ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

~~~~~~*~~~~~~

BÁO CÁO ĐỒ ÁN MÔN HỌC

Final Semester Project

Môn học: HỌC THỐNG KÊ

Giảng viên lý thuyết: Thầy Lê Long Quốc
Giảng viên thực hành: Thầy Ngô Minh Nhựt

HỒ CHÍ MINH – Ngày 08 tháng 05 năm 2025
~~~~~~

Contents

Lời nói đầu

Trong bối cảnh toàn cầu hóa ngày càng sâu rộng, nhu cầu giao tiếp và trao đổi thông tin giữa các ngôn ngữ khác nhau trở nên vô cùng cấp thiết. Trong đó, dịch máy (Machine Translation) – một nhánh quan trọng của lĩnh vực xử lý ngôn ngữ tự nhiên (Natural Language Processing – NLP) – đóng vai trò ngày càng lớn trong việc hỗ trợ con người vượt qua rào cản ngôn ngữ. Với sự phát triển mạnh mẽ của các phương pháp học sâu (Deep Learning), các mô hình dịch máy hiện đại đã đạt được những bước tiến vượt bậc về chất lượng bản dịch, mở ra tiềm năng ứng dụng rộng lớn trong nhiều lĩnh vực như giáo dục, thương mại, văn hóa và công nghệ thông tin.

Đề tài này tập trung nghiên cứu và triển khai giải pháp dịch máy từ tiếng Anh sang tiếng Việt bằng cách ứng dụng mô hình học sâu T5 (Text-to-Text Transfer Transformer), một kiến trúc tiên tiến dựa trên mô hình Transformer do nhóm nghiên cứu của Google đề xuất. Mô hình T5 có khả năng xử lý nhiều tác vụ NLP khác nhau dưới dạng bài toán sinh văn bản, nhờ đó cho phép tận dụng tối đa sức mạnh của học sâu trong bài toán dịch ngôn ngữ.

Để huấn luyện mô hình, nhóm sử dụng bộ dữ liệu song ngữ *PhoMT* — một trong những tập dữ liệu lớn và tiêu chuẩn nhất hiện nay cho tác vụ dịch Anh – Việt. PhoMT bao gồm hàng trăm nghìn cặp câu song ngữ, được thu thập và xử lý kỹ lưỡng nhằm đảm bảo chất lượng ngữ nghĩa và tính phù hợp ngữ cảnh giữa hai ngôn ngữ. Việc lựa chọn PhoMT làm tập dữ liệu huấn luyện không chỉ đảm bảo độ tin cậy của mô hình mà còn giúp mô hình học được nhiều dạng cấu trúc câu và ngữ cảnh phong phú trong ngôn ngữ tiếng Việt.

Báo cáo này trình bày chi tiết về quá trình thực hiện đề tài, bao gồm mô tả bộ dữ liệu PhoMT, kiến trúc và phương pháp huấn luyện mô hình T5, cách thức phân chia dữ liệu (train/validation/test) cũng như các kết quả đánh giá mô hình bằng những độ đo phù hợp như BLEU, ROUGE và METEOR. Qua đó, báo cáo không chỉ minh họa tiềm năng của mô hình học sâu trong dịch máy Anh – Việt mà còn góp phần cung cấp thêm tư liệu tham khảo cho cộng đồng nghiên cứu và phát triển ứng dụng dịch máy trong nước.

Lời cảm ơn

Những kiến thức từ môn học **Học Thống Kê** đã được chúng em vận dụng hiệu quả vào đồ án này nhờ sự hướng dẫn tận tình của các thầy **Lê Long Quốc**, **Ngô Minh Nhựt**. Chúng em vô cùng biết ơn sự nhiệt huyết và tận tâm của các thầy, không chỉ trong việc hỗ trợ thực hiện đồ án mà còn trong việc cung cấp nền tảng kiến thức quý giá.

Bên cạnh đó, nhóm cũng đã tham khảo nhiều nguồn tài liệu trực tuyến như **GitHub**. Những thuật toán, bài viết và thảo luận trên các diễn đàn đã đóng góp không nhỏ vào quá trình hoàn thiện đồ án.

Thành phố Hồ Chí Minh, ngày 08 tháng 05 năm 2025

I. GIỚI THIỆU:

1.1 Thông tin thành viên nhóm:

1.Nguyễn Văn Phước - 22120285
2.Nguyễn Lê Anh Phúc -22120276
3. Nguyễn Hoài Phú - 22120269

1.2 Đánh giá mức độ hoàn thành yêu cầu:

Mức độ hoàn thành yêu cầu 100%

II. TỔNG QUAN ĐỀ TÀI:

Bài toán: Dịch máy Anh–Việt (hoặc Việt–Anh)
Dataset: *PhoMT* (Vietnamese-English Parallel Corpus)
Mô hình: *T5 (Text-To-Text Transfer Transformer)*
Công việc:

- Huấn luyện lại mô hình T5 trên PhoMT
- Đánh giá mô hình
- Viết báo cáo mô tả chi tiết

1. Mô tả tập dữ liệu PhoMT:

PhoMT là tập dữ liệu song ngữ Anh – Việt được xây dựng nhằm phục vụ nghiên cứu và phát triển hệ thống dịch máy. Bộ dữ liệu gồm khoảng **600.000 cặp câu song ngữ**, thu thập từ nhiều nguồn như văn bản chính phủ, tài liệu học thuật, tin tức, và tài liệu kỹ thuật. Các cặp câu đã được làm sạch và chuẩn hóa, đảm bảo chất lượng cao và độ đa dạng về chủ đề như kinh tế, công nghệ, y tế, xã hội. Dữ liệu có định dạng văn bản với cột tiếng Anh và cột tiếng Việt tương ứng, phù hợp để huấn luyện mô hình dịch máy học sâu như T5.

2. Mô tả mô hình T5:

- T5 (Text-to-Text Transfer Transformer) là mô hình học sâu do Google phát triển, dựa trên kiến trúc Transformer và được huấn luyện theo cách thống nhất mọi bài toán xử lý ngôn ngữ tự nhiên dưới dạng bài toán chuyển đổi văn bản thành văn bản.
- T5 có khả năng xử lý nhiều tác vụ như dịch máy, tóm tắt, trả lời câu hỏi bằng cùng một mô hình.

- Với kiến trúc encoder-decoder mạnh mẽ, T5 hỗ trợ học chuyển tiếp (transfer learning) hiệu quả, giúp mô hình thích ứng tốt với các tập dữ liệu mới như PhoMT để giải quyết bài toán dịch Anh – Việt.

3. Chuẩn bị dữ liệu:

- Bộ dữ liệu PhoMT được chia thành 3 tập:
  1) **Tập huấn luyện(Train):80%**
  2) **Tập kiểm tra(Validation):10%**
  3) **Tập kiểm thử(Test):10%**
- Làm sạch văn bản: Loại bỏ các câu trùng lặp, câu chứa ký tự không hợp lệ (ví dụ: ký hiệu lạ, emoji), và loại bỏ câu quá ngắn hoặc quá dài để đảm bảo tính đồng đều của dữ liệu.
- Chuẩn hóa văn bản: Áp dụng chuẩn hóa như chuyển về chữ thường, loại bỏ khoảng trắng thừa và chuẩn hóa dấu câu (ví dụ: dấu ngoặc kép, dấu chấm câu).
- Token hóa: Sử dụng tokenizer của mô hình T5 (T5Tokenizer) để mã hóa văn bản thành chuỗi token phù hợp với kiến trúc Transformer.

4. Huấn luyện mô hình:

5. Kết quả của huấn luyện:

6. Đáng giá mô hình:

~~~~~~~~~~HẾT~~~~~~~~~~