TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

Thành viên:
21120209 - Phạm Bách Chiến
21120278 - Phùng Đoàn Khôi
21120369 - Nguyễn Minh Vũ

BÁO CÁO ĐỒ ÁN: ỨNG DỤNG MÔ HÌNH NGÔN NGỮ LỚN TRONG BÀI TOÁN XÁC ĐỊNH TƯƠNG ĐỒNG VĂN BẢN TIẾNG VIỆT

CSC15006 – Nhập môn xử lý ngôn ngữ tự nhiên

Tp. Hồ Chí Minh, tháng 12/2023

Mục lục

1 Giới thiệu

Tương đồng ngữ nghĩa là bài toán đo lường độ tương đồng về mặt ngữ nghĩa giữa hai đoạn văn bản cho trước. Xác định độ tương đồng giữa hai đoạn văn bản có tác dụng quan trọng trong nhiều tác vụ khác của lĩnh vực NLP. Mức độ tương đồng có thể đi từ nhiều cấp độ, từ từng từ, đến cụm từ, một câu và tới một đoạn văn bản hoặc bài diễn văn dài.

1.1 Phát biểu bài toán

Bài toán đầu vào của ta sẽ là cặp câu. Đầu ra bài toán là một mức điểm thể hiện cho mức độ tương đồng của cặp câu đầu vào dựa trên thang đo độ tương đồng do ta đặt ra.

1.2 Thang đo tương đồng

Mức độ đánh giá được đánh điểm từ 0 tới 5, trong đó:

- Điểm 5: Hai câu giống nhau hoàn toàn về ngữ nghĩa
- Điểm 4: Gần như giống nhau, khác biệt ở trạng ngữ hoặc những thông tin bổ sung không quá quan trọng.
- Điểm 3: Có điểm giống nhau giữa hai câu. Tuy nhiên, có khác biệt ở những thông tin quan trọng, hoặc chủ ngữ, hoặc vị ngữ hay những thông tin bổ sung vào.
- Điểm 2: Tồn tại vài sự liên quan giữa hai câu. Tuy nhiên, có nhiều sự khác biệt trong thành phần câu, ở chủ ngữ, vị ngữ hoặc những thông tin bổ sung khác.
- Điểm 1: Khác nhau về mặt ngữ nghĩa, có thể liên quan về một chủ đề hoặc một hoạt động.
- Điểm 0: Không tương đồng ngữ nghĩa.

1.3 Ứng dụng

Phân tích tương đồng ngữ nghĩa có thể áp dụng trong nhiều ứng dụng khác nhau:

- Truy xuất thông tin (Information Retrieval): Cải thiện khả năng của hệ thống tìm kiếm thông tin, không chỉ dựa trên từ khóa (keywords) mà còn trên sự tương đồng ngữ nghĩa giữa các đoạn văn bản.
- Dịch máy (Machine translation): Hỗ trợ trong việc hiểu và diễn đạt ý nghĩa câu văn tốt hơn.
- Tóm tắt văn bản (Text Summary): Rút gọn văn bản bằng cách lấy những câu quan trọng và sử dụng những câu có sự tương đồng với văn bản gốc.
- Ngoài ra, phân tích tương đồng ngữ nghĩa còn có thể ứng dụng trong những công việc như: kiểm tra đạo văn, phân loại văn bản, phân tích ý kiến và đánh giá...

1.4 Dataset

1.4.1 Vi-STS

Bộ data này được nhóm tạo ra bằng cách sử dụng bộ data STS Benchmark khá phổ biến. Nhóm sử dụng model VietAI-envit5-translation [8] để dịch sang tiếng Việt.

Sentence 1	Sentence 2	Label
Một người phụ nữ cắt bông cải xanh .	Một người phụ nữ đang xắt bông cải xanh với một con dao .	4.25
Một cậu bé đang hát và chơi piano	Một cậu bé chơi piano	3
Một người đàn ông đang cột dây giày .	Một người đàn ông cột dây giày.	5
Một người đang bóc một củ hành tây.	Một người bóc vỏ cà chua.	2

Bảng 1: Một vài dòng dữ liệu của bộ data Vi-STS

1.4.2 vnPara

VnPara [1] là bộ ngữ liệu về hai đoạn văn bản có được diễn giải lại (paraphrase) hay không. Bộ data bao gồm 3083 cặp câu và được đánh nhãn 0: không tương đồng (non-paraphrase) và 1: tương đồng (paraphrase).

Sentence 1	Sentence 2	Label
Song có thể nói ngay rằng cuộc “ mổ xẻ ” tìm nguyên nhân thất bại của bóng đá chuyên nghiệp Việt Nam chưa đi đến nơi đến chốn .	Cả nền bóng đá Việt Nam đang khủng hoảng và đáng nói hơn khi cuộc khủng hoảng này vẫn chưa có dấu hiệu " chạm đáy " .	0
Trong các nước Đông Nam Á Việt Nam chỉ đứng sau Singapore Malaysia Brunei .	Mức độ triển khai chính phủ điện tử của Việt Nam đã vươn lên đứng thứ 4 trong các quốc gia khu vực Đông Nam Á .	0
Hà Nội và Thành phố Hồ Chí Minh đã lọt vào danh sách 10 thành phố mới nổi về gia công phần mềm .	Đáng chú ý hai thành phố Hà Nội và Tp.HCM đã lọt vào danh sách 10 thành phố mới nổi về gia công phần mềm .	1

Bảng 2: Một vài dòng dữ liệu của bộ data vnPara

1.4.3 VNPC

VNPC [2] là bộ data có cấu trúc giống như vnPara. Bộ data VNPC bao gồm 3134 cặp câu, trong đó gồm 2748 cặp được đánh là tương đồng (nhãn 1) và 386 cặp câu không tương đồng (nhãn 0).

Sentence 1	Sentence 2	Label
Trường hợp có tế bào ung thư , bệnh nhi sẽ được điều trị tiếp theo .	Nếu có tế bào ung thư , bệnh nhi sẽ có hướng điều trị tiếp theo .	1
Từ khi lập nước , Cụ Hồ đã thực hiện , nhưng sau này , cách dùng người sáng suốt đó dần dần bị lãng quên , nên thông điệp " tạo cơ hội cho con cháu nông dân , công nhân , người nghèo " của Thủ tướng vẫn là rất mới .	Bài phát biểu tại Quốc hội sau khi tái đắc cử , Thủ tướng đã đưa ra một thông điệp đáng chú ý là phải làm sao để con cháu của nông dân , công nhân , người nghèo đều có cơ hội học tập , tiến thân , kể cả cơ hội trở thành lãnh đạo của đất nước trong tương lai .	0
Trong đó , có 916 công trình đã đưa vào hoạt động , 151 công trình đang thi công , 8 công trình đang tạm dừng hoạt động .	Trong đó , 916 công trình đã đưa vào hoạt động , 151 công trình đang thi công , 8 công trình đang tạm dừng hoạt động .	1

Bảng 3: Một vài dòng dữ liệu của bộ data VNPC

2 Mô hình Pre-trained

2.1 BERT

2.1.1 Mô hình Bert

Mô hình BERT, viết tắt của Bidirectional Encoder Representations from Transformers, là một mô hình ngôn ngữ "encoder-only" dựa trên kiến trúc transformer. Trong đó, BERT đã áp dụng một kỹ thuật mới có tên Masked Language Model (MLM) cho phép huấn luyện hai chiều trong các mô hình mà trước đây không thể.[3]

2.1.2 Mô hình RoBERTa

Một kiểu cải tiến của mô hình Bert, điều chỉnh siêu tham số và kích thước tập huấn luyện, bao gồm: huấn luyện mô hình lâu hơn và sử dụng nhiều dữ liệu hơn; loại bỏ việc dự đoán câu kế tiếp; huấn luyện tuần tự lâu hơn; thay đổi mask trên tập dữ liệu huấn luyện.[5]

2.1.3 Mô hình PhoBert

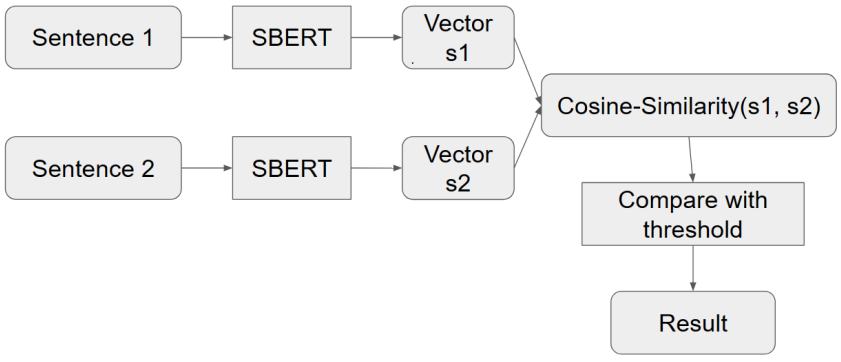
Mô hình đơn ngữ được pre-train dành cho tiếng Việt và phát triển bởi người Việt dựa trên mô hình RoBERTa. Mô hình này được huấn luyện trên 20GB cấp độ từ trong bộ ngữ liệu của tiếng Việt.[6]

2.1.4 Mô hình SBert

SBert hay Sentence-Bert là một mô hình đã được tinh chỉnh của BERT, sử dụng kiến trúc mạng siamese và triplet để lấy embedding ngữ nghĩa của một câu và dùng nó để so sánh với nhau thông qua tính toán cosine-similarity.[4]

2.2 Framework

Vì vấn đề của bài toán là so sánh tính tương đồng giữa 2 câu với nhau nên sử dụng mô hình SBert sẽ phù hợp nhất. Minh chứng cho sự lựa chọn này là để tìm được cặp câu tương đồng nhất trong bộ dữ liệu, BERT sẽ mất trung binh 60 giờ trong khi SBert chỉ mất 5 giây mà vẫn giữ được độ chính xác như BERT.[4]



Framework của mô hình SBert

2.3 Fine-tuning

Việc fine-tune mô hình SBert được thực hiện trên Google Colab T4 GPU cùng với 16GB VRAM. Nhóm áp dụng trên bộ dữ liệu Vi-STS.
Để tính toán độ hiệu quả của fine-tuning, nhóm sử dụng hàm loss với:

$$\text{Loss} = |\text{labels output} - \text{cosine\_similarity(sentence 1, sentence 2)}|$$

3 Mô hình ngôn ngữ lớn

3.1 Các mô hình sử dụng

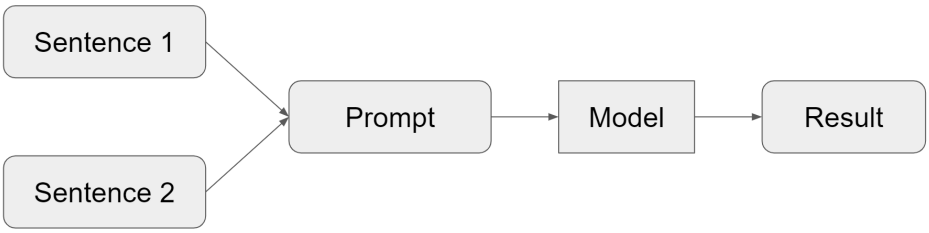
3.1.1 Mistral 7B

Mistral 7B [10] là mô hình ngôn ngữ lớn được tạo bởi Mistral AI, ra mắt vào tháng 10 năm 2023. Tuy chỉ có 7.3 tỉ tham số nhưng Mistral AI có hiệu suất cao hơn mô hình LLama 2 13B [12] trong tất cả benchmark và LLama 1 34B [13] trong nhiều benchmark.

3.1.2 PhoGPT 7B5

PhoGPT [11] là mô hình ngôn ngữ lớn dành riêng cho tiếng Việt, được phát triển bởi VinAI. PhoGPT có 7.5 tỉ tham số và có hiệu suất cao hơn trong nhiều benchmark so với các mô hình ngôn ngữ lớn tiếng Việt trước đây.

3.2 Framework



3.3 Prompting

Ở trong bài toán này, mô hình PhoGPT sử dụng prompt tiếng Việt, còn Mistral sử dụng cả prompt tiếng Việt và Anh.

3.3.1 Prompt tiếng Việt

Hai câu sau có tương đồng về mặt ngữ nghĩa không?
Câu 1:
Câu 2:
Trả lời "Có" hoặc "Không"

3.3.2 Prompt tiếng Anh

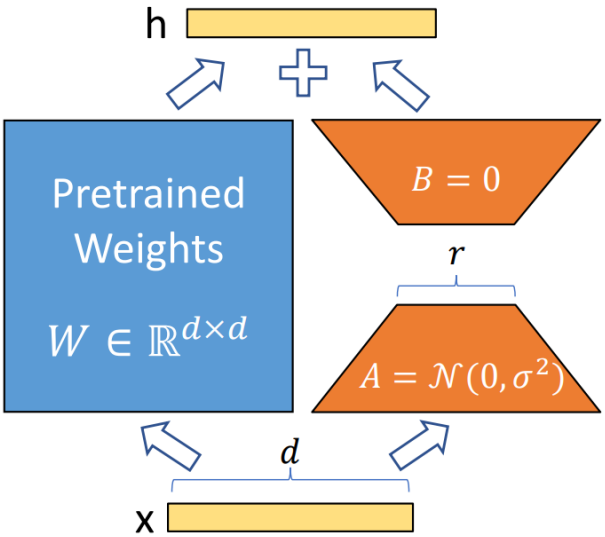
Do these two sentences have the same meaning?
Sentence 1:
Sentence 2:
Answer with "Yes" or "No".

3.4 Fine-tuning

Cả hai mô hình đều được train trên bộ dataset vnPara bằng kĩ thuật QLoRa với cấu hình GPU P100 16GB VRAM trên Kaggle.

3.4.1 LoRa

LoRa (Low-Rank adaptation of LLMs) [7] là kĩ thuật cost-effective fine-tuning kết hợp giữa việc sử dụng adapter và giảm số chiều ma trận.



Thay vì tính toán và cập nhật ma trận trọng số W' thì ta sẽ tách W' thành $W + \Delta W$, trong đó W là ma trận đã được pre-train của model. Những ma trận trọng số của LLM thường có intrinsic rank thấp nên ΔW có thể được xấp xỉ bằng tích của hai ma trận $A \times B$ có rank thấp hơn nhiều so với ma trận ΔW. Và vì A và B có rank thấp nên tổng lượng parameter sẽ ít hơn nhiều so với ma trận ΔW, điều này giúp việc tính toán và lưu trữ được cải thiện rất nhiều.

3.4.2 QLoRa

QLoRa (Quantized LoRa) [9] là kĩ thuật fine-tuning kết hợp giữa việc sử dụng quantization và LoRa. Quantization là cách biểu diễn thông tin dưới dạng kiểu dữ liệu khác với số lượng thông tin ít hơn, và ở kĩ thuật này sẽ là biểu diễn số thực. Thay vì lưu trữ và tính toán các parameter dưới dạng 32 bit hoặc 16 bit thì quantization sẽ giúp ta giảm xuống còn 8, thậm chí là 4 bit với kiểu dữ liệu NF4 (NormalFloat4). Bằng việc kết hợp quantization và LoRa thì ta hoàn toàn có thể huấn luyện các mô hình 7 tỉ tham số với chỉ 16GB VRAM.

4 Kết quả thử nghiệm

4.1 Vi-STS

Trên mô hình pre-trained:

Model	Pearson Correlation
SBERT-PhoBERT-base	0.4913
pm-MiniLM-L12-v2	0.7959
pm-mpnet-base-v2	0.822
MiniLM-L6-v2	0.4907
SBERT-PhoBERT-base*	0.8005
pm-MiniLM-L12-v2*	0.8123
pm-mpnet-base-v2*	0.8253
MiniLM-L6-v2*	0.6979

Bảng 4: Kết quả của mô hình pre-trained trên tập Vi-STS

4.2 vnPara

4.2.1 Mô hình pre-trained

Model	Accuracy(%)	F1-score(%)
SBERT-PhoBERT-base	93.56	93.48
pm-MiniLM-L12-v2	94.55	94.52
pm-mpnet-base-v2	95.12	95.04
MiniLM-L6-v2	87	87.5
SBERT-PhoBERT-base*	95.91	95.84
pm-MiniLM-L12-v2*	95.34	95.36
pm-mpnet-base-v2*	95.69	95.65
MiniLM-L6-v2*	90.25	90.33

Bảng 5: Kết quả của mô hình pre-trained trên tập vnPara

4.2.2 Mô hình ngôn ngữ lớn

Model	Accuracy(%)	F1-score(%)
Phogpt	50.88	64.12
mistral-7b-instruct-v2(en)	82	80.5
mistral-7b-instruct-v2(vi)	74.76	78.89
Phogpt*	51.12	65.37
mistral-7b-instruct-v2(en)*	83.5	884.4
mistral-7b-instruct-v2(vi)*	73.31	78.04

Bảng 6: Kết quả của mô hình ngôn ngữ lớn trên tập vnPara

4.3 VNPC

4.3.1 Mô hình pre-trained

Model	Accuracy(%)	F1-score(%)
SBERT-PhoBERT-base	85.64	92.22
pm-MiniLM-L12-v2	70.49	81.33
pm-mpnet-base-v2	75.59	85.1
MiniLM-L6-v2	78.47	87.6
SBERT-PhoBERT-base*	71.93	81.96
pm-MiniLM-L12-v2*	67.78	78.42
pm-mpnet-base-v2*	66.98	77.66
MiniLM-L6-v2*	56.77	68.08

Bảng 7: Kết quả của mô hình pre-trained trên tập VNPC

4.3.2 Mô hình ngôn ngữ lớn

Model	Accuracy(%)	F1-score(%)
Phogpt	77.83	87.42
mistral-7b-instruct-v2(en)	79.11	87.49
mistral-7b-instruct-v2(vi)	84.68	91.38
Phogpt*	80.22	88.89
mistral-7b-instruct-v2(en)*	79.43	87.65
mistral-7b-instruct-v2(vi)*	84.53	91.23

Bảng 8: Kết quả của mô hình ngôn ngữ lớn trên tập VNPC

5 Kết luận và hướng phát triển

Những mô hình ngôn ngữ lớn có tiềm năng lớn trong việc giải bài toán xác định tương đồng văn bản tiếng Việt nói riêng và tương đồng văn bản nói chung.
Tuy chỉ sử dụng những mô hình với số lượng tham số thấp nhưng kết quả đạt được khá tốt.
Nhược điểm tồn đọng chính là độ chính xác chưa cao như các mô hình pre-trained SBERT, và thời gian chạy cao.
Do đó hướng phát triển của nhóm trong đề tài này sẽ là cải thiện những mô hình ngôn ngữ lớn để đạt hiệu suất cao hơn.

References

[1] Ngo Xuan Bach et al. "Paraphrase identification in Vietnamese documents". In: *2015 Seventh International Conference on Knowledge and Systems Engineering (KSE)*. IEEE. 2015, pp. 174–179.

[2] Hoang-Quoc Nguyen-Son et al. "Vietnamese Paraphrase Identification Using Matching Duplicate Phrases and Similar Words". In: *Future Data and Security Engineering: 5th International Conference, FDSE 2018, Ho Chi Minh City, Vietnam, November 28–30, 2018, Proceedings 5*. Springer. 2018, pp. 172–182.

[3] Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *aclanthology.org* (2018).

[4] Nils Reimers and Iryna Gurevych. "Sentence-bert: Sentence embeddings using siamese bert-networks". In: *arXiv preprint arXiv:1908.10084* (2019).

[5] Yinhan Liu Myle Ott NamanGoyal Jingfei Du Mandar Joshi† Danqi Chen OmerLevy MikeLewis Luke Zettlemoyer† Veselin Stoyanov. "RoBERTa: ARobustly Optimized BERT Pretraining Approach". In: *arXiv preprint arXiv:1907.11692* (2019).

[6] Dat Quoc Nguyen and Anh Tuan Nguyen. "PhoBERT: Pre-trained language models for Vietnamese". In: *arXiv preprint arXiv:2003.00744* (2020).

[7] Edward J Hu et al. "Lora: Low-rank adaptation of large language models". In: *arXiv preprint arXiv:2106.09685* (2021).

[8] Chinh Ngo et al. *MTet: Multi-domain Translation for English and Vietnamese*. 2022. DOI: 10.48550/ARXIV.2210.05610. URL: https://doi.org/10.48550/arxiv.2210.05610.

[9] Tim Dettmers et al. "Qlora: Efficient finetuning of quantized llms". In: *arXiv preprint arXiv:2305.14314* (2023).

[10] Albert Q Jiang et al. "Mistral 7B". In: *arXiv preprint arXiv:2310.06825* (2023).

[11] Dat Quoc Nguyen et al. "PhoGPT: Generative Pre-training for Vietnamese". In: *arXiv preprint arXiv:2311.02945* (2023).

[12] Hugo Touvron et al. "Llama 2: Open foundation and fine-tuned chat models". In: *arXiv preprint arXiv:2307.09288* (2023).

[13] Hugo Touvron et al. "Llama: Open and efficient foundation language models". In: *arXiv preprint arXiv:2302.13971* (2023).