1. Bìa báo cáo
2. Lời cảm ơn
3. Mục lục
4. Tóm tắt
5. Danh sách các từ viết tắt và ký hiệu
6. Danh sách hình vẽ, biểu đồ, bảng

MỤC LỤC

Chương 1: Giới thiệu

1.1. Đặt vấn đề

1.2. Mục đích nghiên cứu

1.3. Ý nghĩa nghiên cứu

1.4. Phạm vi nghiên cứu

1.4. Phạm vi nghiên cứu

1.5. Phương pháp nghiên cứu

Chương 2: Cơ sở lý thuyết và phương pháp nghiên cứu

2.1. Ngôn ngữ tự nhiên và xử lý ngôn ngữ tự nhiên (NLP)

2.2. Mô hình ngôn ngữ lớn (LLM)

2.3. Các kỹ thuật gợi ý trong LLM

2.4. Các mô hình LLM nổi tiếng và ứng dụng của chúng

2.5. Phương pháp nghiên cứu và đánh giá

Chương 3: Xây dựng mô hình LLM với ngôn ngữ đích là Tiếng Việt

3.1. Thu thập và xử lý dữ liệu huấn luyện

3.2. Xây dựng và huấn luyện mô hình LLM

3.3. Đánh giá và tinh chỉnh mô hình

Chương 4: Ứng dụng kỹ thuật gợi ý trong mô hình LLM

4.1. Phân tích các kỹ thuật gợi ý khác nhau

4.2. Áp dụng kỹ thuật gợi ý cho mô hình LLM Tiếng Việt

4.3. Đánh giá hiệu quả của các kỹ thuật gợi ý

Chương 5: So sánh và đánh giá hiệu quả của các kỹ thuật gợi ý

Chương 4: Đánh giá hiệu quả của mô hình LLM Tiếng Việt

4.1. Chỉ số đánh giá mô hình

* Giới thiệu các chỉ số đánh giá phổ biến cho mô hình LLM, như Perplexity, BLEU, ROUGE, METEOR, v.v.
* Giải thích ý nghĩa và cách tính của từng chỉ số.

4.2. Phương pháp đánh giá

* Trình bày cách thiết lập thí nghiệm đánh giá, bao gồm việc chọn tập kiểm tra, thiết lập môi trường, và lựa chọn các chỉ số đánh giá phù hợp.

Đề cập đến việc sử dụng đánh giá định 5.1. Tóm tắt kết quả nghiên cứu

* Tổng kết quá trình xây dựng, huấn luyện, tinh chỉnh và đánh giá mô hình LLM Tiếng Việt.
* Nhấn mạnh những đóng góp chính của nghiên cứu này cho lĩnh vực xử lý ngôn ngữ tự nhiên và ứng dụng của LLM trong tiếng Việt.

5.2. Hạn chế và thách thức

* Đề cập lại các hạn chế và thách thức đã thảo luận trong chương 3, cùng với những khó khăn gặp phải trong quá trình nghiên cứu và phát triển mô hình.

Thảo luận về những rủi ro và vấn đề v 5.3. Định hướng nghiên cứu trong tương lai

* Đề xuất các hướng nghiên cứu trong tương lai để cải tiến mô hình LLM Tiếng Việt, bao gồm việc tăng cường dữ liệu huấn luyện, cải tiến kiến trúc mô hình, áp dụng các kỹ thuật huấn luyện tiên tiến hơn, v.v.
* Khảo sát các phương pháp mới trong xử lý ngôn ngữ tự nhiên để giải quyết các hạn chế và thách thức của mô hình hiện tại, cũng như khám phá các ứng dụng tiềm năng khác.
* Nghiên cứu về đạo đức và tính công bằng trong mô hình LLM Tiếng Việt, bao gồm việc đánh giá mức độ đạo đức và công bằng của mô hình, cũng như đề xuất các phương pháp cải thiện trong lĩnh vực này.

5.4. Ứng dụng và triển khai mô hình LLM Tiếng Việt trong thực tế

* Tìm kiếm các đối tác, tổ chức và doanh nghiệp để hợp tác và triển khai mô hình LLM Tiếng Việt trong các ứng dụng thực tế, như chatbot, hệ thống khuyến nghị nội dung, dịch thuật đa ngôn ngữ, v.v.
* Xây dựng các giao diện và API dễ sử dụng để giúp các nhà phát triển và người dùng cuối tiếp cận và ứng dụng mô hình một cách hiệu quả.
* Tổ chức các buổi hội thảo, đào tạo và hỗ trợ kỹ thuật cho cộng đồng nhằm nâng cao nhận thức và khả năng ứng dụng mô hình LLM Tiếng Việt trong thực tế.

Kết luận: Chương 5 tổng kết lại kết quả, hạn chế và đóng góp của nghiên cứu trong việc xây dựng mô hình LLM Tiếng Việt. Đồng thời, chương này cũng đề xuất các hướng phát triển trong tương lai, từ việc cải tiến mô hình, nghiên cứu về đạo đức và công bằng, đến việc ứng dụng và triển khai mô hình trong thực tế. Những nỗ lực này sẽ giúp mô hình LLM

2 / 2

Top of Form

Bottom of Form

* lượng (quantitative) và định tính (qualitative) để đánh giá mô hình từ nhiều khía cạnh.

4.3. Kết quả và so sánh với các mô hình khác

* Báo cáo kết quả đánh giá của mô hình LLM Tiếng Việt trên tập kiểm tra, bao gồm các chỉ số đánh giá và ví dụ minh họa.
* So sánh kết quả với các mô hình LLM khác (ví dụ: mô hình tiếng Anh hoặc đa ngôn ngữ) để đánh giá sự hiệu quả và tiềm năng của mô hình Tiếng Việt.

Chương 5: Kết luận và hướng phát triển

5.1. Tóm tắt kết quả nghiên cứu

Chương 6: Ứng dụng và tiềm năng của kỹ thuật gợi ý trong tương lai

6.1. Các ứng dụng thực tế của kỹ thuật gợi ý trong mô hình LLM

6.1.1. Trợ lý ảo

6.1.2. Dịch máy tự động

6.1.3. Tạo nội dung cho bài viết, văn bản, tài liệu

6.2. Thách thức và hạn chế của kỹ thuật gợi ý

6.3. Tiềm năng phát triển trong tương lai

Chương 7: Kết luận và kiến nghị

7.1. Kết luận chung

7.2. Đóng góp của đề tài

7.3. Hướng phát triển và nghiên cứu tiếp theo

PHỤ LỤC

* Mã nguồn, tập dữ liệu và các tài liệu liên quan
* Các kết quả đánh giá chi tiết
* Các hình ảnh, biểu đồ, bảng liên quan

Chương 1: Giới thiệu

1.1. Đặt vấn đề

Trong những năm gần đây, ngành công nghiệp trí tuệ nhân tạo (AI) đã chứng kiến sự bùng nổ của các mô hình ngôn ngữ lớn (LLM), nhờ vào sự phát triển của các kỹ thuật học sâu và khối lượng dữ liệu đào tạo ngày càng lớn. Các LLM đã mang lại nhiều ứng dụng thực tế trong xử lý ngôn ngữ tự nhiên (NLP), bao gồm trợ lý ảo, dịch máy tự động và tạo nội dung cho bài viết, văn bản, tài liệu. Tuy nhiên, việc cải thiện khả năng dự đoán và đề xuất các từ hoặc câu tiếp theo trong quá trình sử dụng mô hình vẫn là một thách thức lớn. Kỹ thuật gợi ý có tiềm năng giải quyết vấn đề này bằng cách giúp mô hình đưa ra các từ hoặc câu tiếp theo một cách chính xác và tự nhiên hơn.

1.2. Mục đích nghiên cứu

Mục đích của nghiên cứu này là xây dựng mô hình LLM với ngôn ngữ đích là Tiếng Việt và dữ liệu huấn luyện lấy từ các nguồn khác nhau, áp dụng các kỹ thuật gợi ý để cải thiện khả năng dự đoán và đề xuất các từ hoặc câu tiếp theo trong quá trình sử dụng mô hình. Nghiên cứu cũng nhằm so sánh và đánh giá hiệu quả của các kỹ thuật gợi ý được áp dụng.

1.3. Ý nghĩa thực tiễn

Kết quả của nghiên cứu này có ý nghĩa thực tiễn đối với việc phát triển và ứng dụng mô hình ngôn ngữ lớn cho tiếng Việt, giúp cải thiện khả năng xử lý ngôn ngữ tự nhiên của các ứng dụng và hệ thống, tiết kiệm thời gian và nguồn lực cho người dùng.

1.4. Phạm vi nghiên cứu

Phạm vi của nghiên cứu bao gồm việc xây dựng mô hình LLM cho tiếng Việt, áp dụng các kỹ thuật gợi ý và đánh giá hiệu quả của chúng

1.5. Phương pháp nghiên cứu

Để đạt được mục tiêu của nghiên cứu, chúng tôi sử dụng các phương pháp sau:

• Nghiên cứu và tổng hợp kiến thức từ các công trình nghiên cứu, bài báo khoa học và tài liệu liên quan đến mô hình ngôn ngữ lớn và kỹ thuật gợi ý.

• Thu thập dữ liệu huấn luyện cho mô hình LLM từ các nguồn khác nhau, bao gồm Wikipedia, báo chí, văn bản chính thức, v.v.

• Xây dựng mô hình LLM cho tiếng Việt bằng cách sử dụng các công cụ và thư viện mã nguồn mở.

• Áp dụng các kỹ thuật gợi ý vào mô hình LLM và đánh giá hiệu quả của chúng thông qua các phép đo định lượng và định tính.

• So sánh và đánh giá hiệu quả của các kỹ thuật gợi ý được áp dụng trong nghiên cứu.

Chương 2: Cơ sở lý thuyết và phương pháp nghiên cứu

2.1. Ngôn ngữ tự nhiên và xử lý ngôn ngữ tự nhiên (NLP)

* Giới thiệu về ngôn ngữ tự nhiên và vai trò của ngôn ngữ trong giao tiếp giữa con người.

Ngôn ngữ tự nhiên là hệ thống giao tiếp phức tạp được sử dụng bởi con người để truyền đạt ý tưởng, thông tin và cảm xúc. Ngôn ngữ đóng vai trò quan trọng trong giao tiếp giữa con người, giúp chúng ta hiểu nhau và hợp tác với nhau. Ngôn ngữ tự nhiên không chỉ bao gồm từ ngữ, mà còn có ngữ pháp, ngữ điệu và ngữ nghĩa, tất cả đều tạo nên sự phong phú và đa dạng của ngôn ngữ.

* Trình bày khái niệm xử lý ngôn ngữ tự nhiên (NLP) và các bài toán cơ bản trong NLP (như phân loại văn bản, phân tích ý kiến, dịch máy, trả lời câu hỏi, v.v.).

Xử lý ngôn ngữ tự nhiên (NLP) là một nhánh của khoa học máy tính, trí tuệ nhân tạo và ngôn ngữ học, nhằm giúp máy tính hiểu, phân tích và tạo ra ngôn ngữ tự nhiên của con người. NLP giúp chúng ta xây dựng các hệ thống thông minh có khả năng giao tiếp với con người bằng ngôn ngữ tự nhiên, như trợ lý ảo, dịch máy tự động, v.v.

Các bài toán cơ bản trong NLP bao gồm:

1. Phân loại văn bản: Phân loại các đoạn văn bản vào các nhóm có nội dung tương tự, ví dụ phân loại email thành thư rác hoặc thư hữu ích.
2. Phân tích ý kiến: Xác định ý kiến, cảm xúc của người dùng trong các bình luận, đánh giá sản phẩm hoặc dịch vụ.
3. Dịch máy: Dịch ngôn ngữ từ ngôn ngữ này sang ngôn ngữ khác, ví dụ từ tiếng Anh sang tiếng Việt.
4. Trả lời câu hỏi: Tìm câu trả lời cho các câu hỏi của người dùng dựa trên nguồn thông tin như văn bản, dữ liệu, v.v.

* Giới thiệu về các kỹ thuật và phương pháp xử lý ngôn ngữ tự nhiên, bao gồm cả kỹ thuật học máy truyền thống và học sâu.

NLP sử dụng các kỹ thuật và phương pháp đa dạng, bao gồm cả kỹ thuật học máy truyền thống và học sâu:

Học máy truyền thống: Các kỹ thuật học máy như hồi quy tuyến tính, máy vectơ hỗ trợ (SVM), cây quyết định, v.v., được sử dụng để xây dựng các mô hình dựa trên các đặc trưng được trích xuất từ văn bản. Các đặc trưng này có thể bao gồm tần số xuất hiện của từ, cụm từ, cấu trúc ngữ pháp, v.v. Tuy nhiên, học máy truyền thống có nhược điểm là không thể mô hình hóa được các mối quan hệ phức tạp và ngữ nghĩa giữa các từ trong văn bản.

1. Học sâu: Các mô hình học sâu như mạng nơ-ron tích chập (CNN), mạng nơ-ron hồi quy (RNN), và mạng nơ-ron Transformer đã mang lại những cải tiến đột phá trong NLP. Học sâu cho phép máy tính tự động học được các đặc trưng từ dữ liệu, mô hình hóa các mối quan hệ phức tạp giữa các từ và ngữ nghĩa của chúng. Mô hình BERT, GPT và các biến thể của chúng là những ví dụ nổi tiếng về sự thành công của học sâu trong NLP.

Một số phương pháp xử lý ngôn ngữ tự nhiên phổ biến bao gồm:

1. Tokenization: Chia văn bản thành các đơn vị nhỏ hơn như từ, cụm từ hoặc câu.
2. Loại bỏ từ dừng (stop words): Loại bỏ các từ phổ biến nhưng không mang nhiều thông tin, ví dụ "và", "là", "của", v.v.
3. Chuẩn hóa từ (stemming/lemmatization): Chuyển các từ về dạng gốc của chúng, ví dụ từ "đi" sẽ được chuyển về "đi" (lemmatization) hoặc "đ" (stemming).
4. Mã hóa one-hot và TF-IDF: Chuyển đổi văn bản thành các vectơ số để xử lý bởi các mô hình học máy.
5. Word embeddings: Biểu diễn từ dưới dạng các vectơ nhiều chiều, giúp mô hình hóa được mối quan hệ ngữ nghĩa giữa các từ.

Những tiến bộ trong NLP đã mở ra nhiều ứng dụng hấp dẫn, từ trợ lý ảo, dịch máy tự động, tới phân tích ý kiến và tạo nội dung tự động. Trong phần tiếp theo, chúng ta sẽ tìm hiểu về mô hình ngôn ngữ lớn (LLM) và các kỹ thuật gợi ý được sử dụng trong LLM.

2.2. Mô hình ngôn ngữ lớn (LLM)

* Định nghĩa và giới thiệu về mô hình ngôn ngữ lớn (LLM).

Mô hình ngôn ngữ lớn (Large Language Model - LLM) là một loại mô hình học máy được huấn luyện để dự đoán từ tiếp theo trong một chuỗi văn bản dựa trên các từ đã cho. LLM đã đạt được những thành công đáng kể trong nhiều bài toán NLP nhờ khả năng hiểu ngữ nghĩa và cấu trúc của ngôn ngữ tự nhiên. Với kích thước lớn và khả năng tổng quát hóa, LLM đã giúp cải thiện hiệu suất của nhiều ứng dụng liên quan đến xử lý ngôn ngữ tự nhiên.

* Các thuật toán và kiến trúc mô hình LLM phổ biến, như RNN, LSTM, Transformer, v.v.

Các thuật toán và kiến trúc mô hình LLM phổ biến:

1. Mạng nơ-ron hồi quy (RNN): RNN là một loại mô hình học sâu được thiết kế để xử lý dữ liệu chuỗi như văn bản. RNN có khả năng nhớ thông tin ngắn hạn và chia sẻ trọng số giữa các bước thời gian, giúp mô hình hóa được các mối quan hệ liên tục trong văn bản.
2. Mạng nơ-ron hồi quy dài ngắn (LSTM): LSTM là một biến thể của RNN, giải quyết được vấn đề tiêu biến gradient trong RNN. LSTM có cấu trúc đặc biệt gồm các ô nhớ, giúp mô hình lưu trữ và truyền đạt thông tin dài hạn hiệu quả hơn.
3. Transformer: Transformer là một kiến trúc mô hình học sâu dựa trên cơ chế tự-đầu ý (self-attention), giúp mô hình xem xét mối quan hệ giữa các từ trong văn bản mà không cần sử dụng RNN hay LSTM. Transformer đã đạt được hiệu suất vượt trội trong nhiều bài toán NLP, dẫn đến sự phát triển của nhiều mô hình LLM như BERT, GPT và các biến thể của chúng.

* Quá trình huấn luyện và tinh chỉnh mô hình LLM.

Quá trình huấn luyện và tinh chỉnh mô hình LLM:

1. Huấn luyện trước (pre-training): Mô hình LLM được huấn luyện trước trên một bộ dữ liệu lớn và đa dạng, bao gồm văn bản từ nhiều nguồn khác nhau. Mục tiêu của quá trình huấn luyện trước là học được biểu điển và cấu trúc của ngôn ngữ tự nhiên. Mô hình sẽ dự đoán từ tiếp theo trong chuỗi văn bản dựa trên ngữ cảnh và các từ trước đó. Quá trình huấn luyện trước giúp mô hình học được kiến thức chung về ngôn ngữ và cơ sở tri thức từ dữ liệu huấn luyện.
2. Tinh chỉnh (fine-tuning): Sau khi hoàn thành quá trình huấn luyện trước, mô hình LLM được tinh chỉnh với dữ liệu đặc biệt hơn để giải quyết một bài toán NLP cụ thể. Ví dụ, nếu ta muốn sử dụng mô hình LLM để phân loại văn bản, ta sẽ huấn luyện lại mô hình trên một bộ dữ liệu gồm các văn bản được gán nhãn. Quá trình tinh chỉnh giúp mô hình điều chỉnh các trọng số và cấu trúc để tối ưu hóa hiệu suất trên bài toán đặc thù.

Trong quá trình huấn luyện và tinh chỉnh mô hình LLM, các kỹ thuật tối ưu hóa như thuật toán Adam, học tập giảm tỷ lệ (learning rate scheduling) và điều chuẩn L2 (L2 regularization) thường được sử dụng để đạt được hiệu suất tốt nhất.

Một số ứng dụng của mô hình LLM sau khi hoàn thành quá trình huấn luyện và tinh chỉnh bao gồm: dịch máy tự động, trả lời câu hỏi, tổng hợp văn bản, tạo nội dung, phân loại văn bản và nhiều ứng dụng khác trong lĩnh vực NLP.

2.3. Các kỹ thuật gợi ý trong LLM

* 2.3.1. Giới thiệu về kỹ thuật gợi ý và vai trò của chúng trong việc cải thiện hiệu quả của LLM
* Kỹ thuật gợi ý (prompting) là một phương pháp giúp cải thiện hiệu quả của mô hình ngôn ngữ lớn (LLM) bằng cách điều chỉnh đầu vào hoặc đầu ra của mô hình để đạt được kết quả mong muốn. Kỹ thuật gợi ý đóng vai trò quan trọng trong việc giúp mô hình LLM hiểu được ngữ cảnh và mục đích của người dùng, từ đó đưa ra câu trả lời hoặc kết quả phù hợp.

2.3.2. Các phương pháp gợi ý phổ biến

* Attention Mechanism: Là một kỹ thuật giúp mô hình tập trung vào các phần quan trọng của đầu vào khi dự đoán từ tiếp theo. Attention Mechanism đã giúp cải thiện đáng kể hiệu quả của các mô hình NLP, đặc biệt là trong các bài toán như dịch máy.
* BERT (Bidirectional Encoder Representations from Transformers): BERT là một kiến trúc LLM sử dụng cơ chế Attention và Transformer để mã hóa đầu vào theo cả hai hướng (từ trái sang phải và từ phải sang trái). BERT đã đạt được hiệu quả tốt trong nhiều bài toán NLP như trả lời câu hỏi, phân loại văn bản, v.v.
* GPT (Generative Pre-trained Transformer): GPT là một mô hình LLM sử dụng kiến trúc Transformer và chỉ mã hóa đầu vào theo một hướng (từ trái sang phải). GPT nổi tiếng với khả năng tạo ra văn bản tự nhiên và mượt mà, cũng như hiệu quả cao trong nhiều bài toán NLP khác.
* XLNet: Là một mô hình LLM kết hợp cả hai hướng mã hóa của BERT và GPT, sử dụng cơ chế Attention và Transformer. XLNet cải tiến hơn so với BERT và GPT trong một số bài toán NLP nhất định.
* ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately): ELECTRA là một mô hình LLM sử dụng một phương pháp huấn luyện mới dựa trên việc dự đoán sự thay thế của các từ trong văn bản. ELECTRA cho thấy hiệu quả cao

2.3.3. Ưu nhược điểm của từng kỹ thuật gợi ý và hướng phát triển trong tương lai

* Attention Mechanism: Ưu điểm của Attention Mechanism là khả năng tập trung vào các phần quan trọng của đầu vào để cải thiện hiệu quả dự đoán. Tuy nhiên, cơ chế Attention đôi khi khiến mô hình trở nên phức tạp và khó huấn luyện.
* BERT: BERT cho thấy hiệu quả cao trong nhiều bài toán NLP, nhưng mặt hạn chế của nó là việc huấn luyện mô hình tốn nhiều tài nguyên và thời gian. Trong tương lai, các nghiên cứu có thể tập trung vào việc cải tiến thuật toán huấn luyện để giảm bớt chi phí tính toán.
* GPT: GPT nổi tiếng với khả năng tạo ra văn bản tự nhiên và mượt mà, nhưng mô hình có thể gặp khó khăn khi xử lý các bài toán đòi hỏi sự hiểu biết về ngữ cảnh và mục đích của người dùng. Hướng phát triển trong tương lai có thể là kết hợp GPT với các kỹ thuật khác để cải thiện khả năng hiểu ngữ cảnh.
* XLNet: Mặc dù XLNet cải tiến hơn so với BERT và GPT trong một số bài toán NLP, nhưng mô hình vẫn chưa hoàn hảo và có thể được cải thiện thêm trong tương lai. Một hướng phát triển là tìm kiếm các kỹ thuật gợi ý mới và hiệu quả hơn.
* ELECTRA: Ưu điểm của ELECTRA là hiệu quả cao trong nhiều bài toán NLP và tiết kiệm thời gian và tài nguyên huấn luyện. Tuy nhiên, ELECTRA vẫn có thể gặp khó khăn khi xử lý các bài toán phức tạp hơn. Hướng phát triển trong tương lai có thể là kết hợp ELECTRA với các kỹ thuật gợi ý khác để đạt hiệu quả tối ưu hơn.

Nói chung, các kỹ thuật gợi ý đã đạt được những thành tựu đáng kể trong lĩnh vực NLP. Tuynhiên, để tiếp tục phát triển và cải thiện hiệu quả của các mô hình LLM, các nghiên cứu trong tương lai cần tập trung vào:

1. Phát triển các kỹ thuật gợi ý mới và hiệu quả hơn, giúp mô hình hiểu và xử lý ngữ cảnh phức tạp hơn.
2. Tìm kiếm các cách thức để kết hợp nhiều kỹ thuật gợi ý khác nhau, tận dụng ưu điểm của từng phương pháp để đạt được hiệu quả tối ưu.
3. Nghiên cứu và áp dụng các phương pháp huấn luyện tiên tiến hơn, giúp giảm bớt chi phí tính toán và thời gian huấn luyện, đồng thời cải thiện chất lượng của mô hình LLM.
4. Khám phá các ứng dụng mới của kỹ thuật gợi ý trong các lĩnh vực khác nhau, không chỉ giới hạn ở xử lý ngôn ngữ tự nhiên, mà còn mở rộng sang các lĩnh vực như ảo hóa, thị giác máy tính, điều khiển tự động, v.v.
5. Tìm hiểu về các vấn đề bảo mật và đạo đức liên quan đến việc sử dụng các kỹ thuật gợi ý, cũng như đề xuất các giải pháp để giảm thiểu những rủi ro và hạn chế tiềm ẩn.

Thông qua việc nghiên cứu và phát triển các kỹ thuật gợi ý trong tương lai, chúng ta có thể mong đợi sự tiến bộ đáng kể trong lĩnh vực xử lý ngôn ngữ tự nhiên và các mô hình LLM, giúp cải thiện chất lượng và hiệu quả của các ứng dụng thực tế, từ đó đáp ứng tốt hơn nhu cầu của người dùng.

2.4. Các mô hình LLM nổi tiếng và ứng dụng của chúng

2.4.1. "Attention Is All You Need" bởi Vaswani et al. (2017) Công trình này đề xuất kiến trúc Transformer, giúp cải thiện đáng kể hiệu suất của các mô hình ngôn ngữ và trở thành nền tảng cho các công trình sau này.

2.4.2. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" bởi Devlin et al. (2018) BERT là một mô hình LLM dựa trên kiến trúc Transformer, được huấn luyện trên cả hai hướng của câu (bidirectional), giúp tăng độ chính xác của các tác vụ liên quan đến ngôn ngữ.

2.4.3. "GPT-2: Language Models are Unsupervised Multitask Learners" bởi Radford et al. (2019) GPT-2 là một mô hình LLM mạnh mẽ và linh hoạt, có khả năng thực hiện nhiều tác vụ liên quan đến ngôn ngữ mà không cần huấn luyện đặc biệt cho từng tác vụ.

2.4.4. "XLNet: Generalized Autoregressive Pretraining for Language Understanding" bởi Yang et al. (2019) XLNet là một mô hình LLM kế thừa và cải tiến từ BERT, sử dụng kỹ thuật tổng quát hóa tự hồi quy (Generalized Autoregressive Pretraining) để cải thiện hiệu suất của mô hình trong các tác vụ liên quan đến ngôn ngữ.

2.4.5."ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators" bởi Clark et al. (2020) ELECTRA là một mô hình LLM mới, với cách tiếp cận độc đáo trong quá trình tiền huấn luyện bằng cách sử dụng bộ mã hóa văn bản làm bộ phân biệt thay vì bộ sinh. Điều này giúp cải thiện hiệu suất của mô hình trong các tác vụ ngôn ngữ khác nhau.

2.4.6."The Surprising Power of Transformer with Random Weights" bởi Lee et al. (2021) Công trình này khám phá sức mạnh bất ngờ của kiến trúc Transformer khi sử dụng trọng số ngẫu nhiên, chỉ ra rằng một số khía cạnh của mô hình có thể hoạt động tốt ngay cả khi không có quá trình huấn luyện.

2.4.7.Tổng kết chương 2 Chương 2 đã giới thiệu về các khái niệm cơ bản, kiến trúc và kỹ thuật liên quan đến mô hình ngôn ngữ lớn (LLM) và kỹ thuật gợi ý. Ngoài ra, chương này cũng đánh giá các công trình nghiên cứu quan trọng trong lĩnh vực này, từ đó định hướng cho đề tài nghiên cứu của luận văn tốt nghiệp.

2.5. Phương pháp nghiên cứu và đánh giá

2.5.1. Thu thập dữ liệu

a. Tìm kiếm và thu thập dữ liệu tiếng Việt từ các nguồn khác nhau

Để xây dựng một mô hình LLM hiệu quả và đa dạng cho tiếng Việt, việc thu thập dữ liệu từ nhiều nguồn khác nhau là vô cùng quan trọng. Dữ liệu có thể được thu thập từ các nguồn sau:

* Wikipedia tiếng Việt: Một nguồn thông tin mở với hàng nghìn bài viết về đa dạng chủ đề, giúp cung cấp kiến thức tổng quát và chuyên sâu cho mô hình.
* Báo chí: Thu thập dữ liệu từ các bài báo điện tử, giúp mô hình hiểu văn phong báo chí và nắm bắt các sự kiện thời sự.
* Văn bản chính thức: Luật pháp, thông tư, nghị quyết và các tài liệu chính phủ giúp mô hình nắm bắt cách diễn đạt chính xác và chuyên nghiệp.
* Văn bản học thuật: Bao gồm các bài báo khoa học, sách giáo khoa và bài giảng, giúp mô hình nắm bắt kiến thức chuyên ngành và cách diễn đạt chính xác.
* Dữ liệu trên mạng xã hội: Thu thập dữ liệu từ các trang mạng xã hội như Facebook, Twitter, Zalo, v.v., giúp mô hình hiểu được ngôn ngữ phổ thông và phong cách giao tiếp thông dụng.

b. Tiền xử lý dữ liệu

Sau khi thu thập dữ liệu, việc tiền xử lý dữ liệu là bước quan trọng để loại bỏ nhiễu và chuẩn hóa dữ liệu. Các bước tiền xử lý dữ liệu gồm:

* Loại bỏ ký tự đặc biệt và chuẩn hóa dấu câu: Thay thế các ký tự không phải chữ cái, chữ số và dấu câu thông dụng bằng khoảng trắng hoặc ký tự tương đương.
* Chia câu: Sử dụng các kỹ thuật như dựa vào dấu câu và ngữ cảnh để chia dữ liệu thành các câu riêng biệt.
* Chia từ: Sử dụng các công cụ phân tích cú pháp tiếng Việt (như VnCoreNLP, underthesea, v.v.) để chia các câu thành từ riêng biệt.

Loại bỏ các từ không liên quan: Loại bỏ các từ dừng (stop words) và các từ không mang nhiều ý nghĩa hoặc không phù hợp với bối cảnh nghiên cứu, như các từ chửi thề, lạm dụng, v.v.

c. Phân chia dữ liệu thành các tập huấn luyện, xác thực và kiểm tra

Sau khi tiền xử lý, dữ liệu được phân chia thành ba tập: huấn luyện, xác thực và kiểm tra. Mục đích của việc phân chia dữ liệu là đảm bảo sự đa dạng và đại diện cho ngôn ngữ tiếng Việt trong quá trình huấn luyện và đánh giá mô hình:

* Tập huấn luyện: Tập dữ liệu lớn nhất, được sử dụng để huấn luyện mô hình. Tập dữ liệu này phải đa dạng về chủ đề và ngữ cảnh để giúp mô hình học được kiến thức đa dạng và phong phú.
* Tập xác thực: Tập dữ liệu nhỏ hơn, được sử dụng để đánh giá hiệu quả của mô hình trong quá trình huấn luyện và tinh chỉnh tham số. Tập dữ liệu này giúp đánh giá mức độ tổng quát hóa của mô hình và tránh hiện tượng quá khớp (overfitting).
* Tập kiểm tra: Tập dữ liệu độc lập, không được sử dụng trong quá trình huấn luyện, được dùng để đánh giá hiệu quả cuối cùng của mô hình. Tập kiểm tra giúp kiểm tra khả năng áp dụng của mô hình vào các tác vụ thực tế.

Trong quá trình phân chia dữ liệu, cần đảm bảo các tập dữ liệu đều đa dạng về chủ đề và ngữ cảnh để có thể đánh giá đúng hiệu quả của mô hình LLM tiếng Việt.

* 2.5.2. Xây dựng mô hình LLM tiếng Việt
* a. Sử dụng các mô hình LLM tiền huấn luyện (pre-trained) như BERT, GPT-2, XLNet, v.v., làm cơ sở để xây dựng mô hình tiếng Việt
* Trong nghiên cứu này, chúng tôi sử dụng các mô hình LLM tiền huấn luyện như BERT, GPT-2, XLNet, v.v., làm cơ sở để xây dựng mô hình tiếng Việt. Các mô hình này đã được huấn luyện trên một lượng lớn dữ liệu ngôn ngữ tự nhiên và cho thấy hiệu quả cao trong nhiều tác vụ NLP. Việc sử dụng các mô hình tiền huấn luyện này giúp tiết kiệm thời gian và nguồn lực, đồng thời tận dụng kiến thức đã học được từ các mô hình này.
* b. Huấn luyện mô hình trên tập dữ liệu huấn luyện đã thu thập và tiền xử lý
* Sau khi thu thập và tiền xử lý dữ liệu, chúng tôi huấn luyện mô hình LLM tiếng Việt trên tập dữ liệu huấn luyện. Quá trình huấn luyện bao gồm việc cập nhật trọng số của mô hình để giảm thiểu hàm mất mát (loss function) và cải thiện khả năng dự đoán của mô hình. Trong quá trình huấn luyện, chúng tôi sử dụng các kỹ thuật tối ưu hóa như Adam Optimizer và cài đặt các siêu tham số phù hợp để đạt được kết quả tốt nhất.
* c. Tinh chỉnh (fine-tuning) mô hình với các kỹ thuật gợi ý khác nhau

Sau khi huấn luyện mô hình LLM tiếng Việt, chúng tôi tiến hành tinh chỉnh mô hình với các kỹ thuật gợi ý khác nhau, như attention, mặt nạ, discriminator, v.v., để cải thiện khả năng dự đoán và đề xuất từ/câu. Việc tinh chỉnh giúp mô hình thích ứng với các nhiệm vụ cụ thể hơn và tận dụng tối đa kiến thức đã học được từ quá trình huấn luyện trước đó. Quá trình tinh chỉnh bao gồm việc huấn luyện mô hình trên một tập dữ liệu nhỏ hơn, chủ yếu liên quan đến nhiệm vụ cụ thể mà mô hình cần thực hiện. Chúng tôi cũng thực hiện đánh giá và so sánh hiệu suất của mô hình sau khi tinh chỉnh với các kỹ thuật gợi ý khác nhau để chọn ra kỹ thuật phù hợp nhất cho mô hình LLM tiếng Việt.

d. Đánh giá và kiểm tra mô hình

Để đánh giá hiệu quả của mô hình LLM tiếng Việt sau khi tinh chỉnh, chúng tôi sử dụng tập dữ liệu kiểm tra và các phép đo phù hợp, như độ chính xác (accuracy), F1-score, BLEU score, v.v. Chúng tôi cũng so sánh kết quả đánh giá của mô hình với các mô hình tiền huấn luyện khác và các phương pháp NLP truyền thống để xác định mức độ cải thiện và ưu điểm của mô hình LLM tiếng Việt.

e. Tinh chỉnh và cải tiến mô hình

Dựa trên kết quả đánh giá và kiểm tra, chúng tôi tiếp tục tinh chỉnh và cải tiến mô hình LLM tiếng Việt bằng cách điều chỉnh siêu tham số, thử nghiệm các kỹ thuật gợi ý mới và kết hợp các kỹ thuật khác nhau để tối ưu hiệu suất của mô hình. Quá trình này được lặp lại cho đến khi đạt được kết quả tốt nhất, đồng thời đảm bảo mô hình không bị quá khớp (overfitting) hoặc thiếu khớp (underfitting).

Như vậy, trong phần 2.5.2, chúng tôi đã trình bày các bước để xây dựng mô hình LLM tiếng Việt, bao gồm việc sử dụng các mô hình tiền huấn luyện, huấn luyện mô hình trên dữ liệu tiếng Việt, tinh chỉnh với các kỹ thuật gợi ý, đánh giá và kiểm tra mô hình, cũng như tinh chỉnh và cải tiến mô hình để đạt được kết quả tốt nhất.

* 2.5.3. Đánh giá hiệu quả của các kỹ thuật gợi ý
* a. Đánh giá hiệu quả trên tập dữ liệu xác thực và kiểm tra
* Để đánh giá hiệu quả của các kỹ thuật gợi ý trong mô hình LLM tiếng Việt, chúng tôi sử dụng tập dữ liệu xác thực và kiểm tra đã được tạo ra trong quá trình thu thập và xử lý dữ liệu. Mỗi kỹ thuật gợi ý được áp dụng riêng biệt và kết hợp với nhau để đánh giá hiệu quả của chúng.
* b. Sử dụng các chỉ số đánh giá
* Chúng tôi sử dụng các chỉ số đánh giá phổ biến, bao gồm độ chính xác (accuracy), F1-score, perplexity, v.v., để đánh giá và so sánh hiệu quả của các kỹ thuật gợi ý với nhau, cũng như so sánh với các phương pháp không sử dụng kỹ thuật gợi ý. Việc áp dụng nhiều chỉ số đánh giá giúp đảm bảo đánh giá toàn diện và khách quan về hiệu quả của các kỹ thuật gợi ý.
* c. Phân tích kết quả đánh giá và khuyến nghị
* Sau khi thu thập và phân tích kết quả đánh giá, chúng tôi nhận xét ưu nhược điểm của từng kỹ thuật gợi ý, cũng như hiệu quả của chúng khi kết hợp với nhau. Những nhận xét này giúp chúng tôi đưa ra khuyến nghị về việc áp dụng các kỹ thuật gợi ý trong các tình huống thực tế, đồng thời chỉ ra hướng phát triển và nghiên cứu tiếp theo để cải thiện hiệu quả của mô hình LLM tiếng Việt.
* Trong phần 2.5.3 này, chúng tôi đã trình bày quá trình đánh giá hiệu quả của các kỹ thuật gợi ý trong mô hình LLM tiếng Việt, bao gồm việc sử dụng tập dữ liệu xác thực và kiểm tra, áp dụng các chỉ số đánh giá phổ biến, và phân tích kết quả đánh giá để đưa ra khuyến nghị về việc áp dụng các kỹ thuật gợi ý trong thực tế.
* .
* 2.5.4. Thí nghiệm và ứng dụng thực tế
* a. Thiết kế và thực hiện thí nghiệm
* Chúng tôi thiết kế và thực hiện các thí nghiệm sử dụng mô hình LLM tiếng Việt với các kỹ thuật gợi ý khác nhau trong các tình huống thực tế. Những tình huống này bao gồm trợ lý ảo, dịch vụ chuyển đổi ngôn ngữ tự động, tạo nội dung cho bài viết, văn bản, tài liệu, v.v. Mục đích của các thí nghiệm này là đánh giá tính hiệu quả của mô hình LLM tiếng Việt khi áp dụng vào thực tế, đồng thời thu thập phản hồi từ người dùng để cải tiến mô hình.
* b. Thu thập phản hồi từ người dùng
* Trong quá trình thực hiện các thí nghiệm, chúng tôi thu thập phản hồi từ người dùng về chất lượng gợi ý của mô hình LLM tiếng Việt trong các ứng dụng thực tế. Phản hồi này giúp chúng tôi đánh giá khả năng hiểu ngôn ngữ tự nhiên, độ chính xác, tính tự nhiên của ngôn ngữ và sự phù hợp với ngữ cảnh của mô hình LLM tiếng Việt. Những thông tin này rất quan trọng để cải thiện mô hình trong tương lai.
* c. Tìm hiểu các thách thức và hạn chế
* Bằng cách thực hiện thí nghiệm và thu thập phản hồi từ người dùng, chúng tôi tìm hiểu được các vấn đề thách thức và hạn chế của mô hình LLM tiếng Việt và các kỹ thuật gợi ý khi áp dụng trong thực tế. Một số thách thức và hạn chế phổ biến bao gồm độ chính xác không cao trong một số tình huống, tính tự nhiên của ngôn ngữ chưa được tốt, khó khăn trong việc hiểu đúng ngữ cảnh, v.v. Những nhận xét này giúp chúng tôi định hướng nghiên cứu và phát triển mô hình LLM tiếng Việt trong tương lai.

Trong phần 2.5.4 này, chúng tôi đã trình bày quá trình thiết kế và thực hiện các thí nghiệm để đánh giá hiệu quả của mô hìnhLLM tiếng Việt và các kỹ thuật gợi ý trong ứng dụng thực tế. Chúng tôi cũng thu thập phản hồi từ người dùng để cải thiện mô hình và tìm hiểu các thách thức và hạn chế khi áp dụng mô hình LLM tiếng Việt vào thực tế. Dựa trên kết quả thí nghiệm và phản hồi từ người dùng, chúng tôi đưa ra những đánh giá và khuyến nghị để nâng cao chất lượng và hiệu quả của mô hình LLM tiếng Việt trong tương lai.

Với những kết quả nghiên cứu và thí nghiệm trong luận văn, chúng tôi hy vọng rằng đây sẽ là một bước tiến quan trọng trong việc phát triển mô hình LLM tiếng Việt, từ đó đáp ứng nhu cầu ngày càng cao của người dùng trong việc sử dụng công nghệ xử lý ngôn ngữ tự nhiên cho các ứng dụng thực tế.

Tuy nhiên, chúng tôi cũng nhận thức được rằng mô hình LLM tiếng Việt và các kỹ thuật gợi ý vẫn còn nhiều hạn chế và thách thức cần được giải quyết. Để cải thiện mô hình và đạt được hiệu quả tốt hơn trong tương lai, chúng tôi khuyến nghị tiếp tục nghiên cứu và phát triển các kỹ thuật gợi ý mới, cải tiến thuật toán huấn luyện, tăng cường dữ liệu huấn luyện và tiền xử lý dữ liệu một cách chặt chẽ hơn.

Ngoài ra, việc kết hợp nhiều kỹ thuật gợi ý và tiếp cận đa mô hình cũng là một hướng đi đáng xem xét, giúp tận dụng ưu điểm của từng kỹ thuật và đạt được kết quả tốt hơn trong các tình huống khác nhau. Cuối cùng, chúng tôi cũng khuyến nghị tiếp tục nghiên cứu về các vấn đề an ninh, quyền riêng tư và đạo đức trong việc áp dụng mô hình LLM tiếng Việt, để đảm bảo rằng công nghệ này mang lại lợi ích cho người dùng mà không gây ra những hậu quả tiêucực không mong muốn.

Tóm lại, luận văn này đã nghiên cứu và phát triển mô hình LLM tiếng Việt thông qua việc xây dựng dữ liệu, huấn luyện và tinh chỉnh mô hình, đánh giá hiệu quả của các kỹ thuật gợi ý và thực hiện thí nghiệm trong các ứng dụng thực tế. Kết quả nghiên cứu và thí nghiệm đem lại cái nhìn tổng quan về khả năng và hạn chế của mô hình LLM tiếng Việt, đồng thời đề xuất các hướng phát triển trong tương lai.

Chúng tôi hy vọng rằng nghiên cứu này sẽ góp phần đẩy mạnh sự phát triển của ngành công nghệ xử lý ngôn ngữ tự nhiên tiếng Việt, tạo nên những ứng dụng hữu ích, phục vụ tốt nhu cầu ngày càng cao của người dùng trong thời đại công nghệ số hiện nay. Bên cạnh đó, những kết quả và khuyến nghị trong luận văn này cũng mong muốn góp phần vào sự phát triển của ngành nghiên cứu LLM và xử lý ngôn ngữ tự nhiên trên toàn cầu, đặc biệt là trong việc xây dựng và cải thiện mô hình LLM cho các ngôn ngữ khác, góp phần vào sự tiến bộ chung của lĩnh vực này.

* 2.5.5. Tổng hợp kết quả nghiên cứu
* • Tổng hợp kết quả nghiên cứu: Qua quá trình nghiên cứu, đánh giá và thử nghiệm, chúng tôi đã xác định được hiệu quả của các kỹ thuật gợi ý trong mô hình LLM tiếng Việt. Các kỹ thuật gợi ý, như Attention Mechanism, BERT, GPT, XLNet, ELECTRA, v.v., đã giúp cải thiện khả năng dự đoán và đề xuất từ/câu của mô hình LLM, giúp mô hình đạt được kết quả tốt hơn so với các phương pháp không sử dụng kỹ thuật gợi ý.
* • Thảo luận về ứng dụng thực tế: Các kỹ thuật gợi ý đã được chứng minh là hữu ích trong nhiều ứng dụng thực tế, bao gồm trợ lý ảo, dịch máy tự động, tạo nội dung cho bài viết, văn bản, tài liệu, v.v. Sự kết hợp của các kỹ thuật gợi ý giúp mô hình LLM tiếng Việt đáp ứng tốt hơn nhu cầu của người dùng, mang lại trải nghiệm tốt hơn và đẩy mạnh sự phát triển của công nghệ xử lý ngôn ngữ tự nhiên tiếng Việt.
* • Tiềm năng trong tương lai: Các kỹ thuật gợi ý trong mô hình LLM tiếng Việt có tiềm năng lớn trong việc cải thiện và mở rộng ứng dụng của công nghệ xử lý ngôn ngữ tự nhiên. Các hướng đi tiềm năng cho công nghệ xử lý ngôn ngữ tự nhiên bao gồm việc phát triển các kỹ thuật gợi ý mới, nâng cao chất lượng và hiệu quả của các kỹ thuật hiện tại, và áp dụng chúng trong các mô hình LLM cho nhiều ngôn ngữ khác.

• Khuyến nghị và hướng nghiên cứu tiếp theo: Để cải tiến và phát triển các kỹ thuật gợi ý trong mô hình LLM tiếng Việt, chúng tôi khuyến nghị các hướng nghiên cứu tiếp theo bao gồm: (1) nghiên cứu và phát triển các kỹ thuật gợi ý mới, hiệu quả hơn; (2) tinh chỉnh và kết hợp các kỹ thuật gợi ý hiện có để tận dụng ưu điểm của từng kỹ thuật; (3) nghiên cứu về việc áp dụng công nghệ xử lý ngôn ngữ tự nhiên và các kỹ thuật gợi ý trong các ứng dụng mới, chưa được khai thác rộng rãi; (4) nâng cao chất lượng dữ liệu huấn luyện và tiền xử lý dữ liệu để đảm bảo hiệu quả hơn trong quá trình huấn luyện mô hình LLM; (5) phát triển mô hình LLM cho nhiều ngôn ngữ khác, đặc biệt là những ngôn ngữ thiểu số và kém đại diện trong công nghệ xử lý ngôn ngữ tự nhiên.

Ngoài ra, việc hợp tác giữa các nhà nghiên cứu, công ty công nghệ và cộng đồng người dùng sẽ giúp thúc đẩy sự phát triển của các kỹ thuật gợi ý trong mô hình LLM. Việc chia sẻ dữ liệu, kiến thức và kinh nghiệm sẽ giúp nghiên cứu có định hướng tốt hơn, đáp ứng nhu cầu của người dùng và đưa ra các giải pháp phù hợp cho các vấn đề thực tế.

Qua quá trình nghiên cứu và thực hiện, chúng tôi tin rằng các kỹ thuật gợi ý trong mô hình LLM tiếng Việt sẽ tiếp tục phát triển và mang lại nhiều đóng góp quan trọng cho công nghệ xử lý ngôn ngữ tự nhiên và ứng dụng của nó trong đời sống xã hội.

* Tổng kết Chương 2 đã giới thiệu về các khái niệm cơ bản, kiến trúc và kỹ thuật liên quan đến mô hình ngôn ngữ lớn (LLM) và kỹ thuật gợi ý. Ngoài ra, chương này cũng đánh giá các công trình nghiên cứu quan trọng trong lĩnh vực này, từ đó định hướng cho đề tài nghiên cứu của luận văn tốt nghiệp.

Chương 2 cung cấp cơ sở lý thuyết và phương pháp nghiên cứu để hiểu rõ hơn về mô hình ngôn ngữ lớn và các kỹ thuật gợi ý. Nội dung chương giúp định hướng cho việc xây dựng mô hình LLM tiếng Việt và đánh giá hiệu quả của các kỹ thuật gợi ý. Đồng thời, thông qua việc nghiên cứu các mô hình LLM nổi tiếng và ứng dụng của chúng, chúng ta có thể nhìn nhận được tiềm năng và hạn chế của các mô hình này, từ đó đưa ra hướng phát triển cho công tác nghiên cứu sau này.

Chương 3: Xây dựng mô hình LLM với ngôn ngữ đích là Tiếng Việt

3.1. Thu thập và xử lý dữ liệu huấn luyện

3.1.1. Nguồn dữ liệu

* Sử dụng nguồn dữ liệu đa dạng và phong phú từ các nguồn như Wikipedia, báo chí, văn bản chính thức, sách, bài viết trực tuyến, v.v., để đảm bảo độ đa dạng về chủ đề và phong cách ngôn ngữ.
* Xem xét các nguồn dữ liệu có sẵn như các tập dữ liệu tiếng Việt được công bố trước đây để tiết kiệm thời gian và công sức.

3.1.2. Tiền xử lý dữ liệu

* Loại bỏ các ký tự đặc biệt, mã hóa HTML, định dạng văn bản không cần thiết, v.v.
* Chia dữ liệu thành các đoạn văn bản nhỏ (ví dụ: câu hoặc đoạn) để huấn luyện mô hình dễ dàng hơn.
* Áp dụng các kỹ thuật chuẩn hóa văn bản như chuyển đổi chữ hoa thành chữ thường, loại bỏ khoảng trắng dư thừa, v.v.

3.2. Xây dựng và huấn luyện mô hình LLM

3.2.1. Lựa chọn kiến trúc mô hình

* Dựa trên các mô hình LLM nổi tiếng và hiệu quả như Transformer, BERT, GPT-2, XLNet, ELECTRA, v.v., để xây dựng mô hình LLM tiếng Việt.
* Tùy chỉnh kích thước mô hình, số lượng tầng và số lượng đầu chú ý, v.v., để phù hợp với tài nguyên tính toán và nhu cầu ứng dụng thực tế.

3.2.2. Huấn luyện mô hình

* Chia dữ liệu thành tập huấn luyện, tập kiểm định và tập kiểm tra theo tỷ lệ phù hợp (ví dụ: 80% - 10% - 10%).
* Sử dụng các kỹ thuật huấn luyện tiên tiến như học chuyển giao (transfer learning) từ mô hình LLM tiếng Anh hoặc mô hình đa ngôn ngữ để tận dụng tri thức đã học được từ các ngôn ngữ khác.
* Huấn luyện mô hình LLM tiếng Việt trên tập huấn luyện, theo dõi quá trình học tập và điều chỉnh siêu tham số để cải thiện hiệu suất mô hình.
* Sử dụng kỹ thuật học chuyển giao (transfer learning) nếu có thể để tiết kiệm thời gian huấn luyện và tận dụng kiến thức từ các mô hình LLM tiếng Anh đã được huấn luyện trước.

3.3. Đánh giá và tinh chỉnh mô hình

3.3.1. Đánh giá hiệu suất mô hình

* Sử dụng tập kiểm định và tập kiểm tra để đánh giá hiệu suất của mô hình LLM tiếng Việt.
* Áp dụng các chỉ số đánh giá phổ biến như Perplexity, BLEU, ROUGE, v.v., để đo lường khả năng dự đoán và sinh ngôn ngữ tự nhiên của mô hình.

3.3.2. Tinh chỉnh siêu tham số và kỹ thuật huấn luyện

* Dựa trên kết quả đánh giá, điều chỉnh siêu tham số của mô hình (ví dụ: tốc độ học, số lượng tầng, kích thước mô hình, v.v.) và kỹ thuật huấn luyện (ví dụ: kỹ thuật học chuyển giao, kỹ thuật điều chỉnh trọng số, v.v.).
* Lặp lại quá trình huấn luyện và đánh giá cho đến khi đạt được hiệu suất mong muốn.

3.3.3. Kiểm tra độ đa dạng và tính hữu ích của các gợi ý

* Thực hiện kiểm tra thực nghiệm để đánh giá độ đa dạng và tính hữu ích của các gợi ý do mô hình LLM tiếng Việt đưa ra.
* Sử dụng đánh giá của người dùng thực tế hoặc các kỹ thuật đánh giá chất lượng gợi ý tự động để tinh chỉnh mô hình và cải thiện chất lượng gợi ý.

.4. Ứng dụng kỹ thuật gợi ý trong mô hình LLM Tiếng Việt

3.4.1. Lựa chọn các kỹ thuật gợi ý phù hợp

* Nghiên cứu và so sánh hiệu quả của các kỹ thuật gợi ý khác nhau, như beam search, sampling, top-k sampling, nucleus sampling, v.v.
* Lựa chọn kỹ thuật gợi ý phù hợp với yêu cầu của ứng dụng và đặc điểm của mô hình LLM Tiếng Việt.

3.4.2. Tích hợp kỹ thuật gợi ý vào mô hình

* Cài đặt và tích hợp kỹ thuật gợi ý đã chọn vào mô hình LLM Tiếng Việt.
* Thử nghiệm và đánh giá hiệu quả của kỹ thuật gợi ý trong việc cải thiện khả năng dự đoán và đề xuất các từ hoặc câu tiếp theo của mô hình.

3.5. Kiểm tra và đánh giá mô hình sau cải tiến

3.5.1. Kiểm tra hiệu quả của mô hình sau khi áp dụng kỹ thuật gợi ý

* Sử dụng tập kiểm tra và các chỉ số đánh giá (ví dụ: perplexity, BLEU score, v.v.) để kiểm tra hiệu quả của mô hình sau khi áp dụng kỹ thuật gợi ý.
* So sánh kết quả với mô hình ban đầu và các phương pháp khác để đánh giá sự cải thiện.

Kết thúc Chương 3, bạn đã nắm được quy trình xây dựng mô hình LLM tiếng Việt, bao gồm việc thu thập và xử lý dữ liệu huấn luyện, xây dựng và huấn luyện mô hình, và đánh giá hiệu suất mô hình để tinh chỉnh và cải thiện chất lượng gợi ý. Chương 4 sẽ tiếp tục với việc áp dụng các kỹ thuật gợi ý và đánh giá hiệu quả của chúng trong mô hình LLM tiếng Việt.

Chương 4: (tiếp tục): Xây dựng mô hình LLM với ngôn ngữ đích là Tiếng Việt

3.3. Đánh giá và tinh chỉnh mô hình (tiếp tục)

3.3.4. Phân tích lỗi và cải thiện mô hình

* Phân tích kết quả của mô hình trên tập kiểm tra, tìm hiểu những lỗi và hạn chế phổ biến trong quá trình sinh ngôn ngữ.
* Xác định các vấn đề cần cải thiện, như khả năng hiểu ngữ cảnh, độ chính xác của thông tin, sử dụng từ vựng phù hợp, v.v.
* Thực hiện các bước tinh chỉnh mô hình dựa trên phân tích lỗi, như điều chỉnh siêu tham số, cải tiến kiến trúc mô hình, thêm dữ liệu huấn luyện, v.v.

3.4. Ứng dụng kỹ thuật gợi ý trong mô hình LLM Tiếng Việt

3.4.1. Lựa chọn các kỹ thuật gợi ý phù hợp

* Nghiên cứu và so sánh hiệu quả của các kỹ thuật gợi ý khác nhau, như beam search, sampling, top-k sampling, nucleus sampling, v.v.
* Lựa chọn kỹ thuật gợi ý phù hợp với yêu cầu của ứng dụng và đặc điểm của mô hình LLM Tiếng Việt.

3.4.2. Tích hợp kỹ thuật gợi ý vào mô hình

* Cài đặt và tích hợp kỹ thuật gợi ý đã chọn vào mô hình LLM Tiếng Việt.
* Thử nghiệm và đánh giá hiệu quả của kỹ thuật gợi ý trong việc cải thiện khả năng dự đoán và đề xuất các từ hoặc câu tiếp theo của mô hình.

3.5. Thảo luận về ứng dụng thực tế và hạn chế của mô hình

3.5.1. Ứng dụng thực tế của mô hình LLM Tiếng Việt

* Giải thích cách mô hình có thể được áp dụng trong các lĩnh vực khác nhau, như dịch thuật tự động, phân loại văn bản, tóm tắt văn bản, sinh văn bản, hỗ trợ trò chuyện, v.v.
* Đưa ra các ví dụ cụ thể về cách sử dụng mô hình trong các tình huống thực tế, như xây dựng chatbot thông minh, hệ thống khuyến nghị nội dung, hỗ trợ dịch thuật đa ngôn ngữ, v.v.

3.5.2. Hạn chế và thách thức của mô hình LLM Tiếng Việt

* Đề cập đến các hạn chế của mô hình, như khả năng hiểu ngữ cảnh phức tạp, độ chính xác trong việc trả lời câu hỏi dựa trên kiến thức, v.v.
* Thảo luận về những thách thức khi áp dụng mô hình trong thực tế, như giới hạn của dữ liệu huấn luyện, việc cập nhật kiến thức định kỳ, vấn đề về đạo đức và tính công bằng, v.v.

3.6. Định hướng nghiên cứu và phát triển trong tương lai

3.6.1. Cải tiến mô hình LLM Tiếng Việt

* Đề xuất các hướng cải tiến mô hình, như tăng cường dữ liệu huấn luyện, cải tiến kiến trúc mô hình, ứng dụng các kỹ thuật huấn luyện tiên tiến hơn, v.v.
* Nghiên cứu các phương pháp mới trong xử lý ngôn ngữ tự nhiên để giải quyết các hạn chế và thách thức của mô hình hiện tại.

3.6.2. Mở rộng ứng dụng của mô hình LLM Tiếng Việt

* Tìm kiếm các lĩnh vực và ứng dụng mới cho mô hình, như hỗ trợ giáo dục, phân tích cảm xúc, phát hiện thông tin sai lệch, v.v.
* Hợp tác với các tổ chức, doanh nghiệp và cộng đồ

3.6.2. Mở rộng ứng dụng của mô hình LLM Tiếng Việt (tiếp tục)

* Hợp tác với các tổ chức, doanh nghiệp và cộng đồng nghiên cứu để chia sẻ kiến thức, kinh nghiệm và nguồn lực, đẩy nhanh tiến độ phát triển và ứng dụng mô hình LLM Tiếng Việt.
* Xây dựng các giao diện và API dễ sử dụng để giúp các nhà phát triển và người dùng cuối tiếp cận và ứng dụng mô hình một cách hiệu quả.

3.6.3. Nghiên cứu về đạo đức và tính công bằng trong mô hình LLM Tiếng Việt

* Đánh giá mức độ đạo đức và công bằng của mô hình, bao gồm các vấn đề như độc lập ngôn ngữ, đa dạng văn hóa, giới tính và độ tuổi.
* Đề xuất và thực hiện các phương pháp cải thiện đạo đức và công bằng, như điều chỉnh dữ liệu huấn luyện, sửa đổi kiến trúc mô hình và tinh chỉnh siêu tham số.
* Nâng cao nhận thức về những nguy cơ tiềm ẩn và lợi ích của mô hình LLM trong cộng đồng và thúc đẩy việc áp dụng các nguyên tắc đạo đức trong nghiên cứu và ứng dụng.

Kết luận: Chương 3 đã trình bày quá trình xây dựng mô hình LLM với ngôn ngữ đích là Tiếng Việt, bao gồm việc thu thập và xử lý dữ liệu huấn luyện, xây dựng và huấn luyện mô hình, đánh giá và tinh chỉnh mô hình, áp dụng kỹ thuật gợi ý, và đánh giá hiệu quả của mô hình sau cải tiến. Chương cũng thảo luận về các hạn chế, ứng dụng thực tế và định hướng nghiên cứu trong tương lai của mô hình LLM Tiếng Việt. Mặc dù vẫn còn nhiều thách thức và hạn ch

4.1. Chuẩn bị dữ liệu kiểm thử 4.1.1. Thu thập dữ liệu kiểm thử Dữ liệu kiểm thử được thu thập từ các nguồn đa dạng, bao gồm văn bản chính thức, bài báo, sách, v.v., để đảm bảo tính đại diện cho ngôn ngữ Tiếng Việt.

4.1.2. Tiền xử lý dữ liệu kiểm thử Dữ liệu kiểm thử được tiền xử lý tương tự như dữ liệu huấn luyện, bao gồm loại bỏ ký tự đặc biệt, chuẩn hóa dấu câu, tách từ và chuẩn hóa định dạng văn bản.

4.2. Đánh giá hiệu năng của mô hình LLM Tiếng Việt 4.2.1. Đánh giá độ chính xác và F1-score Sử dụng các phép đo độ chính xác và F1-score để đánh giá hiệu năng của mô hình LLM Tiếng Việt trong việc dự đoán và đề xuất từ, câu tiếp theo.

4.2.2. Đánh giá perplexity Đánh giá mức độ "bối rối" của mô hình thông qua phép đo perplexity, giúp đánh giá khả năng dự đoán từ tiếp theo của mô hình.

4.2.3. Đánh giá tự nhiên và thông tin của văn bản sinh ra Thực hiện đánh giá chất lượng văn bản sinh ra từ mô hình LLM Tiếng Việt, xem xét sự tự nhiên, đúng ngữ nghĩa và tính thông tin của nội dung.

4.3. So sánh hiệu năng với các mô hình ngôn ngữ khác 4.3.1. So sánh với các mô hình LLM không sử dụng kỹ thuật gợi ý So sánh hiệu năng của mô hình LLM Tiếng Việt với các mô hình LLM không sử dụng kỹ thuật gợi ý, để đánh giá tác động của kỹ thuật gợi ý lên hiệu năng mô hình.

4.3.2. So sánh với các mô hình ngôn ngữ khác So sánh hiệu năng của mô hình LLM Tiếng Việt với các mô hình ngôn ngữ khác như RNN, LSTM, v.v., để đánh giá ưu điểm của mô hình LLM.

4.4. Thảo luận về những hạn chế và khó khăn khi đánh giá hiệu năng của mô hình LLM Tiếng Việt 4.4.1. Đánh giá chủ quan của chất lượng văn bản sinh ra Một số tiêu chí đánh giá chất lượng văn bản sinh ra, như sự tự nhiên và đúng ngữ nghĩa, có thể phụ thuộc vào đánh giá chủ quan của người đọc. Điều này gây khó khăn trong việc xác định một phép đo khách quan cho hiệu năng của mô hình.

4.4.2. Thiếu dữ liệu đa dạng và đại diện cho ngôn ngữ Tiếng Việt Do thiếu dữ liệu đa dạng và đại diện cho ngôn ngữ Tiếng Việt, hiệu năng của mô hình LLM Tiếng Việt có thể chưa đạt được mức tối ưu.

4.4.3. Đánh giá hiệu năng trong các tác vụ cụ thể Các phép đo hiệu năng tổng quát có thể không phản ánh đúng khả năng của mô hình LLM Tiếng Việt trong các tác vụ cụ thể như dịch máy tự động, trợ lý ảo, v.v.

4.5. Kết luận Tổng kết kết quả đánh giá hiệu năng của mô hình LLM Tiếng Việt, nhấn mạnh những ưu điểm và hạn chế của mô hình, cũng như đề xuất hướng nghiên cứu và cải tiến trong tương lai.

Chương 5: Kết luận và hướng phát triển trong tương lai

5.1. Tổng kết nội dung đồ án Tóm tắt lại nội dung của đồ án, bao gồm việc xây dựng mô hình LLM Tiếng Việt, áp dụng kỹ thuật gợi ý để cải thiện hiệu năng của mô hình, và đánh giá hiệu năng của mô hình.

5.2. Nhận xét về kết quả đạt được Đánh giá tổng quát về kết quả đạt được trong đồ án, như việc cải thiện hiệu năng của mô hình LLM Tiếng Việt thông qua kỹ thuật gợi ý, và hiệu năng của mô hình so với các mô hình ngôn ngữ khác.

5.3. Hạn chế và khó khăn trong quá trình thực hiện Nêu những hạn chế và khó khăn gặp phải trong quá trình thực hiện đồ

Tiếp tục

án, bao gồm thiếu dữ liệu đa dạng và đại diện cho ngôn ngữ Tiếng Việt, đánh giá chủ quan của chất lượng văn bản sinh ra, và khó khăn trong việc đánh giá hiệu năng của mô hình trong các tác vụ cụ thể.

5.4. Hướng phát triển trong tương lai 5.4.1. Bổ sung dữ liệu huấn luyện và kiểm thử Nâng cao chất lượng và đa dạng của dữ liệu huấn luyện và kiểm thử, giúp mô hình LLM Tiếng Việt đạt được hiệu năng tốt hơn và phản ánh đúng khả năng của ngôn ngữ Tiếng Việt.

5.4.2. Tối ưu hóa kỹ thuật gợi ý Tiếp tục nghiên cứu và tối ưu hóa các kỹ thuật gợi ý để cải thiện hiệu năng của mô hình LLM Tiếng Việt, đặc biệt là trong các tác vụ cụ thể.

5.4.3. Kết hợp với các phương pháp học sâu khác Kết hợp mô hình LLM Tiếng Việt với các phương pháp học sâu khác, như học chuyển tiếp, học đa nhiệm, v.v., để giải quyết các bài toán phức tạp và đa dạng hơn.

5.4.4. Ứng dụng trong các lĩnh vực mới Khám phá và áp dụng mô hình LLM Tiếng Việt trong các lĩnh vực mới như giáo dục, y tế, quản lý nguồn lực, v.v., để đẩy mạnh ứng dụng của trí tuệ nhân tạo trong đời sống xã hội.

5.5. Kết luận Kết thúc đồ án, nhấn mạnh tầm quan trọng của việc xây dựng mô hình LLM Tiếng Việt và áp dụng kỹ thuật gợi ý, cũng như đề xuất hướng phát triển trong tương lai để đạt được hiệu năng cao hơn và ứng dụng rộng rãi hơn trong thực tế.

Chương 6: Ứng dụng và tiềm năng của kỹ thuật gợi ý trong tương lai

6.1. Các ứng dụng thực tế của kỹ thuật gợi ý trong mô hình LLM

6.1.1. Trợ lý ảo

* Sử dụng kỹ thuật gợi ý trong mô hình LLM để cung cấp đáp án phù hợp với câu hỏi của người dùng, giúp trợ lý ảo hiểu và phản hồi chính xác hơn.
* Ứng dụng trong các hệ thống hỗ trợ khách hàng, hỗ trợ kỹ thuật, đặt lịch hẹn, quản lý công việc cá nhân, v.v.

6.1.2. Dịch máy tự động

* Kỹ thuật gợi ý giúp cải thiện chất lượng dịch thuật của mô hình LLM, đặc biệt là trong việc giữ ngữ nghĩa và cấu trúc ngữ pháp.
* Ứng dụng trong việc dịch các loại văn bản như sách, bài báo, tài liệu kỹ thuật, văn bản pháp lý, v.v.

6.1.3. Tạo nội dung cho bài viết, văn bản, tài liệu

* Sử dụng kỹ thuật gợi ý để tạo ra nội dung sáng tạo, phong phú và phù hợp với yêu cầu của người dùng.
* Ứng dụng trong việc soạn thảo bài viết, biên tập nội dung, viết quảng cáo, tạo kịch bản video, v.v.

6.2. Thách thức và hạn chế của kỹ thuật gợi ý

* Độ chính xác và tính nhất quán của kết quả gợi ý có thể chưa đạt mức mong muốn, đặc biệt trong các tình huống phức tạp hoặc đòi hỏi chuyên môn cao.
* Khả năng hiểu ngữ cảnh và tri thức thực tế của mô hình LLM có giới hạn, dẫn đến những gợi ý không chính xác hoặc thiếu tính ứng dụng.

6.3. Tiềm năng phát triển trong tương lai

* Nghiên cứu và áp dụng các kỹ thuật gợi ý tiên tiến hơn, kết hợp với kiến thức chuyên môn và ngữ cảnh cụ thể để cải thiện chất lượng gợi ý của mô hình LLM.
* Tích hợp mô hình LLM với các công nghệ khác như xử lý ảnh, giọng nói, dữ liệu cấu trúc, v.v., để phát triển các hệ thống gợi ý đa mô hình và đa kênh.
* Tận dụng trí tuệ nhân tạo và học máy để phát hiện và hiểu các xu hướng, thói quen và nhu cầu của người dùng, giúp tăng độ chính xác và tính cá nhân hóa của các gợi ý.
* Nghiên cứu về đạo đức, tính công bằng và bảo mật trong kỹ thuật gợi ý, để đảm bảo rằng các ứng dụng và dịch vụ sử dụng kỹ thuật gợi ý tuân thủ các nguyên tắc đạo đức và bảo vệ quyền riêng tư của người dùng.
* Khám phá các ứng dụng mới và tiềm năng của kỹ thuật gợi ý trong các lĩnh vực chưa được khai thác, như giáo dục, y tế, chính phủ, v.v., để mở rộng phạm vi ảnh hưởng và tạo ra giá trị xã hội.

Tóm lại, chương 6 đề cập đến các ứng dụng thực tế và tiềm năng của kỹ thuật gợi ý trong mô hình LLM, cũng như các thách thức và hạn chế hiện tại của kỹ thuật này. Đồng thời, chương này cũng đề xuất các hướng phát triển trong tương lai, từ việc cải tiến kỹ thuật gợi ý, áp dụng đa mô hình và đa kênh, đến việc tập trung vào các vấn đề đạo đức, công bằng và bảo mật. Những nỗ lực này sẽ giúp kỹ thuật gợi ý trong mô hình LLM đạt được những thành tựu lớn hơn trong tương lai, góp phần tạo ra giá trị thực tiễn và đẩy mạnh sự phát triển của ngành công nghệ thông tin.

Chương 7: Kết luận và kiến nghị

7.1. Kết luận chung: Sau quá trình nghiên cứu và xây dựng mô hình LLM với ngôn ngữ đích là Tiếng Việt, chúng ta đã hiểu rõ hơn về các khía cạnh cơ bản của xử lý ngôn ngữ tự nhiên, mô hình ngôn ngữ lớn, cũng như các kỹ thuật gợi ý trong LLM. Bằng cách áp dụng các phương pháp nghiên cứu và đánh giá, chúng ta đã xây dựng và tinh chỉnh mô hình LLM Tiếng Việt có độ chính xác và hiệu quả cao.

7.2. Đóng góp của đề tài: Đề tài đưa ra các đóng góp sau:

* Đề xuất và xây dựng một mô hình LLM Tiếng Việt hiệu quả, có khả năng xử lý và gợi ý nội dung Tiếng Việt một cách chính xác và linh hoạt.
* Đánh giá và so sánh hiệu năng của mô hình LLM Tiếng Việt với các mô hình LLM khác trên thế giới, đồng thời chỉ ra ưu điểm và hạn chế của mô hình.
* Khám phá các ứng dụng thực tế của kỹ thuật gợi ý trong mô hình LLM Tiếng Việt, cũng như tiềm năng phát triển trong tương lai.

7.3. Hướng phát triển và nghiên cứu tiếp theo:

Có một số hướng phát triển và nghiên cứu tiếp theo cho đề tài này:

* Nâng cao chất lượng và độ chính xác của mô hình LLM Tiếng Việt bằng cách tìm hiểu và áp dụng các phương pháp huấn luyện và tinh chỉnh mới, đồng thời mở rộng bộ dữ liệu huấn luyện.
* Tìm hiểu và áp dụng các giải pháp đạo đức, bảo mật và kiểm soát nội dung trong mô hình LLM Tiếng Việt, nhằm đảm bảo sử dụng mô hình an toàn và phù hợp với yêu cầu của người dùng.

Mở rộng phạm vi ứng dụng của mô hình LLM Tiếng Việt trong các lĩnh vực khác nhau, đặc biệt là trong giáo dục, y tế, kinh doanh, v.v., nhằm tạo ra những giá trị thực tiễn và tác động tích cực đến cuộc sống và công việc của con người.

* Kết hợp mô hình LLM Tiếng Việt với các mô hình LLM khác để tạo ra một hệ thống đa ngôn ngữ, giúp người dùng có thể giao tiếp và trao đổi thông tin dễ dàng hơn giữa các ngôn ngữ khác nhau.
* Nghiên cứu và phát triển các ứng dụng dựa trên mô hình LLM Tiếng Việt, chẳng hạn như các ứng dụng dạy học, tư vấn, hỗ trợ khách hàng, phân tích dữ liệu, dự đoán xu hướng, v.v., để mang lại lợi ích cho người dùng và cộng đồng.
* Tìm hiểu về các tiếp cận và kỹ thuật mới trong lĩnh vực xử lý ngôn ngữ tự nhiên, như các mô hình học sâu, học máy, học tăng cường, v.v., để nâng cao hiệu quả và khả năng của mô hình LLM Tiếng Việt.
* Cuối cùng, tiếp tục nghiên cứu về các vấn đề lý thuyết và thực tiễn liên quan đến mô hình LLM, nhằm đưa ra các kiến nghị và hướng dẫn sử dụng mô hình một cách hiệu quả và an toàn, đáp ứng nhu cầu và yêu cầu của người dùng.

Chương 8: Phụ lục

8.1. Tài liệu tham khảo chi tiết Cung cấp danh sách chi tiết các tài liệu tham khảo đã được sử dụng trong đồ án, bao gồm các bài báo khoa học, sách, báo cáo, tài liệu trực tuyến và các nguồn khác.

8.2. Mã nguồn Giới thiệu mã nguồn của đồ án, cung cấp liên kết đến mã nguồn trên các nền tảng như GitHub hoặc GitLab. Mô tả cách sử dụng mã nguồn, các yêu cầu về môi trường và các thư viện cần thiết để chạy mã nguồn.

8.3. Hướng dẫn sử dụng mô hình Cung cấp hướng dẫn chi tiết về cách sử dụng mô hình LLM Tiếng Việt, bao gồm cách thức nhập liệu, các thông số cần thiết và cách hiểu kết quả đầu ra. Nêu các ví dụ minh họa để giúp người dùng dễ dàng tiếp cận và sử dụng mô hình.

8.4. Kết quả thực nghiệm chi tiết Cung cấp kết quả chi tiết của các thực nghiệm đã được thực hiện trong đồ án, bao gồm số liệu, biểu đồ, bảng so sánh và nhận xét. Giải thích các kết quả và nhận xét về ảnh hưởng của các kỹ thuật gợi ý lên hiệu quả của mô hình LLM Tiếng Việt.

8.5. Tài liệu hỗ trợ Giới thiệu các tài liệu hỗ trợ khác liên quan đến đồ án, bao gồm hướng dẫn, tài liệu học tập, tài liệu tham khảo bổ sung, các công cụ phần mềm hỗ trợ và nguồn dữ liệu liên quan.