

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ



NGUYỄN VĂN QUYỀN

XÂY DỰNG MÔ HÌNH BÁN HÀNG TỰ ĐỘNG
TRÊN INTERNET

Ngành: Công nghệ thông tin

Chuyên ngành: Hệ thống thông tin

Mã số: 14025060

TÓM TẮT LUẬN VĂN THẠC SĨ CÔNG NGHỆ THÔNG TIN

HÀ NỘI - 2016

MỤC LỤC

MỤC LỤC.....	1
DANH MỤC KÝ HIỆU VÀ CÁC CHỮ VIẾT TẮT	2
TÓM TẮT NỘI DUNG.....	3
GIỚI THIỆU CHUNG	4
CHƯƠNG 1: TỔNG QUAN BÁN HÀNG TỰ ĐỘNG	5
1.1 Các mô hình trả lời bán hàng tiêu biểu.....	5
1.2 Các vấn đề cần giải quyết và cải tiến	6
CHƯƠNG 2: MẠNG NƠ-RON VÀ PHƯƠNG PHÁP SEQ2SEQ ..	9
2.1 Lý thuyết mạng nơ-ron.....	9
2.2 Phương pháp học chuỗi Seq2Seq	12
CHƯƠNG 3: GIẢI PHÁP VÀ XÂY DỰNG MÔ HÌNH ĐỀ XUẤT ISALES	16
3.1 Phương án xây dựng mô hình đề xuất.....	16
3.2 Nhiệm vụ cụ thể của từng pha.....	17
3.2.1 Pha thu thập dữ liệu.....	17
3.2.2 Pha tiền xử lý dữ liệu.....	18
3.2.3 Pha phân mảnh dữ liệu	19
3.2.4 Pha huấn luyện dữ liệu.....	20
3.2.5 Pha sinh câu trả lời.....	22
CHƯƠNG 4: THỬ NGHIỆM, ĐÁNH GIÁ KẾT QUẢ	24
4.1 Phát biểu usecase	24
4.2 Thử nghiệm iSales.....	24
4.3 Đánh giá kết quả.....	27
KẾT LUẬN	29
TÀI LIỆU THAM KHẢO	30

DANH MỤC KÝ HIỆU VÀ CÁC CHỮ VIẾT TẮT

Từ viết tắt	Từ chuẩn	Diễn giải
Chatbot	Chatbot	Hệ thống trả lời tự động
Page	Page	Khái niệm trang thông tin trên facebook
Post	Post	Khái niệm bài viết trên facebook
Comment	Comment	Khái niệm bình luận trên facebook
NLP	Natural Language Processing	Xử lý ngôn ngữ tự nhiên
ANN	Artificial Neural Network	Mạng nơ ron nhân tạo
RNN	Recurrent Neural Network	Mạng nơ ron tái phát
LSTM	Long short-term memory	Mạng cải tiến để giải quyết vấn đề phụ thuộc quá dài
Seq2Seq	sequence to sequence	Phương pháp học chuỗi liên tiếp trong DeepLearning
Chatbot	Chatbot	Hệ thống trả lời tự động

TÓM TẮT NỘI DUNG

Trong những năm gần đây, thương mại điện tử bùng nổ mạnh, kèm theo đó là nhu cầu giải đáp khi mua hàng tăng cao. Điều này dẫn tới cần có một mô hình bán hàng, hỗ trợ người bán đưa ra các câu trả lời tự động, giảm thiểu công sức tư vấn cũng như tăng khả năng tương tác giữa người dùng và website thương mại điện tử.

Bên cạnh đó, trí tuệ nhân tạo ngày càng phát triển, dần đi vào các lĩnh vực ứng dụng đời sống hàng ngày. Với mục đích nghiên cứu một nhánh mới trong công nghệ học máy, tôi đã đề xuất và được phép nghiên cứu đề tài “***xây dựng mô hình bán hàng tự động trên Internet***”

Hiện tại, mô hình iSales được thiết kế dựa trên mạng nơ-ron, kết hợp phương pháp học chuỗi liên tiếp seq2seq, có khả năng hiểu Tiếng Việt, tự học từ các đoạn đối thoại được thu thập trên mạng hoặc cung cấp bởi người bán hàng và có thể sinh ra câu trả lời tự động. Mặc dù nghiên cứu hiện tại chưa thể đáp ứng cho sản phẩm thương mại nhưng iSales đã có một số kết quả nhất định, đặc biệt là ý nghĩa trong việc áp dụng phương pháp mới trong học máy.

GIỚI THIỆU CHUNG

Mạng xã hội ngày càng phát triển, và con người có thể bán hàng trực tiếp thông qua internet, tuy nhiên không phải lúc nào chúng ta cũng đủ nguồn nhân lực để sẵn sàng kết nối, tư vấn với khách hàng. Do đó, cần có một hệ thống trả lời và bán hàng tự động. Các hệ thống bán hàng tự động hiện nay chỉ dừng lại ở mức độ trả lời những câu hỏi đơn giản có sẵn, việc hỗ trợ Tiếng Việt không đầy đủ, khó khăn trong việc cải tiến. Những bất cập này làm cho việc vận hành và sử dụng hệ thống không mang lại nhiều lợi ích thiết thực.

Dựa vào mô hình nơ-ron và phương pháp học chuỗi liên tiếp Seq2seq, chúng tôi sẽ nghiên cứu và xây dựng mô hình trả lời tự động trên Internet cho Tiếng Việt để phục vụ riêng cho nghiệp vụ bán hàng. Mô hình cho phép học hỏi từ bộ dữ liệu sẵn có của doanh nghiệp mà không mất công sức xây dựng các mẫu hỏi đáp thủ công.

Cấu trúc luận văn được tổ chức như sau:

- Chương 1: Tổng quan bán hàng tự động
- Chương 2: Mạng nơ-ron và phương pháp seq2seq
- Chương 3: Giải pháp và xây dựng mô hình đề xuất iSales
- Chương 4: Thử nghiệm, đánh giá kết quả.

Cuối cùng, tôi xin đưa ra Kết luận trong quá trình nghiên cứu và xây dựng “mô hình bán hàng tự động trên Internet”.

CHƯƠNG 1: TỔNG QUAN BÁN HÀNG TỰ ĐỘNG

1.1 Các mô hình trả lời bán hàng tiêu biểu

“*Mô hình trả lời bán hàng*”, đây là một khái niệm có lẽ không còn mới. Tuy nhiên, thời điểm khái niệm này được đề cập nhiều có lẽ được bắt đầu khi ngành thương mại điện tử phát triển vượt bậc. Sự xuất hiện của Amazon, Skype, Facebook hoặc các số lượng lớn các website bán hàng trực tuyến kéo theo nhu cầu tương tác giữa người bán và người mua được nâng lên rõ rệt. Trong quá trình nghiên cứu, tôi đã tham khảo hai mô hình trả lời bán hàng khá nổi tiếng: **Messenger** của Facebook và **Chatbot Skype** của Microsoft. Về cơ bản, hai mô hình này đều chỉ “forward” thông tin và bắt buộc có sự xuất hiện của người bán hàng trong vai trò tư vấn. Mặc dù ở “mô hình bán hàng tự động” Skype của Microsoft đã rất tiến bộ và “thông minh” hơn so với Messenger của Facebook, nhưng nó vẫn còn có hạn chế trong việc cần nhập sẵn một số mẫu thông tin cố định để chatbot có thể truy xuất và trả lời thông qua những thông tin đó. Việc này có thể làm hạn chế khả năng giao tiếp của chatbot chỉ dừng lại ở một số câu chữ không thể có những đoạn hội thoại dài giống như giao tiếp giữa hai con người.

Việt Nam cũng bắt đầu phát triển mạnh nền thương mại điện tử trong một vài năm trở lại đây, đúng theo xu thế của thế giới. Với lợi thế không mất phí, có thể sử dụng trên môi trường internet dù ở bất kỳ đâu, các mô hình này dần chiếm được cảm tình của người sử dụng, giúp cho việc tương tác giữa đôi bên thuận lợi hơn. **uhChat**, và **subiz**, hai sản phẩm của Việt Nam vẫn chưa “tự động” thật sự, chỉ “tự động” ở mức độ kết nối với người dùng và đưa ra một số câu chào hỏi theo mẫu. So sánh với nhau, mô hình **subiz** cũng được phát triển theo luồng tương tác

tương tự **uhChat**. Nhưng điểm khác biệt, subiz cho phép người dùng đặt một “thư viện câu trả lời”, với những mẫu câu trả lời có sẵn. Thư viện này được xây dựng bởi người bán hàng. Với những câu hỏi có nội dung tương tự nhau, người bán hàng sẽ lựa chọn câu trả lời trong thư viện để trả lời mà không cần gõ lại thông tin nhiều lần.

Tóm lại, các “mô hình trả lời bán hàng” hiện nay trên thế giới và Việt Nam, đa số vẫn chỉ dừng lại ở mức ứng dụng chat, là forwarder trong phiên giao tiếp bán hàng, vẫn yêu cầu sự có mặt và xử lý của người bán hàng. Chatbot của Skype đã có sự tiến bộ hơn, đã có thể thay thế người bán hàng ở một mức độ nào đó, tuy nhiên cũng chỉ dừng lại ở những mẫu hội thoại ngắn và có nội dung ngắn.

1.2 Các vấn đề cần giải quyết và cải tiến

Trong quá trình tìm hiểu các mô hình trả lời bán hàng trên, dựa trên các tính năng hỗ trợ cũng như hạn chế của từng mô hình, ta có thể tạm phân loại các mô hình làm 3 nhóm.

Nhóm đầu tiên có thể gọi là nhóm “mô hình đơn giản”. Tiêu biểu cho mô hình này là Messenger của Facebook, uhChat....Ở những mô hình này, vai trò quyết định đưa ra câu trả lời là người bán hàng. Người bán hàng sẽ phải tiếp nhận câu hỏi, phân tích và đưa ra câu trả lời phụ thuộc vào kinh nghiệm. Họ cần online, theo dõi màn hình, ai yêu cầu thì tư vấn. Ở đây, ứng dụng chat được tích hợp vào chỉ mang tính chất forwarder. Điểm mạnh của những mô hình này là dễ sử dụng, đơn giản trong việc tích hợp trên website thương mại đơn giản và thường hỗ trợ nhiều tính năng. Tuy nhiên, điểm yếu của các mô hình này là không tự động, ta sẽ không hướng đến nhóm này. Nhóm thứ 2 là nhóm “Mô hình bán tự động” gồm subiz và các mô hình

tương tự. Với một số mẫu dữ liệu cài sẵn được nhập từ người bán hàng, và được lưu trữ vào cơ sở dữ liệu. Nếu có câu hỏi, ứng dụng chat sẽ phân tích và tìm kiếm trong tập mẫu, nếu có sẽ sử dụng làm câu trả lời, nếu không tìm thấy thì trả lại một câu mặc định nào đó. Khi đó, sự xuất hiện của người bán hàng là cần thiết để đưa ra câu trả lời hợp lý. Các thuật toán sử dụng cho mô hình này thường là thuật toán mapping. Ưu điểm lớn nhất là mô hình có thể trả lời những câu hỏi một cách hợp lý nếu tìm thấy câu trả lời trong tập mẫu. Tuy nhiên, mô hình này mang tính chất “học vẹt”. Nhóm thứ 3, là nhóm có trí thông minh hơn hẳn 2 nhóm trước. Nhóm “mô hình tự động” như chatbotSkype. Ở các mô hình tự động này, người bán hàng không cần tham gia phiên tư vấn trả lời, chatbot sẽ tự động phân tích và đưa ra câu trả lời. Điều bắt buộc ở các mô hình này là cần có sự xuất hiện của AI (Artificial intelligence), cho phép “hiểu” ngữ cảnh câu hỏi và đưa ra câu trả lời sau khi phân tích dữ liệu đầu vào và dữ liệu mẫu.

Trong 3 nhóm nêu trên, mô hình tự động là mô hình mà chúng ta sẽ nghiên cứu và hướng đến. Tuy nhiên, nếu dừng lại ở mức độ chatbot Skype thì còn có khá nhiều nhược điểm và vấn đề cần cải tiến

Thứ nhất, vấn đề phát sinh những câu hỏi nằm ngoài “tập dữ liệu mẫu”, mô hình sẽ không trả lời, hoặc trả lời những câu mẫu được định nghĩa sẵn cho những trường hợp này.

Thứ hai, vấn đề xây dựng một mô hình có thể hiểu được ngôn ngữ tiếng Việt, mô hình chatbot Skype hoàn toàn không có khả năng này.

Thứ ba, vấn đề câu hỏi dài và phức tạp, các mô hình hiện tại đều gặp vấn đề này. Ở chatbot Skype đang dừng lại mô

hình đưa ra gợi ý ngắn gọn cho người mua hàng lựa chọn, những đoạn đối thoại sinh ra thường ngắn.

Thứ tư, vấn đề tái sử dụng dữ liệu tư vấn: dữ liệu tư vấn giữa người mua hàng và người bán hàng được hầu hết các mô hình nêu trên lưu trữ lại. Tuy nhiên chỉ đáp ứng cho mục tiêu thống kê và báo cáo thì khá phí phạm, vì đây là tập dữ liệu lớn và chuẩn, có thể tái sử dụng trong các bài toán Machine Learning, Big Data.

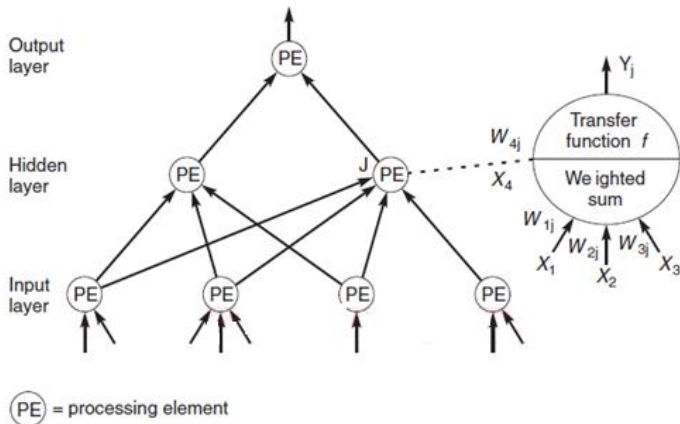
Cuối cùng là vấn đề “tự động”: là vấn đề mấu chốt của luận văn, không chỉ dừng lại ở mức độ tự động trả lời với những câu hỏi – câu trả lời có sẵn trong cơ sở dữ liệu, mà mô hình cần tự động trong quá trình “phân tích” câu hỏi, quá trình sinh mới câu trả lời nằm ngoài dữ liệu có sẵn. Nói cách khác, mô hình có thể “suy nghĩ” và “tư vấn” như một con người, thay thế sự xuất hiện của người bán hàng trong pha tư vấn.

Rõ ràng, chúng ta cần thay đổi phương pháp sử dụng Machine Learning để cho mô hình tự học được một lượng lớn dữ liệu từ người dùng, tạo nên một bộ “tri thức” và ứng dụng trong quá trình phân tích và sinh câu trả lời, thì những vấn đề nêu trên có thể được giải quyết.

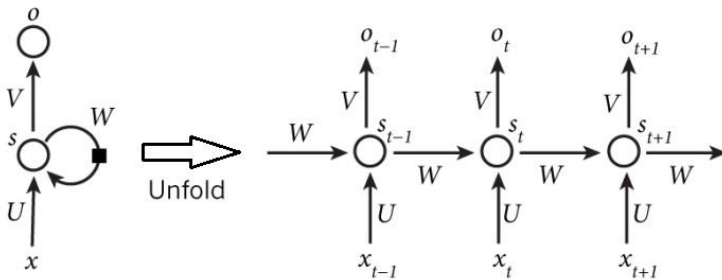
CHƯƠNG 2: MẠNG NƠ-RON VÀ PHƯƠNG PHÁP SEQ2SEQ

2.1 Lý thuyết mạng nơ-ron

Mạng nơ-ron nhân tạo (Artificial Neural Network – ANN) hay thường được gọi ngắn gọn là mạng nơ-ron là một mô hình xử lý thông tin được mô phỏng dựa trên hoạt động của hệ thống thần kinh của sinh vật, bao gồm số lượng lớn các nơ-ron được gắn kết để xử lý thông tin. ANN được giới thiệu năm 1943 bởi nhà thần kinh học Warren McCulloch và nhà logic học Walter Pits, nó hoạt động giống như bộ não của con người, được học bởi kinh nghiệm (thông qua việc huấn luyện), có khả năng lưu giữ các tri thức và sử dụng các tri thức đó trong việc dự đoán các dữ liệu chưa biết (unseen data) [5]. Processing Elements (PE) của ANN gọi là nơ-ron, nhận các dữ liệu vào (inputs) xử lý chúng và cho ra một kết quả (output) duy nhất. Kết quả xử lý của một nơ-ron có thể làm input cho các nơ-ron khác.

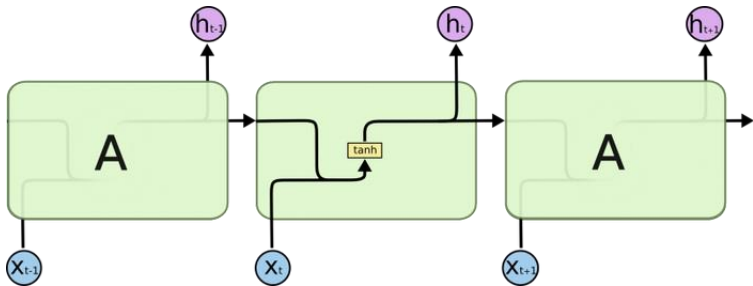


Mạng nơ-ron tái phát (RNN - Recurrent Neural Network) là một trong những mô hình Deep Learning được đánh giá có nhiều ưu điểm trong các tác vụ xử lý ngôn ngữ tự nhiên (NLP). Ý tưởng của RNN đó là thiết kế một mạng nơ-ron sao cho có khả năng xử lý được thông tin dạng chuỗi như câu hỏi. Recurrent có nghĩa là thực hiện lặp lại cùng một tác vụ cho mỗi thành phần trong chuỗi. Trong đó, kết quả đầu ra tại thời điểm hiện tại phụ thuộc vào kết quả tính toán của các thành phần ở những thời điểm trước đó. Nói cách khác, RNNs là một mô hình có trí nhớ, có khả năng nhớ được thông tin đã tính toán trước đó, không như các mạng nơ-ron truyền thống đó là thông tin đầu vào (input) hoàn toàn độc lập với thông tin đầu ra (output). Về lý thuyết, RNNs có thể nhớ được thông tin của chuỗi có chiều dài bất kì, nhưng trong thực tế mô hình này chỉ nhớ được thông tin ở vài bước trước đó.

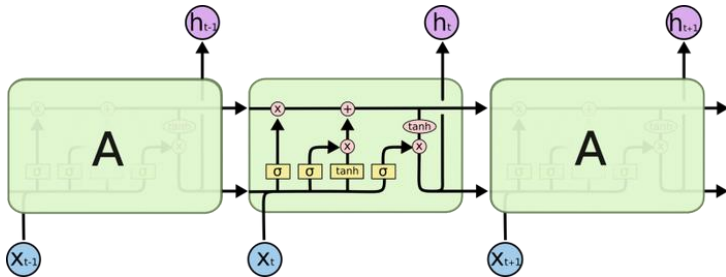


Long Short Term Memory là trường hợp đặc biệt của RNN, có khả năng học long-term dependencies. Mô hình này được giới thiệu bởi Hochreiter & Schmidhuber (1997) [8], và được cải tiến lại. Sau đó, mô hình này dần trở nên phổ biến nhờ vào các công trình nghiên cứu gần đây. Mô hình này có khả năng tương thích với nhiều bài toán nên được sử dụng rộng rãi ở các ngành liên quan. LSTM được thiết kế nhằm loại bỏ vấn

đề phụ thuộc quá dài. Ta quan sát lại mô hình RNN bên dưới, các layer đều mắc nối với nhau. Trong RNN chuẩn, module repeating này có cấu trúc rất đơn giản chỉ gồm một lớp đơn giản tanh layer.



LSTM cũng có cấu trúc mắt xích tương tự, nhưng các module lặp có cấu trúc khác hẳn. Thay vì chỉ có một layer neural network, thì LSTM có tới bốn layer, tương tác với nhau theo một cấu trúc cụ thể.

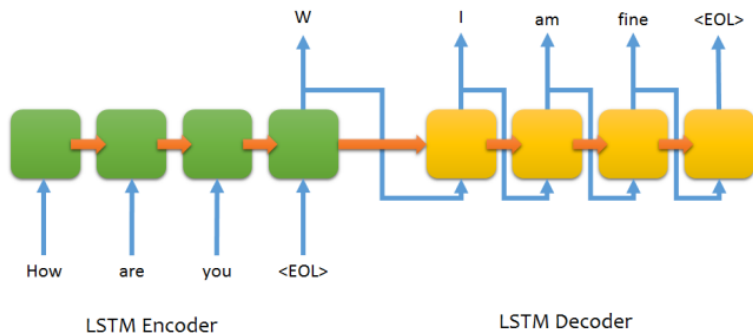


Trong đó, các ký hiệu sử dụng trong mạng LSTM gồm có: hình chữ nhật là các lớp ẩn của mạng nơ-ron, hình tròn biểu diễn toán tử Pointwise, đường kẻ gộp lại với nhau biểu thị phép nối các toán hạng, và đường rẽ nhánh biểu thị cho sự sao chép từ vị trí này sang vị trí khác. Mô hình thiết kế của LSTM là một bảng mạch số, gồm các mạch logic và các phép toán logic trên

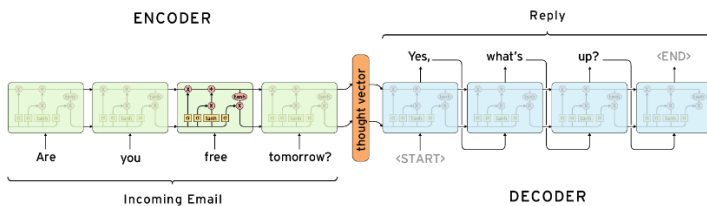
đó. Thông tin, hay nói khác hơn là tần số của dòng điện di chuyển trong mạch sẽ được lưu trữ, lan truyền theo cách thiết kế bảng mạch. Mấu chốt của LSTM là cell state (trạng thái nhớ), đường kẻ ngang chạy dọc ở trên top diagram. Cell state giống như băng chuyền, chạy xuyên thẳng toàn bộ mắc xích, chỉ một vài tương tác nhỏ tuyến tính (minor linear interaction) được thực hiện. Điều này giúp cho thông tin ít bị thay đổi xuyên suốt quá trình lan truyền.

2.2 Phương pháp học chuỗi Seq2Seq

Bản thân mô hình seq2seq [6] bao gồm hai mạng RNN: Một cho bộ mã hóa, và một cho bộ giải mã. Bộ mã hóa nhận một chuỗi (câu) đầu vào và xử lý một phần tử (từ trong câu) tại mỗi bước. Mục tiêu của nó là chuyển đổi một chuỗi các phần tử vào một vectơ đặc trưng có kích thước cố định mà nó chỉ mã hóa thông tin quan trọng trong chuỗi và bỏ qua các thông tin không cần thiết. Có thể hình dung luồng dữ liệu trong bộ mã hóa dọc theo trục thời gian, giống như dòng chảy thông tin cục bộ từ một phần tử kết thúc của chuỗi sang chuỗi khác.



Mỗi trạng thái ẩn ảnh hưởng đến trạng thái ẩn tiếp theo và trạng thái ẩn cuối cùng được xem như tích lũy tóm tắt về chuỗi. Trạng thái này được gọi là bối cảnh hay vector suy diễn, vì nó đại diện cho ý định của chuỗi. Từ bối cảnh đó, các bộ giải mã tạo ra một chuỗi, một phần tử (word) tại một thời điểm. Ở đây, tại mỗi bước, các bộ giải mã bị ảnh hưởng bởi bối cảnh và các phần tử được sinh ra trước đó.



Có một vài thách thức trong việc sử dụng mô hình này. Một trong những vấn đề đáng ngại nhất là các mô hình không thể xử lý được các chuỗi dài. Bởi vì hầu như tất cả các ứng dụng chuỗi sang chuỗi, bao gồm cả độ dài các chuỗi. Vấn đề tiếp theo là kích thước từ vựng. Bộ giải mã phải chạy hàm softmax hơn trên một tập rất lớn các từ vựng (khoảng 20,000 từ) cho mỗi một từ xuất ra. Điều này sẽ làm chậm quá trình huấn luyện, cho dù phần cứng có thể đáp ứng được khả năng xử lý. Đại diện của một từ là rất quan trọng. Làm thế nào để có thể biểu diễn được các từ trong chuỗi? Sử dụng one-hot vector có nghĩa là chúng ta phải đối mặt với các vector thưa thớt lớn, do kích thước vốn từ vựng lớn mà không có ý nghĩa về mặt ngữ nghĩa của từ được mã hóa bên trong các vector one-hot. Sau đây là cách giải quyết một số vấn đề mà chúng ta sẽ gặp phải.

PADDING – Tạo độ dài cố định

Trước khi huấn luyện, chúng ta cần chuyển đổi độ dài của các phần tử trong chuỗi thành các chuỗi có độ dài cố định, bằng việc thêm vào các phần tử đệm PADDING. Các phần tử đệm đặc biệt mà chúng ta sẽ sử dụng:

- **EOS:** Kết thúc câu (End of sentence)
- **PAD:** Phần đệm bù (Filler)
- **GO:** Bắt đầu giải mã (Start decoding)
- **UNK:** Unknown; từ không biết, không có trong từ điển từ vựng

BUCKETING – Tránh lu mờ thông tin

Bộ đệm đã giải quyết được vấn đề độ dài của các chuỗi, nhưng hãy xem xét một trường hợp các câu có độ dài lớn. Nếu câu dài nhất trong tập dữ liệu có độ dài là 100, chúng ta cần mã hóa tất cả các chuỗi còn lại bằng độ dài 100, để không mất thông tin của bất kỳ từ nào. Như vậy, chuyện gì xảy ra với chuỗi từ “How are you ?”. Sẽ có 97 phần tử đệm PAD được sử dụng khi mã hóa một chuỗi câu. Điều này sẽ làm lu mờ thông tin thực tế trong câu. Bucketing giải quyết vấn đề này bằng việc đặt các câu vào các xô buckets có kích thước khác nhau. Ví dụ ta có một danh sách các xô buckets: [(5, 10), (10, 15), (20, 25), (40, 50)]. Nếu độ dài của mẫu hỏi là 4 như ví dụ trên sẽ được đặt vào xô (5, 10). Mẫu hỏi sẽ được đệm với độ dài 5 và đáp án được đệm với độ dài 10. Trong lúc chạy mô hình (huấn luyện hoặc dự đoán), chúng ta sẽ sử dụng một mô hình khác cho mỗi bucket, tương ứng với các độ dài của mẫu hỏi và câu trả lời. Tất cả

những mô hình này chia sẻ các tham số giống nhau và do đó hoạt động chính xác theo cùng một cách.

Word Embedding – Mật độ dày đặc

Word Embedding là một kỹ thuật cho việc học mật độ dày đặc thông tin đại diện của từ trong một không gian vector có số chiều nhỏ hơn. Mỗi một từ có thể được xem như là một điểm trong không gian này, được đại diện bởi một vector có độ dài cố định. Word Embedding thường được thực hiện trong lớp đầu tiên của mạng: Trong đó lớp embedding sẽ ánh xạ một từ (chỉ số index của từ trong từ điển từ vựng) từ từ điển sang một vector dày đặc với kích thước đã cho. Trong mô hình seq2seq, trọng số của lớp embedding được huấn luyện giống nhau với các tham số khác của mô hình.

Qua chương 2, ta đã có cái nhìn tổng quan các khái niệm nền tảng, các công cụ cần thiết phục vụ cho xây dựng mô hình bán hàng tự động. Dựa trên những kiến thức trên, chương 3 sẽ đưa ra giải pháp và phương án thực hiện.

CHƯƠNG 3: GIẢI PHÁP VÀ XÂY DỰNG MÔ HÌNH ĐỀ XUẤT ISALES

3.1 Phương án xây dựng mô hình đề xuất

Mô hình bán hàng tự động được xây dựng cần qua 5 pha:



- Pha thu thập dữ liệu: sử dụng graphAPI để crawl dữ liệu từ page trên Facebook, trích xuất dữ liệu, chỉ lấy những comment trong các post nhằm có được dữ liệu thô cho quá trình huấn luyện. Kết quả pha được lưu trữ lại trong file text.
- Pha tiền xử lý dữ liệu: pha này cần đọc dữ liệu từ file kết quả của pha thu thập, tiến hành xóa bỏ những cặp câu bất thường/vô nghĩa, xóa bỏ những cặp câu bị trùng/có khoảng trắng và ghi kết quả vào file.
- Pha tiền phân mảnh dữ liệu: nhiệm vụ pha là đọc dữ liệu từ file kết quả của pha xử lý dữ liệu, tách từ theo nghĩa tiếng Việt. Lưu kết quả cho quá trình huấn luyện.
- Pha huấn luyện dữ liệu: cho phép đọc dữ liệu từ file kết quả của pha phân mảnh dữ liệu, bóc tách Word2Id và tạo ra từng cặp hội thoại, đưa kết quả vào mạng LSTM để huấn luyện.
- Pha sinh câu trả lời: nhận dữ liệu input là câu hỏi đầu vào, tiến hành phân tích câu hỏi dựa trên các ngữ cảnh

học được. Sử dụng mô hình sinh ra câu trả lời và hiển thị kết quả.

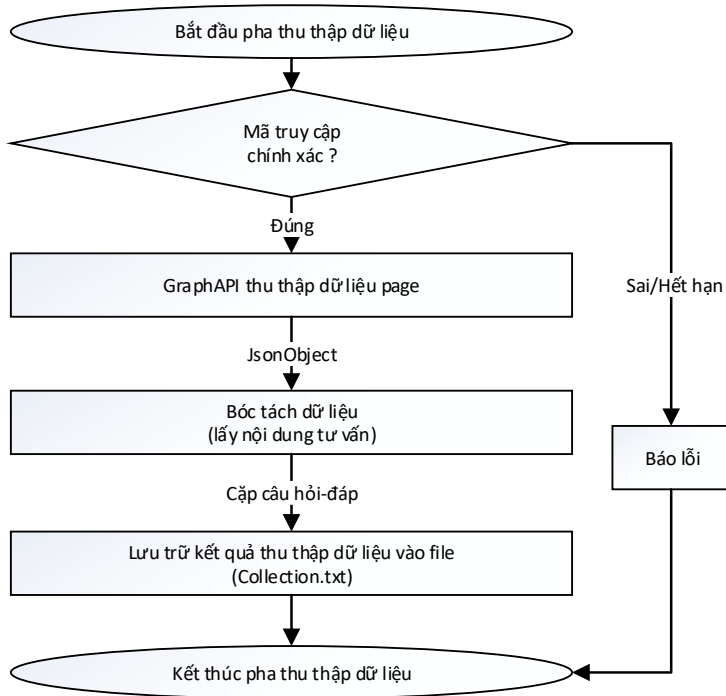
Dựa trên nghiệp vụ của từng pha, tôi sử dụng những công cụ sau để thực hiện xây dựng mô hình đề xuất:

- NLTK: Công cụ xử lý ngôn ngữ tự nhiên mã nguồn mở.
- VNTK: Vietnamese language toolkit, do chúng tôi xây dựng và phát triển để xử lý các vấn đề cơ bản của tiếng Việt.
- Dongdu: Thư viện hỗ trợ tách từ tiếng Việt [13], của tác giả Lưu Tuấn Anh
- Tensorflow: Framework machine learning
- Python: Ngôn ngữ lập trình để xây dựng mô hình đối thoại tiếng Việt.

3.2 Nhiệm vụ cụ thể của từng pha

3.2.1 Pha thu thập dữ liệu

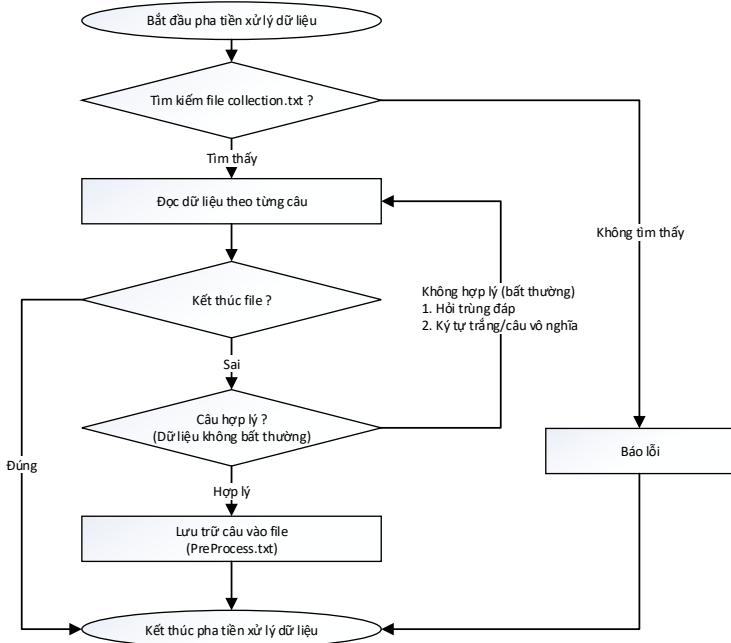
Sử dụng graphAPI là cần một mã truy cập (*user access token*). Người dùng muốn lấy được dữ liệu cần phải đăng ký trở thành nhà phát triển, mỗi khi sử dụng graphAPI, Facebook sẽ cấp một mã truy cập có thời hạn. Sau khi người dùng khai báo đúng mã truy cập trên ứng dụng, dữ liệu được thu thập được thông qua ứng dụng. Từ đó việc trích xuất và lưu trữ có thể thực hiện một cách đơn giản.



Kết quả mong muốn của pha thu thập dữ liệu sẽ là một tập tin chứa các cặp câu hỏi đáp theo đúng thứ tự tư vấn trên Page.

3.2.2 Pha tiền xử lý dữ liệu

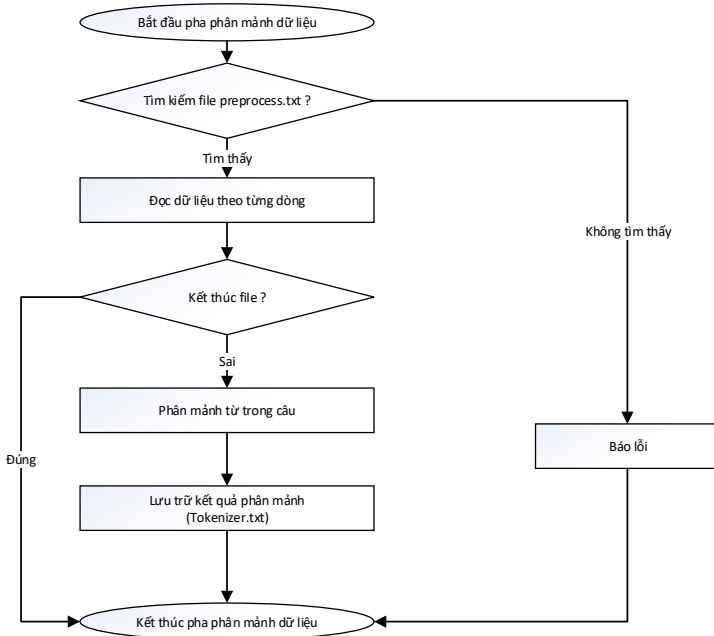
Dữ liệu được thu thập và trích xuất từ Page của các website bán hàng có thể đáp ứng được những yêu cầu cơ bản để sử dụng huấn luyện trong mô hình bán hàng tự động. Tuy nhiên, với hiện trạng trên mạng, những comments trên các post đôi khi không có ý nghĩa, hoặc bị lặp từ. Việc tiền xử lý dữ liệu sẽ chọn lọc các thuộc tính phù hợp cho mô hình (*Filtering Attributes*) và làm sạch (*clean data*) dữ liệu thô được thu thập sau bước thu thập. Nghiệp vụ của pha này gồm có:



Sau pha này, kết quả thu được là một file dữ liệu đã được xử lý, tất cả những cặp câu tư vấn không có ý nghĩa, hoặc bị trùng giữa câu hỏi và câu trả lời sẽ bị xóa.

3.2.3 Pha phân mảnh dữ liệu

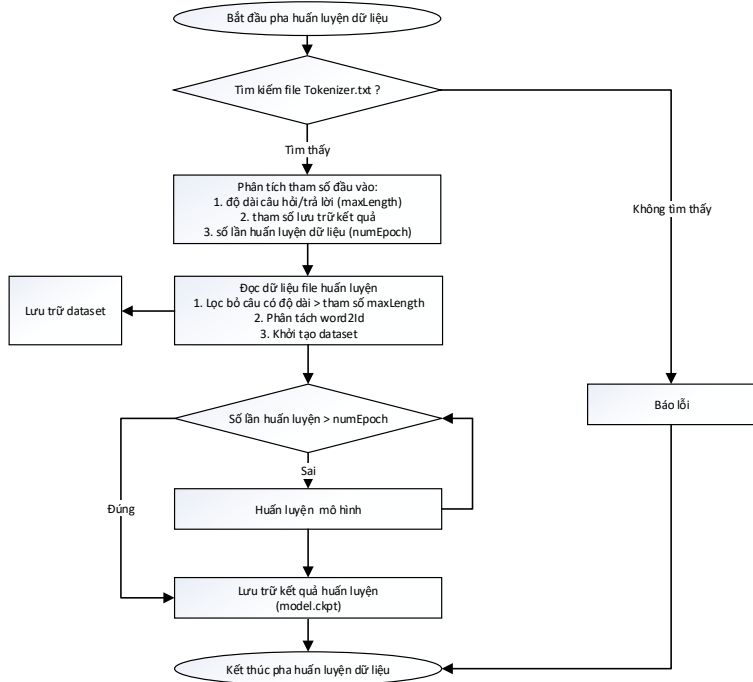
Một trong những thử thách trong mô hình bán hàng tự động qua mạng tại Việt Nam là thử thách về ngôn ngữ. Tiếng Việt phức tạp hơn tiếng Anh ở nhiều khía cạnh. Ở tiếng Anh, định nghĩa từ và sự kết hợp từ đơn giản, trong khi đó với Tiếng Việt lại khác. Ví dụ câu chào đơn giản: “Welcome to HaNoi” và “Chào mừng tới Hà Nội”, cùng một ý nghĩa nhưng phương pháp xử lý ở hai câu là khác nhau.



Kết quả nhận được là file segment chứa tập các cặp hỏi-đáp đã được tách từ theo ngôn ngữ Tiếng Việt. Tuy kết quả chỉ được $> 90\%$, nhưng rõ ràng có thể sử dụng làm input tốt cho pha huấn luyện dữ liệu.

3.2.4 Pha huấn luyện dữ liệu

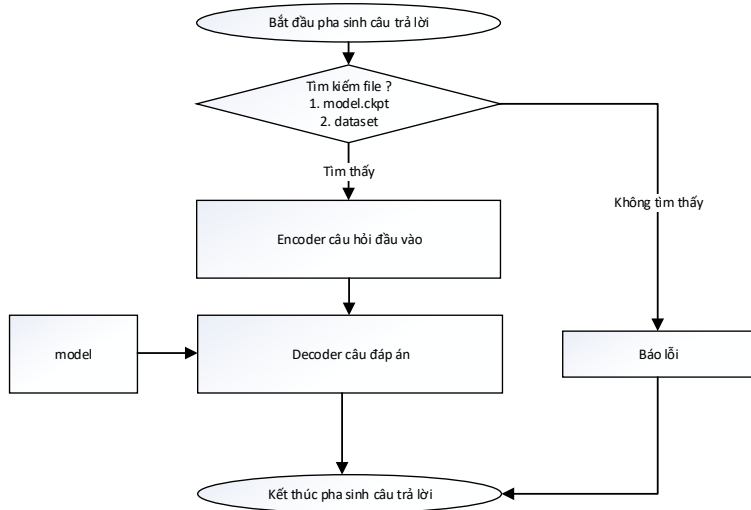
Pha huấn luyện dữ liệu là pha mấu chốt, quan trọng nhất trong xây dựng mô hình bán hàng tự động. Sau bước thu thập dữ liệu, 2 pha tiền xử lý và phân mảnh dữ liệu chỉ mang lại tác dụng làm cho kết quả của mô hình được tốt hơn, thì pha huấn luyện này là pha không thể thiếu.



Output của pha là dataset được tạo ra từ dữ liệu thu thập và mô hình iSales đã qua huấn luyện. Mô hình này đã giải quyết vấn đề xây dựng một mô hình có thể hiểu được ngôn ngữ tiếng Việt. Ở hàm `extractConversation`, `nltk` được sử dụng để tách các từ thành mã (`word2id`). Các từ được tách bằng thành mã sẽ không phân biệt ngôn ngữ tiếng Anh hay tiếng Việt. Hơn nữa, pha phân mảnh dữ liệu đã đảm bảo các từ tách ra hoàn toàn dễ hiểu khi đứng một mình.

3.2.5 Pha sinh câu trả lời

Dựa trên mô hình được xây dựng qua bước huấn luyện dữ liệu, có thể ứng dụng mô hình dữ liệu này để dùng cho việc sinh câu trả lời tư vấn cho các câu hỏi đầu vào. Khi câu hỏi được đặt ra, hệ thống sẽ phân tách câu hỏi ra từng từ, mapping theo dataset được tạo trong pha trước đó. Từ đó sẽ sử dụng bộ Encoder và Decoder để sinh ra câu trả lời phù hợp với ngữ cảnh nhất.



Output của pha là câu trả lời được tạo ra từ mô hình dữ liệu đã qua huấn luyện. Rõ ràng, ở pha này, LSTM đã giúp giải quyết vấn đề phát sinh những câu hỏi nằm ngoài “tập dữ liệu mẫu”: việc huấn luyện mô hình với phương pháp LSTM cho phép mô hình có khả năng hiểu ngữ cảnh, sinh câu trả lời dựa trên các tri thức học được. Một vấn đề khác được xử lý là câu hỏi dài và phức tạp, mạng LSTM giải quyết được bài toán trí

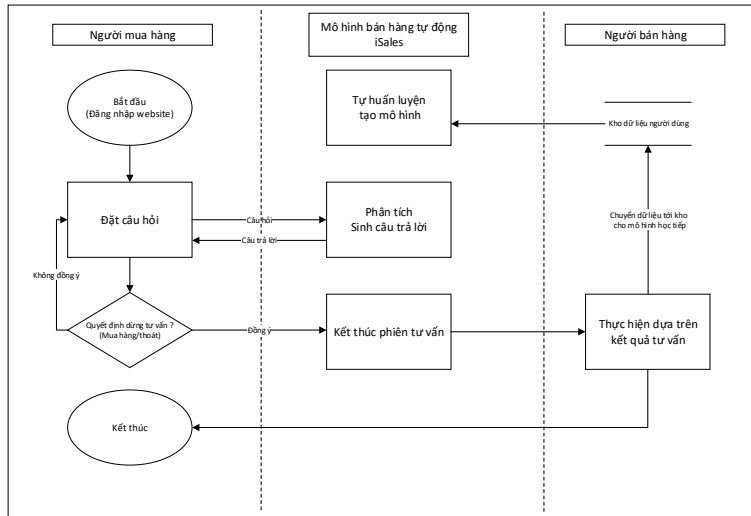
nhớ dài hạn. Bên cạnh đó, huấn luyện mô hình bằng Tensorflow cho phép tham số đầu vào là độ dài của câu hỏi/trả lời học. Đó là lời giải thích đáng cho vấn đề này.

iSales là giải pháp đề xuất để giải quyết bài toán mô hình bán hàng tự động. Mô hình này có tổng thể 5 pha, được xây dựng dựa trên phương pháp học chuỗi seq2seq và mạng nơ-ron. Trên lý thuyết, iSales đáp ứng các yêu cầu ban đầu, có khả năng tự học dữ liệu người dùng, tự sinh câu trả lời trong phạm vi thương mại điện tử mà không gặp khó khăn với các câu hỏi dài. Và để có những đánh giá khách quan nhất, tôi sẽ thử nghiệm iSales với dữ liệu thực tế trên mạng, với danh sách các câu hỏi bất kỳ trong chương 4.

CHƯƠNG 4: THỬ NGHIỆM, ĐÁNH GIÁ KẾT QUẢ

4.1 Phát biểu usecase

Để đánh giá khách quan cho mô hình, cần có một usecase áp dụng cụ thể. Sau quá trình khảo sát, tìm hiểu, tôi xin đưa ra được quy trình nghiệp vụ của việc bán hàng tự động trên Internet như sau:



Phần quan trọng nhất trong quy trình là hệ thống tự động phân tích, đưa ra câu trả lời phù hợp mà không cần đến người bán hàng.

4.2 Thử nghiệm iSales

Thử nghiệm mô hình iSales trên máy tính cá nhân với 5 pha: thu thập dữ liệu, tiền xử lý, phân mảnh, huấn luyện dữ liệu và sinh câu trả lời. Usecase đặt ra ở phần trước yêu cầu thực hiện đầy đủ và cần ít nhất 10 câu hỏi. Tuy nhiên, khả năng trả lời câu hỏi của iSales cũng như đa phần các ứng dụng trí tuệ nhân tạo khác, đều bị ảnh hưởng bởi dữ liệu huấn luyện, thời

gian huấn luyện và thuật toán sử dụng. Để có kết quả chi tiết hơn, tôi sẽ áp dụng usecase trên với 3 kết quả huấn luyện khác nhau về thời gian.

Với mong muốn đảm bảo lượng dữ liệu tạm chấp nhận cho quá trình huấn luyện, tôi thực hiện pha thu thập dữ liệu trên hệ điều hành window với danh sách 10 page bán hàng như sau.

Tên page	Mô tả
cuahangcongnghecom	Trang bán hàng công nghệ
laroseshop1990	Trang bán hàng quần áo
Hilheoshop96	Trang bán hàng quần áo
92wear	Trang bán hàng quần áo
banhangtructuyen360	Trang bán hàng quần áo
ChilloutVillage.vn	Trang bán hàng quần áo
ChoNinhHiep.net	Trang bán hàng quần áo
T.Shirt.Pro	Trang bán hàng quần áo
adoredressdesign	Trang bán hàng quần áo
cloud.cuckoo.shop	Trang bán hàng quần áo

Kết quả của pha là file collection.txt với số lượng dòng comment thu được 5683 dòng. Nhận xét chung là dữ liệu xấu, có dòng trống, có comment không có ý nghĩa, xuất hiện các dòng lặp liên kề nhau. Thực hiện tiếp pha tiền xử lý dữ liệu, kết quả thu được là file preprocess.txt với 5148 dòng dữ liệu tạm ổn, không còn dòng trống, các dòng lặp bị xóa, comment không có ý nghĩa bị loại bỏ. Output của pha phân mảnh dữ liệu là tokenizer.txt, với số dòng dữ liệu không thay đổi, tuy nhiên dữ liệu không còn dòng trống, các dòng lặp bị xóa, comment không có ý nghĩa bị loại bỏ, các từ đã được tách theo chuẩn Tiếng Việt. Mặc dù còn có một số câu tách chưa chuẩn như “Dạ giá

850k_a”...nhưng đánh giá sơ bộ kết quả này có thể áp dụng cho huấn luyện dữ liệu.

Kết quả huấn luyện có thể liệt kê như bảng sau. Rõ ràng, cùng một khối dữ liệu như nhau, nếu tăng các tham số huấn luyện lên kéo theo thời gian huấn luyện tăng lên.

Thực hiện sinh câu trả lời với danh sách gồm 15 câu hỏi về usecase mua áo sơ mi, gồm các câu hỏi liên quan đến màu sắc, cỡ, giá cả của áo, yêu cầu các thông tin khi ship về địa phương. Các câu hỏi được đặt một cách ngẫu nhiên

Kết quả sinh câu trả lời không được tốt. Với các trường hợp test có maxLength nhỏ, thì iSales sẽ không trả lời. Những câu khác trả lời được nhưng câu trả lời không hợp lý.

Kết quả tốt nhất thu thập được:

Human: xin chào
iSales (case03): đẹp
Human: mình muốn mua áo sơ mi
iSales (case03) : ok a
Human: có màu trắng không
iSales (case03): lảng ah
Human: còn màu đen không
iSales (case03): ok
Human: còn màu khác không
iSales (case03): ok
Human: cỡ áo l còn không
iSales (case03): còn b nhé
Human: cỡ áo s hết à bạn
iSales (case03): hiccc
Human: áo trắng bn ạ
iSales (case03): mũ 80k ạ
Human: giá màu đen bao nhiêu
iSales (case03): sỉ vâng
Human: giá chiếc màu trắng bao nhiêu

iSales (case03): 40

Human: thế là mình tự phải ra mua đúng ko ạ

iSales (case03): 40

Human: khi nào thì hàng đi vậy

iSales (case03): yeah

Human: nếu ship ra tỉnh khác thì mất thêm khoảng bao nhiêu vậy bạn

iSales (case03): dạ vâng ạ

Human: ship về vĩnh phúc hết bao lâu

iSales (case03): giá 550k ạ

Human: chuyển qua bình xuyên, vĩnh phúc bạn nha

iSales (case03): dạ không ạ

4.3 Đánh giá kết quả

Sau khi thực hiện thử nghiệm mô hình iSales trên usecase mua áo sơ mi, kết quả thu được khi đặt câu hỏi cho iSales là tỷ lệ đáp án chấp nhận được quá thấp < 10%. Tuy nhiên, iSales đã giải quyết được 1 số vấn đề:

- Vấn đề phát sinh những câu hỏi nằm ngoài “tập dữ liệu mẫu”, phần câu hỏi đặt ra hoàn toàn mang tính chất khách quan. Tuy nhiên có thể thấy, iSales đã tự đưa ra câu trả lời được với tất cả các câu hỏi phù hợp.
- Vấn đề xây dựng một mô hình có thể hiểu được ngôn ngữ tiếng Việt, iSales đã có thể hiểu được câu hỏi bằng tiếng Việt, và trả lời bằng tiếng Việt.
- Vấn đề câu hỏi dài và phức tạp, tùy thuộc vào tham số độ dài huấn luyện, iSales đã thể hiện như trong 3 trường hợp. Huấn luyện với tham số độ dài càng cao, thì độ dài câu mà iSales càng có khả năng đưa ra câu trả lời.
- Vấn đề “tự động”, iSales đã làm được điều này, nhưng đáp án đưa ra chưa tốt, cần cải thiện rất nhiều.

Rõ ràng, với những kết quả hiện tại, iSales chưa thể trở thành ứng dụng trên thị trường. Luận văn thành công ở phương diện phương án xây dựng bằng một phương pháp mới, nhưng vẫn chưa đảm bảo được kết quả tốt. Nếu so sánh với chatbot Skype dưới con mắt người dùng, chatbot Skype vẫn tốt hơn, vì có thể hoàn thành được phiên giao dịch nếu người mua hàng làm đúng theo chỉ dẫn của nó. Tuy nhiên, iSales cũng là một thử thách hay trong lĩnh vực ứng dụng mạng nơ-ron nói riêng và trong ngành vực trí tuệ nhân tạo nói chung.

KẾT LUẬN

Luận văn tập trung tìm hiểu thực trạng về các mô hình bán hàng trên thế giới và Việt Nam, đưa ra các nhược điểm, các thành phần cần cải tiến, từ đó đề xuất giải pháp, xây dựng mô hình bán hàng iSales có tính tự động, không phụ thuộc vào người bán hàng trong quá trình tư vấn dịch vụ. Mô hình iSales hiện tại, chưa thể đưa ra thành ứng dụng thương mại. Tuy có thể đáp ứng được các yêu cầu có khả năng hiểu Tiếng Việt và sinh câu trả lời tự động với độ dài câu hỏi không giới hạn, nhưng kết quả tư vấn iSales đưa ra còn ngô nghê, khó hiểu.

Hướng nghiên cứu tiếp theo của luận văn tập trung vào việc cải thiện chất lượng câu trả lời của iSales.

- Mở rộng khả năng thu thập và tiền xử lý dữ liệu làm cho dữ liệu thu thập được nhiều hơn, sạch hơn. Phương án đưa ra là sử dụng BigData để quản lý dữ liệu và cho phép tìm kiếm, lọc dữ liệu nhanh hơn [12], hoặc lựa chọn các nguồn dữ liệu như dữ liệu sau tư vấn được lưu trữ của các doanh nghiệp.
- Cải tiến pha huấn luyện dữ liệu để giảm thời gian huấn luyện và tăng hiệu quả đầu ra của mô hình.

Đồng thời, tôi muốn phối hợp nghiên cứu và triển khai ứng dụng với các đơn vị bên ngoài, đầu tư mở rộng phần cứng để giảm thời gian huấn luyện. Những hướng phát triển trên nhằm hướng tới một mục đích chung là phát triển hệ thống iSales có tầm thế. Từ đó đóng góp, phát triển nền trí tuệ nhân tạo thế giới nói chung và Việt Nam nói riêng.

TÀI LIỆU THAM KHẢO

- [1] Facebook Messenger, September 2016, “Facebook Messenger”.
- [2] Microsoft, 31 Mar 2016, “Build 2016: Microsoft Skype bots preview announced”.
- [3] Martín Abadi, 14 Mar 2016, “TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems”.
- [4] Google Blog, 09 November 2015, “TensorFlow - Google’s latest machine learning system, open sourced for everyone”.
- [5] S.M. Al-Alawi, H.A. Al-Hinai, May–August 1998, “An ANN-based approach for predicting global radiation in locations with no direct measurement instrumentation”.
- [6] Ondřej Dušek, Filip Jurčiček, 17 Jun 2016, “Sequence-to-Sequence Generation for Spoken Dialogue via Deep Syntax Trees and Strings”.
- [7] Razvan Pascanu, Tomas Mikolov, Yoshua Bengio, May 2013, “On the difficulty of training recurrent neural networks”.
- [8] James Ryan, September 2016, “Translating Player Dialogue into Meaning Representations Using LSTMs”.
- [9] Jay Parikh, August 2012, “Facebook processes more than 500 TB of data daily”.
- [10] Facebook, August 2016, “The Graph API”.
- [11] Lư Tuấn Anh, Yamamoto Kazuhide, 16 Feb 2013, “Pointwise for Vietnamese Word Segmentation”.

[12] Xue-Wen Chen, Xiaotong Lin, 16 May 2014, “Big Data Deep Learning: Challenges and Perspectives”.

[13] Francois Chaubard, Rohit Mundra, Richard Socher, Spring 2015, “CS 224D: Deep Learning for NLP”.

[14] Sepp Hochreiter; Jürgen Schmidhuber, 1997, "Long short-term memory".

[15] Feigenbaum, Edward A. 2003, “Some challenges and grand challenges for computational intelligence”.