERTa** của Facebook được Facebook giới thiệu giữa năm 2019.
* Tương tự như BERT, PhoBERT cũng có 2 phiên bản là "base" và "large" với 24 transformers block.
* PhoBERT được train trên khoảng 20GB dữ liệu bao gồm khoảng 1GB Vietnamese Wikipedia corpus và 19GB còn lại lấy từ Vietnamese news corpus. Đây là một lượng dữ liệu khả ổn để train một mô hình như BERT.
* PhoBERT sử dụng RDRSegmenter của VnCoreNLP để tách từ cho dữ liệu đầu vào trước khi qua BPE encoder.
* Như đã nói ở trên, do tiếp cận theo tư tưởng của RoBERTa, PhoBERT chỉ sử dụng task Masked Language Model để train, bỏ đi task Next Sentence Prediction.
* 05/2020 là BERTweet: [A pre-trained language model for English Tweets](https://github.com/VinAIResearch/BERTweet), cùng 2 tác giả PhoBert. Khá là giống với PhoBert chỉ khác là dữ liệu train thôi.
* Các thư viên hỗ trợ để triển khai PhoBert:
* fairseq: Là project của Facebook chuyên hỗ trợ các nghiên cứu và dự án liên quan đến model seq2seq. Và giúp đơn giản hóa quá trình tiền xử lý dữ liệu, định nghĩa mô hình, huấn luyện mô hình và triển khai mô hình
* fastBPE: Là package hỗ trợ tokenize từ (word) thành các từ phụ (subwords) theo phương pháp mới nhất được áp dụng cho các pretrain model NLP hiện đại như BERT và các biến thể của BERT.
* vncorenlp: Là một package NLP trong Tiếng Việt, hỗ trợ tokenize và các tác vụ NLP khác.
* transformers: Là một project của Huggingface hỗ trợ huấn luyện các model dựa trên kiến trúc transformer như BERT, GPT-2, RoBERTa, XLM, DistilBert, XLNet, T5, CTRL… phục vụ cho các tác vụ NLP trên cả nền tảng Pytorch và TensorFlow.
* Họ đã sử dụng kích thước lô (batch size) là 1024 trên 4 GPU V100, mỗi GPU có dung lượng 16GB, và tốc độ học tối đa (peak learning rate) là 0.0004 cho PhoBERT base. Đối với PhoBERT large, họ đã sử dụng kích thước lô là 512 và tốc độ học tối đa là 0.0002. Quá trình huấn luyện đã chạy trong 40 epoch, trong đó tốc độ học được điều chỉnh trong 2 epoch, tổng cộng khoảng 540K bước huấn luyện cho PhoBERTbase và 1.08M bước huấn luyện cho PhoBERT large. Chúng tôi đã huấn luyện trước PhoBERT base trong 3 tuần và trong 5 tuần.



A screenshot of a computer

Description automatically generated

2.4 Ưu và nhược điểm

|  |  |  |
| --- | --- | --- |
|  | Ưu điểm | Nhược điểm |
| BERT | Hiệu suất cao trên nhiều tác vụ NLP nhờ kiến trúc Transformer hai chiều, có khả năng xử lý tốt ngữ cảnh.  Được cộng đồng sử dụng rộng rãi, nhiều tài nguyên và công cụ hỗ trợ.  Có nhiều phiên bản tiền huấn luyện có sẵn, dễ dàng fine-tune cho các tác vụ cụ thể. | Yêu cầu tài nguyên tính toán lớn, thời gian huấn luyện dài.  Hiệu suất có thể không tốt bằng các mô hình mới hơn như RoBERTa trong một số tác vụ.  Có thể gặp khó khăn với ngữ cảnh dài và ngữ cảnh phức tạp. |
| RoBERTa | Cải tiến từ BERT với việc huấn luyện lâu hơn và trên lượng dữ liệu lớn hơn, không sử dụng nhiệm vụ NSP (Next Sentence Prediction), tối ưu hóa quá trình huấn luyện.  Hiệu suất vượt trội trên các tác vụ NLP, đặc biệt là khi so sánh với BERT.  Khả năng xử lý ngữ cảnh tốt hơn nhờ vào các cải tiến trong quy trình huấn luyện.  Được train với data lớn hơn nên mô hình có khả năng tổng quát tốt hơn và đạt được kết quả cao. | Yêu cầu tài nguyên tính toán lớn hơn BERT, thời gian huấn luyện dài hơn.  Sử dụng lượng dữ liệu và thời gian huấn luyện lớn, không phù hợp với những hệ thống có tài nguyên hạn chế.  Không sử dụng nhiệm vụ NSP, có thể là một nhược điểm trong một số tác vụ yêu cầu hiểu biết về quan hệ câu. |
| PhoBERT | Được huấn luyện đặc biệt cho tiếng Việt, đạt hiệu suất tốt hơn các mô hình đa ngữ như XLM-R trên các tác vụ NLP tiếng Việt.  Công khai phát hành mô hình giúp thúc đẩy nghiên cứu và ứng dụng trong lĩnh vực NLP tiếng Việt.  Được tối ưu hóa cho tiếng Việt, phù hợp với các yêu cầu ngôn ngữ đặc thù. | Cần nhiều tài nguyên tính toán và thời gian huấn luyện dài, đặc biệt là với mô hình PhoBERTlarge.  Chỉ tối ưu cho tiếng Việt, không sử dụng tốt cho các ngôn ngữ khác.  Chưa được ứng dụng rộng rãi, ít tài nguyên và công cụ hỗ trợ hơn so với BERT và RoBERTa. |

TÀI LIỆU THAM KHẢO

Vaswani, A., Shazeer, N., Parmar, N., & Jones, J. (2017). Attention is All You Need. arXiv. preprint **arXiv:1706.03762**.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT Pre-training of Deep Bidirectional Transformersfor Language Understanding. arXiv preprint ***arXiv:1810.04805****.*

Dat Q. Nguyen, Anh T. Nguyen (2020). PhoBERT: Pre-trained language models for Vietnamese. arXiv preprint **arXiv: 2003.00744**

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2019) RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint, ***arXiv:1907.11692***

Top of Form

**PHỤ LỤC**

Phần này bao gồm những nội dung cần thiết nhằm minh họa hoặc hỗ trợ cho nội dung luận văn như số liệu, biểu mẫu, tranh ảnh. . . . nếu sử dụng những câu trả lời cho một *bảng câu hỏi thì bảng câu hỏi mẫu này phải được đưa vào phần Phụ lục ở dạng nguyên bản* đã dùng để điều tra, thăm dò ý kiến; **không được tóm tắt hoặc sửa đổi**. Các tính toán mẫu trình bày tóm tắt trong các biểu mẫu cũng cần nêu trong Phụ lục của luận văn. Phụ lục không được dày hơn phần chính của luận văn