Mô hình dịch ngôn ngữ Việt-Lào

Phùng Quang Tiến
21020090

Đặng Nguyễn Duy Trúc
22021126

Kiều Minh Tuấn
21020394

***Tóm tắt nội dung*—Đây là báo cáo cho quá trình chuẩn bị dữ liệu, huấn luyện cho mô hình ngôn ngữ để dịch từ tiếng Việt sang tiếng Lào, dựa trên mô hình ngôn ngữ ViNMT.**

I. Giới thiệu

Trong những năm gần đây, dịch máy đã đạt được nhiều thành tựu nhờ sự phát triển mạnh mẽ của các mô hình học sâu, đặc biệt là kiến trúc Transformer. Tuy nhiên, phần lớn các thành tựu này tập trung vào các cặp ngôn ngữ có tài nguyên phong phú như Anh-Pháp, Anh-Đức, hoặc Anh-Trung. Trong khi đó, các cặp ngôn ngữ khu vực Đông Nam Á như Việt-Lào vẫn là một bài toán thách thức do thiếu dữ liệu song ngữ và tài nguyên ngôn ngữ học. Tiếng Việt và tiếng Lào là hai ngôn ngữ thuộc các hệ ngôn ngữ khác nhau, tuy nhiên chúng có nhiều điểm tương đồng về mặt ngữ pháp, cấu trúc câu và văn hoá giao tiếp, tạo ra tiềm năng lớn cho các ứng dụng dịch tự động. Mặc dù vậy, sự thiếu dữ liệu song ngữ chất lượng cao khiến cho việc huấn luyện các mô hình dịch hiệu quả trở nên khó khăn.

II. Training Data

Dữ liệu huấn luyện ban đầu bao gồm 100.000 câu song ngữ Việt Lào từ bộ dữ liệu trong cuộc thi VLSP 2023. Sau đó bộ dữ liệu được mở rộng ra:

- 20.000 câu song ngữ Việt-Lào từ bộ song ngữ Việt Lào từ bộ dữ liệu ALT.
- 200.000 câu song ngữ Việt-Lào sinh tự động.

Tất cả dữ liệu huấn luyện tổng cộng 320.000 câu song ngữ Việt Lào được đi qua tiền xử lý dữ liệu, áp dụng Byte-Pair Encoding cho ngôn ngữ đích, chuẩn hóa viết hoa, dấu câu cho ngôn ngữ nguồn.

III. Tổng quan mô hình

Mô hình dịch Việt Lào sẽ được xây dựng dựa trên mô hình đa ngôn ngữ ViNMT [1] (Vietnam Neural Machine Translation Toolkit). ViNMT được xây dựng dựa trên kiến trúc Transformer, đồng thời cung cấp một số cải tiến đáng chú ý như:

- Dễ dàng mở rộng và tùy biến: Nhờ cấu trúc mô-đun với các giao diện rõ ràng, người dùng có thể dễ dàng thay thế, kế thừa hoặc nâng cấp từng thành phần riêng lẻ trong mạng nơ-ron (như hàm kích hoạt, attention, positional encoding, v.v.) mà không ảnh hưởng đến toàn bộ hệ thống. Điều này đặc biệt hữu ích khi cần thử nghiệm các cải tiến mô hình.
- Giảm thiểu sự phụ thuộc và lỗi phát sinh: Mỗi mô-đun chỉ đảm nhiệm một chức năng duy nhất, tránh việc dồn nhiều trách nhiệm vào một lớp. Nhờ đó, việc bảo trì và sửa lỗi trở nên đơn giản.
- Tối ưu cho nghiên cứu ngôn ngữ tài nguyên thấp: Việc dễ dàng tích hợp hoặc thay đổi mô-đun giúp ViNMT phù hợp với các dự án thử nghiệm nhanh các cấu hình mô hình khác nhau, như dịch Việt-Lào, một cặp ngôn ngữ ít tài nguyên.Điều chỉnh kích thước từ vựng, số epoch... đều thực hiện thuận tiện.

Mô hình ViNMT được điều chỉnh và tối ưu để hỗ trợ hiệu quả hơn cho dịch từ tiếng Việt sang tiếng Lào.

IV. Thí nghiệm

Quá trình huấn luyện mô hình dịch Việt-Lào được thực hiện theo từng bước cải tiến liên tục, với mục tiêu nâng cao chất lượng bản dịch. Dưới đây tóm tắt các thay đổi cấu hình mô hình cùng với kết quả tương ứng, từ mô hình cơ sở đến cấu hình đã được tối ưu:

1) Mô hình cơ sở sử dụng tập huấn luyện nhỏ (100.000 cặp câu), vocab size 32.000. Kết quả BLEU khá thấp (5.18)
2) Tăng độ dài câu huấn luyện từ 50 lên 100 cho thấy hiệu quả rõ rệt, BLEU tăng lên 6.88. Mặc dù thời gian huấn luyện tăng (từ 7h đến 11h), việc mở rộng dữ liệu đã giúp mô hình học được nhiều đặc trưng hơn.
3) Giảm kích thước từ vựng từ 32.000 xuống 16.000 trong khi giữ nguyên tập dữ liệu huấn luyện tiếp tục nâng cao chất lượng bản dịch (BLEU = 12.44). Việc giảm kích thước từ vựng giúp mô hình tập trung vào những từ phổ biến, tránh phân mảnh ngữ liệu.
4) Với số lượng epoch mặc định là 20 chưa đủ để mô hình hội tụ Tiếp tục huấn luyện lên 50 epoch, mô hình đạt BLEU = 28.24 và sacreBLEU = 15.19
5) Hợp nhất dữ liệu với tập ALT (tổng cộng khoảng 120k câu) và huấn luyện trong 80 epoch giúp tăng nhẹ các chỉ số (BLEU = 29.97; sacreBLEU = 17.96; BLEURT = 0.2233). Mô hình giờ đây đã có khả năng tổng quát hóa tốt hơn trên nhiều cấu trúc ngữ pháp.
6) Thay hàm kích hoạt từ ReLU ở Feed Forward layer sang GELU đem lại cải thiện nhẹ nhưng ổn định (BLEU = 30.56, sacreBLEU = 19.11). Dù BLEURT giảm nhẹ (0.2066 so với 0.2233), kết quả BLEU tăng cho thấy cải thiện về độ chính xác của bản dịch.

V. Kết luận

Báo cáo đã trình bày quá trình xây dựng và cải tiến mô hình dịch tự động Việt–Lào dựa trên bộ công cụ ViNMT và kiến trúc Transformer đa ngôn ngữ. Bắt đầu từ một mô hình cơ sở

với chất lượng dịch còn hạn chế, qua từng bước cải thiện hiệu suất của mô hình thông qua nhiều hướng tiếp cận:

- Mở rộng quy mô dữ liệu huấn luyện và tinh chỉnh kích thước từ vựng phù hợp.
- Tăng số epoch huấn luyện nhằm tối ưu hóa quá trình học của mô hình.
- Hợp nhất thêm dữ liệu từ nguồn ALT để cải thiện khả năng tổng quát hóa.
- Thử nghiệm thay đổi hàm kích hoạt từ ReLU sang GELU để tăng khả năng biểu diễn phi tuyến tính.
- Tuy nhiên vì hạn chế tài nguyên tính toán nên vẫn chưa có thí nghiệm khách quan nào chứng minh tính hiệu quả của phương pháp tăng cường dữ liệu.

Kết quả cho thấy các cải tiến này đã giúp nâng BLEU score từ 5.18 (baseline) lên đến 30.56, đồng thời sacreBLEU và BLEURT cũng được cải thiện đáng kể. Những kết quả đạt được là nền tảng quan trọng cho việc tiếp tục mở rộng sang các cặp ngôn ngữ ít tài nguyên khác trong tương lai, đồng thời có thể tích hợp vào các hệ thống phục vụ dịch thuật thực tế.

TÀI LIỆU

[1] "ViNMT: Neural Machine Translation Tookit" Nguyen Hoang Quan, Nguyen Thanh Dat, Nguyen Hoang Minh Cong, Nguyen Van Vinh, Ngo Thi Vinh, Nguyen Phuong Thai, and Tran Hong Viet