


THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút):
<https://youtu.be/y956Ce2aXAE>
- Link slides (dạng .pdf đặt trên Github của nhóm):
<https://github.com/anhquan075/CS519.M11.KHCL/blob/main/CS519.M11.DeCuong.FinalReport.AIO.pdf>
- *Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới*
- *Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in*

| | |
|--|---|
| <ul style="list-style-type: none">• Họ và Tên: Nguyễn Lê Anh Quân• MSSV: 19522081  | <ul style="list-style-type: none">• Lớp: CS519.M11.KHCL• Tự đánh giá (điểm tổng kết môn): 9/10• Số buổi vắng: 0• Số câu hỏi QT cá nhân: 10• Số câu hỏi QT của cả nhóm: 5• LinkGithub: https://github.com/anhquan075/CS519.M11.KHCL• Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:<ul style="list-style-type: none">○ Lên ý tưởng bài toán thực hiện○ Phân công công việc cho các thành viên trong nhóm.○ Viết phần giới thiệu bài toán và mục tiêu |
| <ul style="list-style-type: none">• Họ và Tên: Cát Văn Tài• MSSV: 19522147 | <ul style="list-style-type: none">• Lớp: CS519.M11.KHCL• Tự đánh giá (điểm tổng kết môn): 9/10• Số buổi vắng: 0• Số câu hỏi QT cá nhân: 10• Số câu hỏi QT của cả nhóm: 5 |

| | |
|---|--|
|  | <ul style="list-style-type: none"> • Link Github: https://github.com/vantai-nlp/CS519.M11.KHCL • Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm: <ul style="list-style-type: none"> ◦ Viết phần nội dung và phương pháp ◦ Làm slide và thuyết trình |
| <ul style="list-style-type: none"> • Họ và Tên: Nguyễn Khả Tiến • MSSV: 19522337  | <ul style="list-style-type: none"> • Lớp: CS519.M11.KHCL • Tự đánh giá (điểm tổng kết môn): 9/10 • Số buổi vắng: 0 • Số câu hỏi QT cá nhân: 10 • Số câu hỏi QT của cả nhóm: 5 • LinkGithub: https://github.com/TienNguyenKha/CS519.M11.KHCL • Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm: <ul style="list-style-type: none"> ◦ Tìm hiểu đề tài. ◦ Làm poster. ◦ Chỉnh sửa và hỗ trợ viết báo cáo. |

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI: TÌM HIỂU VÀ XÂY DỰNG HỆ THỐNG CHATBOT HỖ TRỢ HỌC NHẬP MÔN LẬP TRÌNH

TÊN ĐỀ TÀI TIẾNG ANH: RESEARCHING AND BUILDING CHATBOT SYSTEM FOR HELPING STUDY INTRODUCTION TO PROGRAMMING

Sinh viên thực hiện:

Nguyễn Lê Anh Quân - 1952081

Lớp: KHCL2019.3

Cát Văn Tài - 19522147

Lớp: KHCL2019.3

Nguyễn Khả Tiến - 19522337

Lớp: KHCL2019.3

Tóm tắt

Hiện nay, với những tiến bộ trong lĩnh vực xử lý ngôn ngữ tự nhiên (Natural Language Processing). Việc xây dựng các chatbot hỗ trợ học tập đã trở nên rất phổ biến. Song, vẫn tồn tại nhiều yếu tố cần quan tâm như thiếu bộ dữ liệu thích hợp, cấu trúc ngữ pháp tiếng Việt phức tạp v.v. Đặc biệt, các hệ thống chatbot hỗ trợ học lập trình vẫn còn rất ít. Vì những lý do trên, chúng tôi đề xuất xây dựng một hệ thống chatbot hỗ trợ học lập trình, cụ thể là Nhập môn lập trình. Bằng cách mô phỏng về công việc giảng dạy lập trình, chatbot của chúng tôi có thể mang lại cho những người mới học một phương thức học tập hiệu quả và tiết kiệm thời gian.

Giới thiệu

Chatbot - một chương trình máy tính được thiết kế để mô phỏng cuộc trò chuyện với người dùng, thông qua mạng Internet [1]. Ngày nay, với sự phát triển của các lĩnh vực như xử lý ngôn ngữ tự nhiên (NLP), thì Chatbot ngày càng được cải tiến và được áp dụng nhiều trong các hệ thống chăm sóc , tư vấn khách hàng, v.v [2]. Đồng thời trong quá trình học tập, việc hỏi đáp, tra cứu thông qua hệ thống Chatbot cũng đã được áp dụng vào một số môn học như Tiếng Anh, Toán học. Tuy nhiên đối với các môn học về lập trình hiện vẫn chưa được phổ biến. Vì vậy chúng tôi muốn nhân cơ hội này để xây dựng một hệ thống Chatbot như là một

“giáo viên ảo” có thể hỗ trợ các bạn sinh viên, học sinh học Nhập môn lập trình.

Việc xây dựng hệ thống này sẽ có thể giúp ích được rất nhiều cho các bạn mới bắt đầu học lập trình nhưng không biết nên học như thế nào, bắt đầu từ đâu, ... Tạo cảm giác thoải mái, tự tin khi đặt câu hỏi và tìm kiếm các kiến thức liên quan. Ngoài ra, trong [3], tác giả đã đề cập đến việc ứng dụng chatbot vào lĩnh vực giáo dục trong thời điểm việc học trực tuyến ngày càng phổ biến. Cụ thể là chatbot có thể giúp cung cấp những thông tin cần thiết trong quá trình học như gửi thông báo, nhắc nhở lịch thi cũng như giúp học sinh hiểu rõ hơn về chương trình giảng dạy.

Để xây dựng một hệ thống chatbot, chúng ta có rất nhiều framework hỗ trợ hiệu quả. RASA là một trong số đó, đây là một công cụ NLP open - source. Rasa có nhiều ưu điểm vượt trội và hiệu quả [4] với mục tiêu xây dựng của chúng tôi. Vì vậy trong đề tài này chúng tôi sẽ chọn RASA để xây dựng hệ thống với đầu vào và đầu ra như sau:

Input: câu hỏi từ người dùng về môn học Nhập môn lập trình.

Output: câu trả lời phù hợp với nhu cầu người dùng.

Tuy nhiên, việc áp dụng các kỹ thuật xử lý ngôn ngữ tự nhiên vào các hệ thống chatbot vẫn còn có nhiều hạn chế như ngôn ngữ tiếng Việt có cấu trúc ngữ pháp phức tạp, dữ liệu cho lĩnh vực này còn hạn chế dẫn đến áp dụng các kỹ thuật học sâu gặp nhiều thách thức. Do đó việc xây dựng một bộ dữ liệu riêng biệt là mục tiêu chính chúng tôi hướng đến. Bộ dữ liệu cho lĩnh vực này hiện tại vẫn chưa có cho tiếng Anh lẫn tiếng Việt. Việc thu thập bộ dữ liệu yêu cầu độ chính xác cao, đúng ngữ cảnh và phong cách nhắn tin của mọi người.

Mục tiêu

- Tìm hiểu tổng quan về hệ thống Chatbot và một số framework cho xây dựng chatbot, trong đó trọng tâm là RASA framework làm cơ sở để xây dựng một hệ thống Chatbot.
- Xây dựng bộ dữ liệu để phục vụ cho việc xây dựng hệ thống Chatbot tập trung vào chủ đề Nhập môn lập trình.
- Xây dựng ứng dụng minh họa.

Nội dung và phương pháp thực hiện

Để đạt được những mục tiêu, chúng tôi sẽ tiến hành nghiên cứu và tìm hiểu về:

- Tổng quan về Chatbot, đặc biệt là kiến trúc thành phần của hệ thống Chatbot nhằm để hiểu rõ về nguyên lý hoạt động của nó.
- Tìm hiểu về framework RASA, một framework khá phổ biến để xây dựng Chatbot, nhằm hiện thực hóa việc xây dựng Chatbot.
- Xây dựng bộ dữ liệu riêng về chủ đề Nhập môn lập trình bằng tiếng Việt nhằm để huấn luyện cho mô hình RASA Chatbot và từ đó xây dựng ứng dụng minh họa cho hệ thống.

Nội dung 1. Kiến trúc thành phần của hệ thống Chatbot



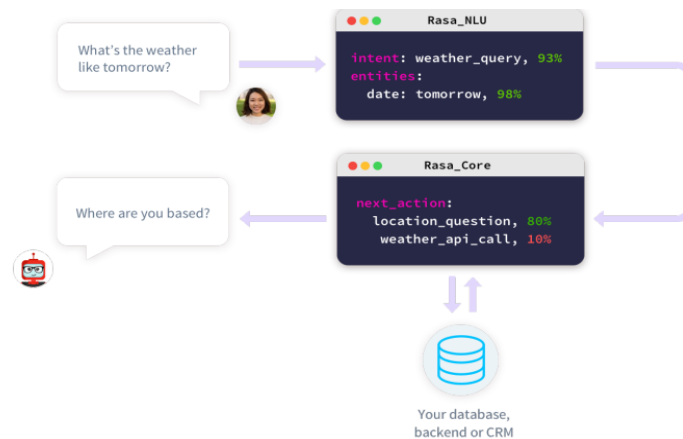
Hình 1. Cấu trúc thành phần cơ bản của hệ thống Chatbot

- Hiểu ngôn ngữ tự nhiên (Natural Language Understanding): Bao gồm việc xử lý ngôn ngữ tự nhiên (NLP) có nhiệm vụ xác định được ý định câu hỏi của người dùng (intent classification) và trích chọn thông tin (entity extraction).
- Quản lý hội thoại (Dialogue Management): Có nhiệm vụ xác định được hành động (action) tiếp theo dựa vào trạng thái hành động trước đó hoặc ngữ cảnh của cuộc hội thoại. Các ngữ cảnh này phải được tham chiếu trong các kịch bản dựng sẵn (history) được đào tạo cho Chatbot.
- Mô hình sinh ngôn ngữ (Natural Language Generation): Là thành phần sinh ngôn ngữ dựa vào chính sách (policy) và hành động được xác định trong DM thông qua các bộ hội thoại. NLG có thể sinh ra câu trả lời dựa vào tập mẫu câu trả lời (pre-defined template) đã đào tạo cho bot.

Nội dung 2. Framework RASA cho việc xây dựng Chatbot

Rasa gồm có 3 mô-đun chính:

- Rasa NLU: Đảm nhiệm các vấn đề liên quan đến Xử lý ngôn ngữ tự nhiên (NLP) như phân loại ý định (Intent Classification) và trích xuất thực thể (Entity Extraction). Nói cách khác, Rasa NLU dùng để hiểu tin nhắn từ người dùng.
- Rasa Core: Bộ khung xử lý các vấn đề xảy ra trong một cuộc hội thoại, Rasa Core sẽ quyết định Chatbot nên thực hiện hành động như thế nào.
- Rasa X: Một công cụ hỗ trợ triển khai mô hình Chatbot đã xây dựng thành sản phẩm thực tế, giúp tiếp cận nhanh chóng đến người dùng.



Hình 2. Ví dụ minh họa về quy trình xử lý của một hệ thống Rasa Chatbot

Mỗi mô-đun trên đều góp phần hình thành nên một bộ khung Rasa hoàn chỉnh. Tổng quan, một hệ thống Rasa Chatbot cơ bản sẽ có các quy trình thực hiện khi tham gia một cuộc đối thoại với người dùng như sau:

1. Tiếp nhận đoạn tin nhắn mà người dùng gửi.
2. Đưa đoạn tin nhắn vừa nhận được vào Rasa NLU để bắt đầu quá trình phân loại ý định và trích xuất thực thể.
3. Từ ý định và thực thể nhận được, bắt đầu xem xét các cách phản hồi phù hợp thông dựa trên các Mẫu hội thoại (Stories) đã được định nghĩa sẵn để từ đó đưa ra quyết định phản hồi tốt nhất.
4. Sau khi đã quyết định được cách trả lời, truy vấn câu trả lời tương ứng trong bộ trả lời đã được định nghĩa sẵn và gửi về hộp hội thoại để trả lời người dùng.

Nội dung 3. Xây dựng bộ dữ liệu phục vụ cho việc xây dựng chatbot Nhập môn lập trình và Ứng dụng minh họa

- Thu nhập dữ liệu từ các trang web tutorial uy tín về lập trình

- Xử lý dữ liệu để phù hợp với framework xây dựng Chatbot, framework chúng tôi sử dụng ở đây là RASA.
- Xây dựng hệ thống dựa theo các module và các phương pháp đã chọn lọc được.
- Đánh giá hiệu suất hệ thống.

Kết quả mong đợi:

- Tài liệu trình bày lại về thiết kế hệ thống Chatbot, thiết kế hệ thống cơ sở dữ liệu cho Chatbot và Framework sử dụng cho việc xây dựng hệ thống Chatbot.
- Tập dữ liệu gồm 6392 câu hội thoại, với mỗi câu trung bình từ 5 tới 7 token, dùng để huấn luyện cho hệ thống Chatbot chuyên về hỗ trợ môn học Nhập môn lập trình nói riêng và các lĩnh vực khác nói chung.
- Ứng dụng của hệ thống Chatbot với các khả năng:
 - Giải đáp các thắc mắc về C++ của người dùng.
 - Đưa ra các ví dụ và tìm kiếm những video tutorial liên quan đến thắc mắc của người dùng.
 - Gợi ý cách giải các bài tập sơ đồ khối.

Tài liệu tham khảo

- [1] Ruane, Elayne, Abeba Birhane, and Anthony Ventresque. “Conversational AI: Social and Ethical Considerations”, 2019.
- [2] M. Huang, X. Zhu, and J. Gao. “Challenges in Building Intelligent Open-domain Dialog Systems”. arXiv:1905.05709, 2020.
- [3] G. Molnar and Z. Szuts. “The Role of Chatbots in Formal Education”. In 2018 IEEE 16th International Symposium on Intelligent Systems and Informatics (SISY).
- [4] A. Jiao. “An Intelligent Chatbot System Based on Entity Extraction Using RASA NLU and Neural Network”. J. Phys.: Conf. Ser., vol.1487, p.012014, 2020.