TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
CHƯƠNG TRÌNH CHẤT LƯỢNG CAO

ĐỀ CƯƠNG KHÓA LUẬN TỐT NGHIỆP

Tên đề tài:

XÂY DỰNG MÔ HÌNH DỊCH MÁY TỪ TIẾNG ANH SANG TIẾNG VIỆT DỰA TRÊN KỸ THUẬT GẮN THẺ DỮ LIỆU DỊCH NGƯỢC

Họ tên, chức danh GVHD: GV.TS. Ngô Huy Biên (Khoa Công nghệ Thông tin)

Sinh viên: Mai Thanh Bình (1753030) – Nguyễn Trần Tuấn Anh (1753028)

Thể loại KLTN: Nghiên cứu có ứng dụng demo.

Thời gian thực hiện: từ ngày 06/01/2021 đến ngày 15/07/2021

Nội dung KLTN:

1. **Giới thiệu về đề tài:**

Đối với sự phát triển mạnh của khoa học công nghệ, cùng nền kinh tế hiện nay, việc học tập, tiếp xúc với nguồn tài liệu, tri thức trở nên rất cần thiết. Sự phát triển nhanh về mạng máy tính, cơ sở hạ tầng mạng, con người ngày càng có nhiều điều kiện để tìm kiếm nguồn tri thức phong phú và đa dạng được thể hiện bởi nhiều ngôn ngữ khác nhau, đặc biệt là tiếng Anh. Tuy nhiên, nhu cầu đọc hiểu thông tin lại trở nên khó khăn vì rào cản ngôn ngữ, việc dịch một văn bản, tài liệu có chất lượng cao tốn rất nhiều thời gian và công sức. Đối với tiếng Việt, bài toán xây dựng một hệ dịch máy tự động từ tiếng Anh sang tiếng Việt cũng đang rất được quan tâm. Nhưng do sự thiếu thốn về kho dữ liệu tri thức, việc nghiên cứu dịch máy về tiếng Việt vẫn còn khá ít, nên vẫn chưa có một hệ dịch Anh-Việt nào cho chất lượng bản dịch tương đối tốt. Nắm được hạn chế trên, chúng em chọn đề tài này nhằm học hỏi, tìm hiểu và áp dụng kiến thức về dịch máy để xây dựng một mô

hình dịch máy từ tiếng Anh sang tiếng Việt với chất lượng bản dịch tương đối, hỗ trợ cho việc đọc hiểu.

2. **Phạm vi của đề tài:**
    - Mô hình chỉ dịch từ tiếng Anh sang tiếng Việt.
    - Mô hình chỉ sử dụng được khi có kết nối mạng internet.
    - Mô hình chỉ tập trung vào kỹ thuật gắn thẻ dữ liệu dịch ngược.
3. **Mục tiêu đề tài:**

Bản luận văn trình bày:

- Các mô hình phổ biến để xử lý việc dịch một văn bản từ tiếng Anh sang tiếng Việt.
- Chi tiết vấn đề gặp phải khi dịch máy và giải pháp được đề xuất bởi một mô hình được chọn làm chuẩn, ví dụ Cross-View Training.
- Chi tiết giải pháp cho cùng vấn đề trên của mô hình Tagged Back-Translation.
- Chi tiết các cải tiến đề xuất dựa trên mô hình Tagged Back-Translation.
- Các kết quả đạt được bằng thực nghiệm của (i) mô hình Cross-View Training, (ii) mô hình Tagged Back-Translation, và (iii) mô hình đề xuất.
- Chi tiết việc so sánh các kết quả đạt được bằng thực nghiệm của mô hình Cross-View Training với mô hình Tagged Back-Translation với mô hình đề xuất.

Phần mềm thu được gồm:

- Dữ liệu tạo thêm/chỉnh sửa được.
- Công cụ tạo ra để xử lý các tác vụ đặc thù.
- Mã nguồn viết/chỉnh được dùng huấn luyện và đánh giá mô hình.
- Các mô hình đào tạo được.
- Mã nguồn hệ thống demo việc sử dụng mô hình đề xuất.

Bài báo 4 trang tóm tắt lại luận văn (không bắt buộc).

4. **Cách tiếp cận dự kiến:**

Các nghiên cứu có liên quan trực tiếp đến đề tài:

- Semi-Supervised Sequence Modeling with Cross-View Training [1].
- Tagged Back-Translation [4].
- Understanding Back-Translation at Scale [5].

Các ứng dụng tương tự:

Google Translate

- Đây là một dịch vụ dịch máy đa ngôn ngữ bằng nơ-ron (Neural machine translation) miễn phí được phát triển và cung cấp bởi Google.
- Tính năng:
    - Google Translate cung cấp mức độ hỗ trợ khác nhau cho 103 ngôn ngữ, tối đa 5000 từ.
    - Giao diện dễ sử dụng, hỗ trợ đa nền tảng và có độ chính xác khá cao.
    - Miễn phí và có hỗ trợ giọng nói, dịch qua hình ảnh, video thời gian thực.
- Link tham khảo: https://translate.google.com

Microsoft Translator

- Đối thủ cạnh tranh xứng đáng với Google Translate, là một dịch vụ dịch máy đa ngôn ngữ miễn phí từ Microsoft (thông qua Bing Translator).
- Tính năng:
    - Hỗ trợ 54 ngôn ngữ, tối đa 5000 từ.
    - Giao diện dễ sử dụng, độ chính xác cao và hỗ trợ đa nền tảng (có hỗ trợ ứng dụng trên Apple Watch và Android Wear).
    - Hỗ trợ giọng nói, hình ảnh, có chế độ dịch đoạn hội thoại thời gian thực.
    - Miễn phí sử dụng.
- Link tham khảo: https://bing.com/translator

Viki Translator

- Đây là trang web dịch riêng biệt dành cho cặp ngôn ngữ Anh - Việt.
- Tính năng:
    - Hỗ trợ dịch cụm từ, đoạn văn từ tiếng Anh sang tiếng Việt và ngược lại, tối đa 2000 từ.

    - Giao diện phép dễ nhìn, có hỗ trợ web, tiện ích trên Chrome, Windows và ứng dụng Android.
    - Miễn phí sử dụng và hỗ trợ giọng nói.
- Link tham khảo: https://vikitranslator.com/

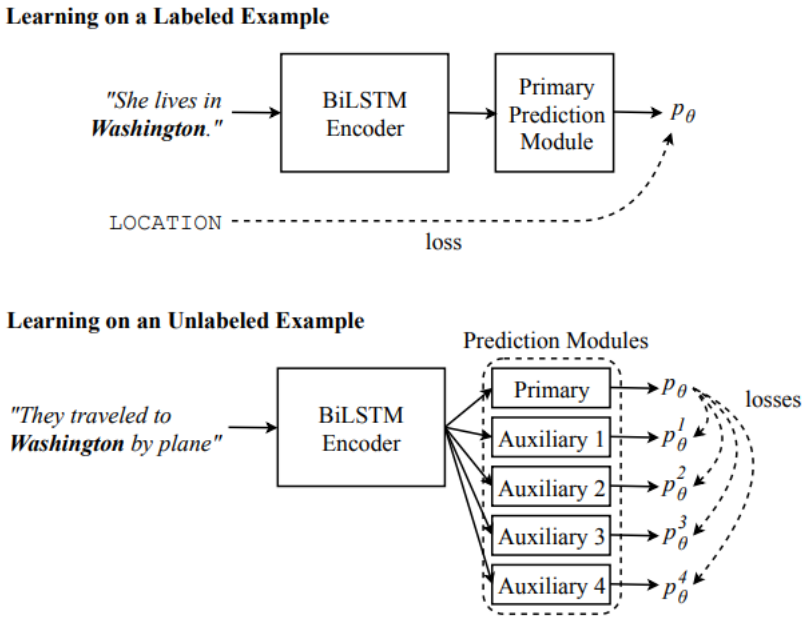
Hạn chế của các Ứng dụng dịch tiếng Anh sang tiếng Việt có sẵn:

- Tôi cần dịch một văn bản dài từ tệp tin .docx hoặc .pdf, nhưng ứng dụng hiện có không hỗ trợ dịch văn bản quá 5000 từ.
- Tôi cần dịch một từ hoặc một câu tiếng Anh có trên trang web đang được mở, nhưng ứng dụng hiện có hỗ trợ tiện ích dịch ngay trên trình duyệt web.
- Tôi muốn chất lượng bản dịch từ tiếng Anh sang tiếng Việt tốt hơn, nhưng ứng dụng hiện có chưa hỗ trợ dịch Anh – Việt tối đa.
- Tôi muốn dùng API dịch Anh – Việt cho ứng dụng cá nhân, nhưng ứng dụng hiện có chỉ cung cấp các dịch vụ có sẵn trên web hoặc ứng dụng của họ.

Kiến trúc trong đề tài này bao gồm một mô hình làm chuẩn là Cross-View Training, mô hình chính dựa trên Tagged Back-Translation và mô hình được cải tiến. Hai mô hình cùng giải quyết vấn đề về thiếu dữ liệu đào tạo song ngữ giữa hai cặp ngôn ngữ Anh-Việt, tăng chất lượng đào tạo để có chất lượng dịch máy tốt hơn.

Mô hình **Cross-View Training**:

- Là một thuật toán bán giám sát dùng để đào tạo các biểu diễn từ phân tán (distributed word representations) mà sử dụng các ví dụ có gắn nhãn hoặc không gắn nhãn.
- Trong CVT, mô hình luân phiên học trên một nhóm nhỏ các ví dụ được gắn nhãn và học trên một nhóm nhỏ các ví dụ không được gắn nhãn.

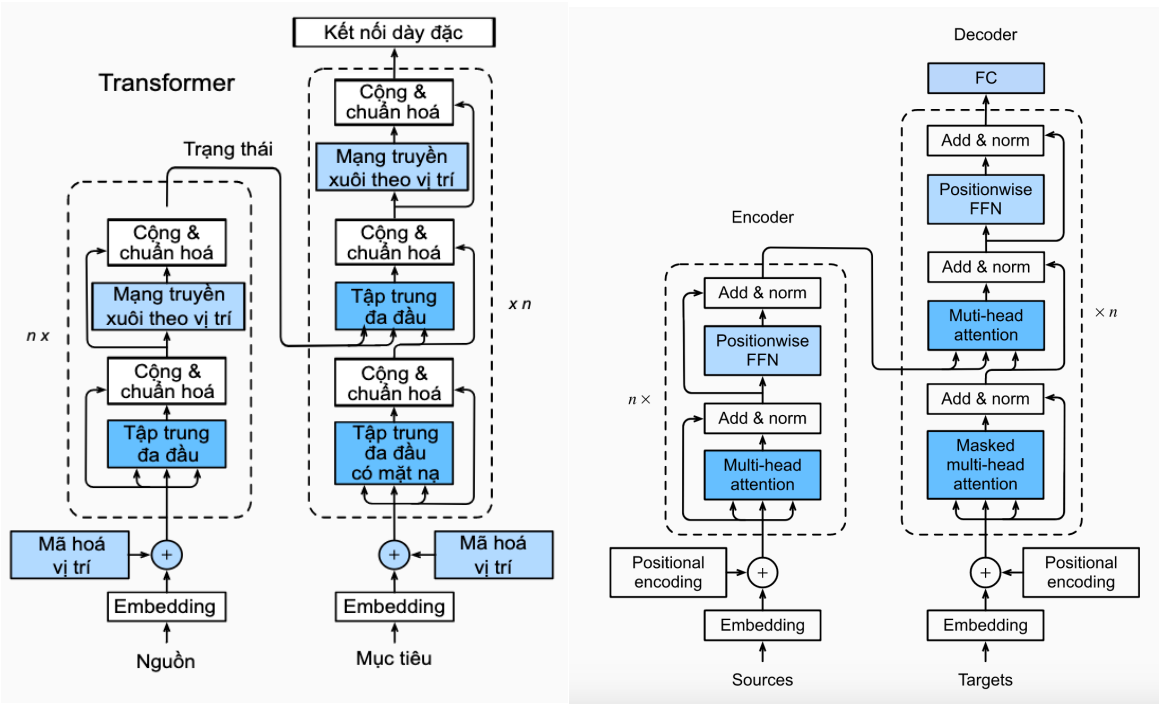


Tổng quan về Cross-View Training [5].

- Mô hình được đào tạo với việc học tập có giám sát trên các ví dụ được dán nhãn. Trên các ví dụ không được gắn nhãn, các mô-đun dự đoán phụ trợ với các quan điểm khác nhau của đầu vào được đào tạo để thống nhất với mô-đun dự đoán chính.

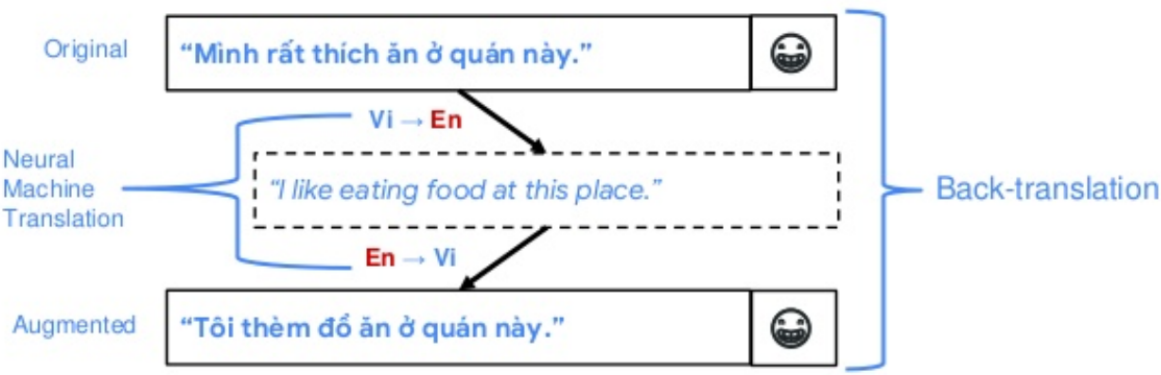
Với mô hình **Tagged Back-Translation**, chúng em sẽ chọn kiến trúc Transformer để triển khai:

- Transformer dựa trên kiến trúc mã hóa-giải mã (encoder-decoder), sử dụng cơ chế tập trung (self-attention) bằng các tầng tập trung đa đầu (multi-head attention), kết hợp thông tin vị trí thông qua biểu diễn vị trí (positional encoding) và áp dụng chuẩn hóa tầng (layer normalization).



Kiến trúc Transformer.

- Với nhiệm vụ ban đầu là xây dựng mô hình dịch máy từ Anh⟶Việt, sử dụng kiến trúc Transformer đào tạo hai mô hình dịch máy Anh⟶Việt và Việt⟶Anh với dữ liệu song ngữ có sẵn. Áp dụng kỹ thuật dịch ngược để dịch từ tiếng Việt sang tiếng Anh, tiếp tục dịch đầu ra từ tiếng Anh sang tiếng Việt (Việt⟶Anh⟶Việt).



- Đối với những chuỗi tiếng Anh có được từ quá trình dịch ngược sẽ được gắn thẻ bằng cách thêm mã thông báo đảo ngược (reversed token) <BT>. Kết hợp lại với tập dữ liệu song ngữ ban đầu, thu được tập dữ liệu đào tạo song ngữ Anh-Việt mới để đào tạo mô hình dịch máy Anh⟶Việt.
- Ví dụ: TaggedBT: <BT> Raise the child, love the child.

Các **nguồn dữ liệu** đào tạo có sẵn:

- Bộ dữ liệu IWSLT[1] (The International Workshop on Spoken Language Translation) English-Vietnamese được ra đời năm 2015. Được chia làm 3 phần với hơn 133 ngàn cặp câu dùng với mục đích huấn luyện, kiểm tra và đánh giá chất lượng mô hình.
- Bộ dữ liệu Binhvq News Corpus[2] của tác giả Vương Quốc Bình trích xuất từ khoảng 14.896.998 bài báo trên internet bao gồm nhiều phần khác như: Only Title (10.787.976 câu), Full TXT (title + description + body) (khoảng 111.274.300 câu) và đã được xử lý cơ bản.
- Bộ dữ liệu EVBCorpus[3] chứa hơn 20.000.000 từ (20 triệu) từ 15 cuốn sách song ngữ, 100 văn bản song song Anh-Việt / Việt-Anh, 250 văn bản luật và pháp lệnh song song, 5.000 bài báo và 2.000 phụ đề phim. Thành phần, chú thích, mã hóa và tính khả dụng của kho ngữ liệu nhằm tạo điều kiện thuận lợi cho sự phát triển của công nghệ ngôn ngữ và các nghiên cứu về trích xuất thuật ngữ song ngữ, chủ yếu cho cặp ngôn ngữ Anh-Việt-Anh.

Mã nguồn có sẵn:

- Pytorch-cvt[4]: mã nguồn mô hình Cross-View Training cho bài toán gán nhãn chuỗi (sequence labeling).

Tài nguyên GPU dự định sẽ dùng:

- Google Colab.
- Máy tính cá nhân GPU Geforce GTX 1650 4GB.

Bản mẫu website demo:

[1] https://github.com/stefan-it/nmt-en-vi
[2] https://github.com/binhvq/news-corpus
[3] https://sites.google.com/a/uit.edu.vn/hungnq/evbcorpus
[4] https://github.com/rezkaaufar/pytorch-cvt

Website Demo Mô Hình Dịch Máy Từ Tiếng Anh Sang Tiếng Việt Bằng Tagged Back-Translation

Tiếng Anh	Tiếng Việt
Nhập nội dung	

Dịch

Cách đánh giá mô hình dịch máy:

- BLEU là một phương pháp dùng để đánh giá chất lượng bản dịch được đề xuất trong công trình "BLEU: a method for Automatic Evaluation of Machine Translation" bởi Papineni K., Roukos S., Ward T., và Zhu Z-J vào tháng 7-2001. Ý tưởng chính của phương pháp là so sánh kết quả bản dịch tự động bằng máy với một bản dịch chuẩn dùng làm bản đối chiếu. Việc so sánh được thực hiện thông qua việc thống kê sự trùng khớp của các từ trong hai bản dịch có tính đến thứ tự của chúng trong câu (phương pháp n-grams theo từ) được nêu ra trong công trình "Toward finely differentiated evaluation metrics for machine translation" bởi Hovy E.H vào năm 1999. Phương pháp này dựa trên hệ số tương quan giữa bản dịch máy và bản dịch chính xác được thực hiện bởi con người để đánh giá chất lượng của một hệ thống dịch. Nhóm chúng em sẽ dùng BLEU để đánh giá và so sánh các mô hình với nhau.
- Khảo sát và đánh giá: thực hiện khảo sát chất lượng bản dịch bằng cách khảo sát người dùng, cho người dùng chấm điểm bản dịch được dịch bởi mô hình của nhóm và các hệ thống có sẵn trên thị trường (Google Translate, Microsoft Translator, Viki Translator).

5. **Kết quả dự kiến của đề tài:**
    - Mô hình dịch máy đề xuất từ tiếng Anh sang tiếng Việt dựa trên kỹ thuật gắn thẻ dữ liệu dịch ngược.
    - Mã nguồn huấn luyện mô hình và dữ liệu chỉnh sửa/tạo thêm.
    - Dịch vụ web API dịch máy văn bản từ tiếng Anh sang tiếng Việt dựa trên mô hình đã được đào tạo.
    - Trang web demo việc dịch văn bản từ tiếng Anh sang tiếng Việt sử dụng API đã xây dựng.

Các mốc thời gian nghiên cứu:

Thời gian	Công việc	Người thực hiện
15/12/2020 – 06/01/2021	Liên hệ giảng viên nhận đề tài. Xây dựng kế hoạch sơ bộ cho các công việc cần thực hiện.	Cả hai thành viên.
07/01/2021 – 31/01/2021	Tìm hiểu và phân tích các yêu cầu về kiến thức cho đề tài. Khảo sát và sử dụng các hệ thống cung cấp dịch vụ tương tự trên thị trường.	Cả hai thành viên.
02/02/2021 – 23/02/2021	Tìm hiểu kiến thức về máy học, dịch máy.	Cả hai thành viên.
24/02/2021 – 09/03/2021	Thu thập dữ liệu ngôn ngữ. Tìm hiểu các mô hình phổ biến để xử lý việc dịch một văn bản từ tiếng Anh sang tiếng Việt. Viết chương 1 luận văn.	Cả hai thành viên.
10/03/2021 – 16/03/2021	Viết đề cương luận văn tốt nghiệp. Tìm hiểu về hai mô hình Cross-View Training và Tagged Back-Translation.	Mai Thanh Bình. Cả hai thành viên.
17/03/2021 – 23/03/2021	Xây dựng và huấn luyện mô hình dịch máy dựa trên Cross-View Training.	Cả hai thành viên.
24/03/2021 – 06/04/2021	Viết chương 2 luận văn. Hoàn thiện mô hình Cross-View Training. Xây dựng mô hình Tagged Back-Translation.	Mai Thanh Bình. Cả hai thành viên.
07/04/2021 – 20/04/2021	Cải thiện và nâng cấp mô hình. Viết chương 3 luận văn.	Cả hai thành viên. Mai Thanh Bình.

21/04/2021 – 04/05/2021	Xây dựng và triển khai hệ thống dịch vụ cung cấp API dịch.	Cả hai thành viên.
05/05/2021 – 18/05/2021	Xây dựng trang web demo. Đề xuất và thử nghiệm các chiến lược khác nhau. Viết chương 4 luận văn.	Cả hai thành viên.
19/05/2021 – 01/06/2021	Kiểm tra lại mô hình và cải thiện hiệu năng của mô hình. Chỉnh sửa, cải thiện dịch vụ API và trang web demo.	Cả hai thành viên.
02/06/2021 – 15/06/2021	Hoàn chỉnh cuốn luận văn. Tổng hợp tài liệu. Chỉnh sửa và cải thiện các sản phẩm luận văn.	Cả hai thành viên
16/06/2021 – 29/06/2021	Hoàn thiện các sản phẩm luận văn.	Cả hai thành viên.
30/06/2021 – 14/07/2021	Hoàn chỉnh slide trình bày. Hoàn chỉnh sản phẩm khóa luận.	Cả hai thành viên.

Tài liệu tham khảo

[1] Isaac Caswell, Ciprian Chelba, David Grangier (2019). Tagged Back-Translation. arXiv, preprint arXiv:1906.06442.

[2] Benjamin Marie, Raphael Rubino, Atsushi Fujita (2020). Tagged Back-translation Revisited: Why Does It Really Work?. ACL 2020: 5990-5997.

[3] Rico Sennrich, Barry Haddow, Alexandra Birch. Improving Neural Machine Translation Models with Monolingual Data (2015). arXiv, preprint arXiv:1511.06709.

[4] Sergey Edunov, Myle Ott, Michael Auli, David Grangier. Understanding Back-Translation at Scale (2018). arXiv, preprint arXiv:1808.09381.

[5] Kevin Clark, Minh-Thang Luong, Christopher D. Manning, Quoc V. Le. Semi-Supervised Sequence Modeling with Cross-View Training (2018). arXiv, preprint arXiv:1809.08370.

Ý kiến của giảng viên hướng dẫn

Chữ ký của giảng viên hướng dẫn

TP. HCM, … / …/ …

Chữ ký (các) sinh viên