ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

Nguyễn Hà Thanh

NGHIÊN CỨU XÂY DỰNG ỨNG DỤNG XỬ LÝ VĂN BẢN LUẬT GIAO THÔNG

KHÓA LUẬN TỐT NGHIỆP ĐẠI HỌC HỆ CHÍNH QUY
Ngành: Công nghệ thông tin

HÀ NỘI – 2015

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

Nguyễn Hà Thanh

NGHIÊN CỨU XÂY DỰNG ỨNG DỤNG XỬ LÝ VĂN BẢN LUẬT GIAO THÔNG

KHÓA LUẬN TỐT NGHIỆP ĐẠI HỌC HỆ CHÍNH QUY

Ngành: Công nghệ thông tin

Cán bộ hướng dẫn: PGS TS. Nguyễn Việt Hà

HÀ NỘI - 2015

VIETNAM NATIONAL UNIVERSITY, HANOI
UNIVERSITY OF ENGINEERING AND TECHNOLOGY

Nguyen Ha Thanh

RESEARCH AND PROPOSE VIETNAMESE TRAFIC LAW PROCESSING METHOD

A THESIS PRESENTED FOR THE DEGREE BACHELOR

Major: Information Technology

Supervisor: Assoc Prof. Nguyen Viet Ha

HA NOI - 2015

TÓM TẮT

Tóm tắt: Mặc dù nhu cầu pháp lý của xã hội ngày một cao, hệ thống pháp luật của Việt Nam vẫn còn nhiều điểm chồng chéo, liên tục thay đổi, gây khó khăn cho việc tiếp cận, áp dụng, sử dụng và thực thi pháp luật. Ngày nay với sự phát triển nhanh chóng của các kỹ thuật học máy đặc biệt là kỹ thuật sử dụng mạng nơron nhân tạo, nhiều ứng dụng thông minh đã ra đời và giúp ích rất nhiều cho cuộc sống con người. Trong giới hạn của một khóa luận tốt nghiệp của sinh viên, đề tài nghiên cứu giải quyết bài toán đặt ra là xây dựng chương trình có khả năng phát hiện các câu luật giao thông có hình thái khác nhau nhưng biểu hiện ý nghĩa giống nhau. Hướng tiếp cận chính để giải quyết vấn đề là sử dụng kỹ thuật nơron nhân tạo trong học máy. Cách thức tiến hành thực nghiệm của đề tài phù hợp để chứng minh tính khả thi của phương pháp và có được những kết quả bước đầu khá ấn tượng, mở ra triển vọng cho các ứng dụng chất lượng cao trong xử lý các vấn đề pháp lý.

Từ khóa: Mạng nơron, xử lý tiếng việt, luật giao thông

SUMARY

Sumary: Nowadays, the demands for legal services in our society are rising sharply. However, the legal system in Vietnam is still greatly overlapping and constantly changing, which creates considerable difficulties for people in accessing, applying and using the law for legitimate reasons. Today, with the rapid development of machine learning, especially the technical uses of artificial neural network, many smart applications were born and became very helpful for human life. Within the scope of a graduation paper for the bachelor degree, this research aims at studying related knowledge and building a program having the capacity to detect the traffic law sentences which are in different morphology but express similar meaning. The main approach to achieve these aims is to use techniques in machine learning artificial neurons. Experimental methods proposed in this research are suitable for proving the method. Initial results are rather impressive, opening up prospects for high quality applications in handling legal issues.
Keyword: Artificial neural network, Vietnamese processing, traffic law

LỜI CAM ĐOAN

Tôi xin cam đoan những đóng góp trong khóa luận được trình bày một cách chính xác và trung thực, tất cả các tài liệu tham khảo, công trình nghiên cứu của người khác được sử dụng trong đề tài đều được ghi rõ nguồn, được liệt kê tại chú thích dưới mỗi trang và được đặt trong danh mục các tài liệu tham khảo của khóa luận.

Những cải tiến, đóng góp trong phương pháp, kỹ thuật lập trình cũng như mã nguồn của chương trình thực nghiệm tự thiết kế không có sự sao chép công trình của người khác. Nếu như những gì tôi nói trên đây là trái sự thật, tôi xin chịu hình thức kỷ luật cao nhất của nhà trường.

Hà Nội, ngày 30/4/2015

Sinh viên

Nguyễn Hà Thanh

LỜI CẢM ƠN

Trước tiên, em muốn gửi lời cảm ơn sâu sắc nhất đến thầy Nguyễn Việt Hà, thầy Nguyễn Lê Minh đã gợi ý cho em một hướng nghiên cứu rất thú vị và tận tình hướng dẫn, đưa những lời khuyên và kinh nghiệm quý báu cho em trong trong quá trình thực hiện khóa luận.

Em cũng xin bày tỏ lời cảm ơn sâu sắc đến các thầy là tác giả đề tài "Nghiên cứu phát triển một số sản phẩm thiết yếu về xử lí tiếng nói và văn bản tiếng Việt" (VLSP), những người đã tạo nền móng quan trọng cho việc tiếp cận xử lý ngôn ngữ tiếng Việt.

Em xin được gửi lời cảm ơn tới các tác giả của các bài báo, nghiên cứu có liên quan. Trong khoa học nói chung và trong khoa học máy tính nói riêng, không có phương pháp, cách thức nào là tuyệt đối tối ưu nhưng nhờ sự nghiên cứu nghiên túc và tận tâm của các tác giả, các cách tiếp cận, xử lý vấn đề trong công trình này đã hội tụ nhanh hơn tới cách tiếp tận hiệu quả trên thế giới hiện nay.

Hà Nội, ngày 30/4/2015

Sinh viên

Nguyễn Hà Thanh

MỤC LỤC

CHƯƠNG 1. MỞ ĐẦU VÀ ĐẶT VẤN ĐỀ

1.1. BỐI CẢNH NGHIÊN CỨU

Xã hội càng phát triển, chất lượng cuộc sống của con người ngày càng được nâng cao thì nhu cầu sử dụng pháp luật của các cá nhân, tổ chức cũng theo đó ngày một gia tăng. Trong hiến pháp nước Cộng hòa xã hội chủ nghĩa Việt Nam năm 2013, chế định về *quyền con người, quyền nghĩa vụ cơ bản của công dân* đã được đưa từ chương 5 lên chương 2 (so với hiến pháp năm 1992), điều đó một lần nữa cho thấy vị trí và vai trò của pháp luật trong đời sống thường ngày đang dần được khẳng định, một xã hội hoạt động theo pháp luật là cơ sở cho một sự phát triển nhanh và ổn định.

Để có thể sử dụng và áp dụng pháp luật, những người làm trong ngành phải đọc rất nhiều và liên tục cập nhật các thông tin từ các văn bản pháp luật mới được thông qua. Các văn bản pháp luật ra đời sau có hiệu lực thay thế, phủ định hoặc bổ sung các văn bản trước đó, việc này hiện nay diễn ra rất thường xuyên và liên tục gây trở ngại lớn cho những người hoạt động pháp lý. Những người dù hành nghề lâu năm cũng không dám tự tin những gì mình biết về một vấn đề pháp lý còn đúng nữa hay không nếu như họ đã không tra cứu vấn đề này trong một thời gian dài. Bên cạnh đó, do cơ chế bảo hiến của hệ thống pháp luật Việt Nam còn nhiều bất cập cho nên vẫn còn những điểm chồng chéo mâu thuẫn giữa các văn bản quy phạm pháp luật. Tại thời điểm khóa luận này được hoàn thành, các văn bản quy phạm pháp luật về thuế đã và đang liên tục được sửa đổi. Thời báo kinh tế Sài Gòn có đoạn viết: "*[...] "Ma trận" các văn bản về thuế đã tạo ra hệ thống văn bản quy phạm pháp luật chồng chéo, chắp vá và gây khó khăn lớn cho đối tượng thực hiện. Chẳng hạn, để biết quy định nào của Luật Thuế TNDN còn hiệu lực thi hành và quy định cụ thể như thế nào, đối tượng thực hiện phải so sánh Luật Thuế TNDN năm 2008, Luật Sửa đổi, bổ sung một số điều của Luật Thuế TNDN năm 2013, Luật Sửa đổi, bổ sung một số điều tại các luật thuế năm 2014, các nghị định và thông tư tương ứng, trong đó có cả nghị định sửa nhiều nghị định và thông tư sửa nhiều thông tư.[...]*[1] "

[1] http://www.thesaigontimes.vn/125339/Ma-tran-van-ban-phap-luat-ve-thue.html

Với những thành tựu rất đáng kể của các hướng nghiên cứu Trí tuệ nhân tạo như Học máy, Xử lý ngôn ngữ tự nhiên trong những năm gần đây, việc áp dụng Công nghệ thông tin để xử lý các văn bản Luật hứa hẹn sẽ tạo ra được một cuộc cách mạng về phương pháp tìm kiếm phục vụ cho việc soạn thảo, sử dụng, áp dụng và thực thi pháp luật. Các hệ thống thông minh còn có thể giúp phát hiện ra những điểm mâu thuẫn, chồng chéo trong hệ thống pháp luật cũng như cung cấp kiến thức chuyên gia để giải quyết một vấn đề pháp luật.

1.2. NHIỆM VỤ CỦA KHÓA LUẬN

Trong giới hạn của một khóa luận tốt nghiệp của sinh viên, nội dung nghiên cứu tập trung giải quyết một bài toán nhỏ liên quan đến xử lý văn bản luật. Trong một hệ thống văn bản pháp luật chồng chéo (ví dụ như hệ thống các quy phạm về thuế trong phần trước), sẽ có những câu luật trong các văn bản khác nhau, được viết theo cách sắp xếp từ khác nhau nhưng lại mang ý nghĩa đồng nhất. Việc phát hiện ra được những cặp câu có tính chất như vậy sẽ là cơ sở của rất nhiều các ứng dụng xử lý pháp luật sau này. Nhiệm vụ của đề tài là khái quát cơ sở lý thuyết, kế thừa các nghiên cứu đã có, đề xuất giải pháp và xây dựng được một chương trình có khả năng phát hiện được những cặp câu luật như vậy trong một ngữ cảnh được giới hạn là các quy phạm pháp luật quy định về giao thông Việt Nam. Đóng góp của đề tài có thể được sử dụng cho các ứng dụng góp phần tăng khả năng tiếp cận các quy định về giao thông cho mọi người, tăng tốc độ tìm kiếm những điều luật liên quan đến công việc của các luật sư, thẩm phán, những cá nhân, tổ chức đang áp dụng, thi hành, sử dụng pháp luật và phát hiện sự chồng chéo trong các văn bản luật.

Hướng tiếp cận chính để giải quyết vấn đề là sử dụng kỹ thuật nơron nhân tạo trong học máy. Cụ thể, công trình sử dụng hai mạng nơron thực hiện hai nhiệm vụ chính, một là vector hóa các từ và hai là phát hiện sự đồng nghĩa của các câu luật được viết đúng chính tả với cấu trúc sắp xếp từ ngẫu nhiên. Công trình chủ yếu học tập ý tưởng của Richard Socher, Andrej Karpathy, Quoc V. Le*, Christopher D. Manning, Andrew Y. Ng. trong bài báo *Grounded Compositional Semantics for Finding and Describing Images with Sentences*. Đóng góp của công trình là đề xuất được một bài toán có ý nghĩa thực tiễn và xây dựng được một hệ thống hoạt động một cách tương đối hiệu quả với dữ liệu là tiếng Việt dựa trên những công cụ, nghiên cứu đã có trước đó và một số cải tiến về kỹ thuật.

1.3. CÁC NỘI DUNG CỦA KHÓA LUẬN

Khóa luận được trình bày trong 5 chương nhằm cung cấp một cái nhìn tổng thể về bối cảnh nghiên cứu, ý nghĩa của đề tài, các cơ sở lý thuyết có liên quan, quy trình, kết quả tiến hành thực nghiệm và một số so sánh với các công trình đã có trên thế giới.

Chương mở đầu nói về ý nghĩa, vị trí của đề tài trong bối cảnh chung xét trên phương diện xu hướng phát triển của xã hội cũng như xu hướng phát triển của các kỹ thuật Trí tuệ nhân tạo mà cụ thể ở đây là Học máy. Phần cuối chương tóm tắt bố cục của khóa luận nhằm giúp cho các thầy cô, các bạn và các em dễ theo dõi, tiện cho việc đánh giá, đối sánh và tham khảo.

Chương 2 nêu ra những cơ sở lý thuyết quan trọng có liên quan đến đề tài. Đầu tiên là những lý thuyết về mạng nơron nhân tạo, phần này nhằm cung cấp cho những người đọc không cùng chuyên ngành có thể dễ dàng nắm bắt được ý tưởng và tiếp tục hiểu được những phần tiếp theo của khóa luận. Tiếp đó là cách thức sử dụng và huấn luyện mạng nơron, phương pháp truyền sai số ngược và cập nhật trọng số mạng bằng giải thuật Gradient descent và cải tiến kỹ thuật của nó (Stochastic gradient descent). Cuối chương 2 khóa luận trình bày một số đặc điểm của ngôn ngữ tiếng Việt, đó là một trong những cơ sở quan trọng để giải thích phương pháp và phân tích kết quả thực nghiệm.

Chương 3 của khóa luận nói về phương pháp được đề xuất để giải quyết bài toán thực nghiệm cụ thể là sử dụng mạng nơron nhân tạo để phát hiện các câu luật mang cùng ý nghĩa. Trong chương này, bài toán thực nghiệm được phát biểu một cách rõ ràng, chính xác bằng ngôn ngữ tự nhiên, ngôn ngữ ký hiệu và có ví dụ minh họa. Tiếp đó là những nghiên cứu được kế thừa và phương pháp được đề xuất để giải quyết bài toán cụ thể đối với các câu luật giao thông Việt Nam. Sau đó, phần cuối chương nêu ra những đánh giá sơ bộ về phương pháp trên phương diện những điểm mạnh, hạn chế và nguyên nhân của chúng.

Chương 4 mô tả lại quy trình và cách thức thực nghiệm bao gồm công cụ, môi trường, dữ liệu và phương pháp tổ chức thực nghiệm. Sau đó, các kết quả của thực nghiệm được trình bày bằng bảng thống kê và một số ví dụ trong tập kiểm thử. Cuối cùng các kết quả thực nghiệm được phân tích, đánh giá một cách tổng thể dựa trên định lượng và định tính để rút ra những điểm đã đạt được, những điểm còn hạn chế và hướng giải quyết các hạn chế đó. Đề tài có nêu lên một số các kết quả của những nghiên cứu có liên

quan để thấy được chất lượng của phương pháp đề xuất trong công trình. Thông qua so sánh, có thể thấy được kết quả khả quan bước đầu của phương pháp được đề xuất.

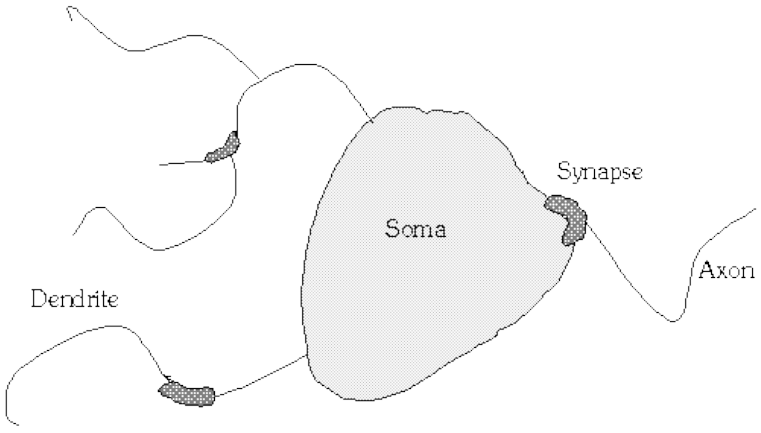
Chương 5 tổng kết lại toàn bộ những gì đã được làm rõ trong khóa luận một cách cô đọng và súc tích nhất, phục vụ cho việc đánh giá tổng quan cả đề tài và hỗ trợ việc tra cứu đối với những người quan tâm đến đề tài nghiên cứu. Chương này tóm lược lại các nội dung về xuất xứ của bài toán, các phương pháp và kết quả thực nghiệm, kết luận lại những ý nghĩa và tiềm năng của kết quả nghiên cứu. Sau đó, các mặt mạnh, điểm hạn chế của khóa luận được nêu ra và cuối cùng là các hướng nghiên cứu tiếp theo để giải quyết những vấn đề còn tồn tại và nâng cấp công trình.

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

2.1. TỔNG QUAN VỀ MẠNG NƠRON NHÂN TẠO

Xây dựng và sử dụng mạng nơron nhân tạo là một kỹ thuật trong Học máy (Machine learning) nó có thể đánh giá, tiên liệu được giá trị đầu ra của các bộ dữ liệu có đầu vào chưa biết trước. Mạng nơron nhân tạo được tạo nên từ các nơron (các nút tính toán) có liên kết với nhau thông qua các đường truyền tín hiệu có trọng số, tùy vào dữ liệu sử dụng trong huấn luyện mạng nơron mà các trọng số này được cập nhật và hình thành nên đặc trưng riêng của mạng nơron đó. Khi mạng nơron đã được huấn luyện thành công, nó có khả năng làm việc với các dữ liệu cùng loại với dữ liệu đã được huấn luyện nhưng với đầu vào chưa biết trước. Mạng nơron nhân tạo được phát minh dựa trên ý tưởng của mạng nơron sinh học (hệ thống thần kinh trung ương của động vật, chủ yếu là não bộ).

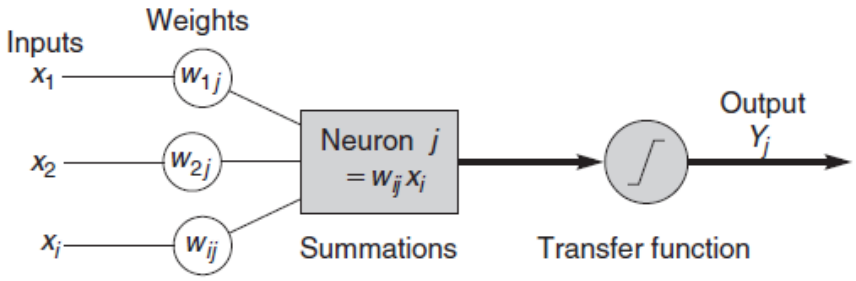
Giới thiệu sơ lược về mạng nơron sinh học, đây là cấu trúc được cấu tạo phức tạp nhưng ta quan tâm đến ba thành phần chính là Soma, Dendrite và Axon. Soma là nhân của nơron, chịu trách nhiệm chính cho việc tính toán và phát ra những xung thần kinh. Dendrite và Axon là các dây dài và mảnh, làm nhiệm vụ dẫn truyền xung thần kinh, đây cũng là lý do tên gọi "dây thần kinh" được ra đời. Hai loại dây này khác nhau ở chỗ Dendrite truyền các xung điện cho nhân Soma xử lý còn Axon truyền các xung điện từ nhân đi ra nếu như điện thế trong nhân vượt quá một ngưỡng nào đó.



Hình 2.1. Mạng nơron sinh học

Hiểu một cách đơn giản, nơron hoạt động bằng cách lấy tổng các xung điện nó nhận được và phát ra một xung điện khác nếu như điện thế trong nơron vượt một ngưỡng nào đó. Các xung điện truyền giữa các nơron thông qua các khớp có tính truyền khác nhau. Các khớp mạnh có khả năng truyền thông tin rất dễ dàng trong khi các khớp yếu làm cản trở thông tin truyền qua.

Được lấy cảm hứng từ mạng nơron sinh học, mạng nơron nhân tạo cũng có cấu tạo và cách hoạt động tương tự như vậy.



Hình 2.2. Mạng nơron nhân tạo

Mỗi thành phần tính toán (nơron) trong mạng nơron nhân tạo cũng có các cửa ngõ nhận thông tin giống Dendrite và Axon. Thông tin được truyền giữa các nơron này là các số thực, trên mỗi mối nối có một trọng số khác nhau để mô phỏng tính truyền của mạng nơron sinh học. Tại mỗi nơron, các tín hiệu đầu vào được công dồn và truyền qua hàm kích hoạt, hàm kích hoạt đóng vai trò tạo ra một ngưỡng tín hiệu cho nơron nhân tạo. Khi tổng của các kích thích đầu vào thỏa mãn điều kiện về độ lớn, nơron nhân tạo mới có thể phát tín hiệu sang nơron kế tiếp nó ở lớp tiếp theo với một cường độ được kiểm soát. Các hàm kích hoạt phải thỏa mãn 3 điều kiện:

- Có tính đơn điệu

- Bị chặn trên và chặn dưới

- Có tính liên tục và trơn

Các hàm kích hoạt được dùng trong mạng nơron nhân tạo được triển khai trên máy tính còn phải thỏa mãn đặc tính là đơn giản trong việc tính đạo hàm, thông thường các

hàm kích hoạt được sử dụng là Hàm ngưỡng, Hàm tuyến tính từng đoạn và các hàm Hyperbolic. Trong công trình sử dụng hàm tanh (thuộc họ hàm Hyperbolic). Công thức của hàm tanh(x) như sau

$$tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

Đạo hàm của hàm tanh(x) được tính rất đơn giản bởi công thức:

$$\frac{d}{dx} tanh(x) = 1 - tanh^{2}(x)$$

2.2. SỬ DỤNG VÀ HUẤN LUYỆN MẠNG NƠRON NHÂN TẠO

Theo *Giáo trình tin học cơ sở* của thầy Đào Kiến Quốc và thầy Trương Ninh Thuận, việc xử lý thông tin trên máy tính không làm tăng lượng tin mà chỉ hướng hiểu biết của con người vào những khía cạnh hữu ích trong hoạt động thực tiễn[2]. Do đó việc xử lý thông tin trên máy tính có thể quy về các hàm tính toán có đầu vào là các thông tin đã biết và một đầu ra là những thông tin có thể suy luận được và phù hợp với nhu cầu sử dụng của con người. Trong tập hợp tất cả các bài toán xử lý dữ liệu, tồn tại những bài toán có hàm tính toán đơn giản và có thể cho ra chính xác đầu ra đối với các đầu vào tương ứng, ví dụ như các bài toán giải phương trình đa thức, bài toán tính lương từ số ngày công hay bài toán chuyển đổi tiền tệ. Tồn tại song song với nó là những bài toán mà chi phí để tìm ra một hàm tính toán chính xác với mọi đầu vào quá lớn so với năng lực hiện tại của máy tính và con người, ví dụ như các bài toán về nhận diện chữ viết tay, nhận diện khuôn mặt, dự đoán ung thư hay các bài toán xử lý ngôn ngữ tự nhiên phức tạp. Để giải quyết phần nào các bài toán này, khoa học về phương pháp tính ra đời với nhiệm vụ tìm được hàm tính toán xấp xỉ đủ tốt so với hàm tính toán chính xác, các hàm này gọi là hàm giả thiết (hypothesis).

Mạng nơron nhân tạo nếu xét trên phương diện xử lý thông tin cũng là một hàm tính toán với đầu vào và đầu ra xác định, ban đầu các trọng số giữa các liên kết nơron được tạo ngẫu nhiên nên khi một đầu vào bất kỳ được truyền cho mạng nơron, kết quả đầu ra sẽ là một giá trị ngẫu nhiên. Cấu trúc mạng nơron nhân tạo ưu việt ở chỗ, nó có thể

[2] Giáo trình tin học cơ sở - Đào Kiến Quốc, Trương Ninh Thuận 6-2010

tự động cập nhật các trọng số liên kết để xấp xỉ hàm tính toán gần đúng với một tập dữ liệu được biết trước (gọi là tập dữ liệu học). Trong quá trình thay đổi trọng số như vậy, mạng nơron sẽ hình thành ra những luật xử lý dữ liệu có khả năng dự đoán kết quả đầu ra với một đầu vào nó chưa từng biết trước. Cách thức học của máy với mạng nơron nhân tạo có điểm tương đồng với cách thức học của con người và các loài động vật khác đó là sử dụng kinh nghiệm đã có để phán đoán những gì chưa biết trong tương lai.

Đa phần các bài toán Mạng nơron nhân tạo hoạt động dựa trên 3 hành vi chính là tính toán thử, xác định sai số và tái cấu trúc mạng. Với một tập dữ liệu học $D_{learn} = \{x^{(i)}, y^{(i)}\}$, với $x^{(i)}$ và $y^{(i)}$ là đầu vào và đầu ra của ví dụ thứ i trong tập dữ liệu học D_{learn}, mạng nơron sẽ tính toán giá trị đầu ra ứng với $x^{(i)}$. Tiếp đó, mạng nơron sẽ tái cấu trúc bằng cách cập nhật lại các trọng số liên kết bằng phương pháp lan truyền ngược (back propagation) với mục tiêu tối thiểu hóa sai số với kết quả đầu ra của mạng, công việc đó được công thức hóa bằng việc tối ưu hàm giá $J_{train}\,(\theta)$

$$J_{train}\,(\theta) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

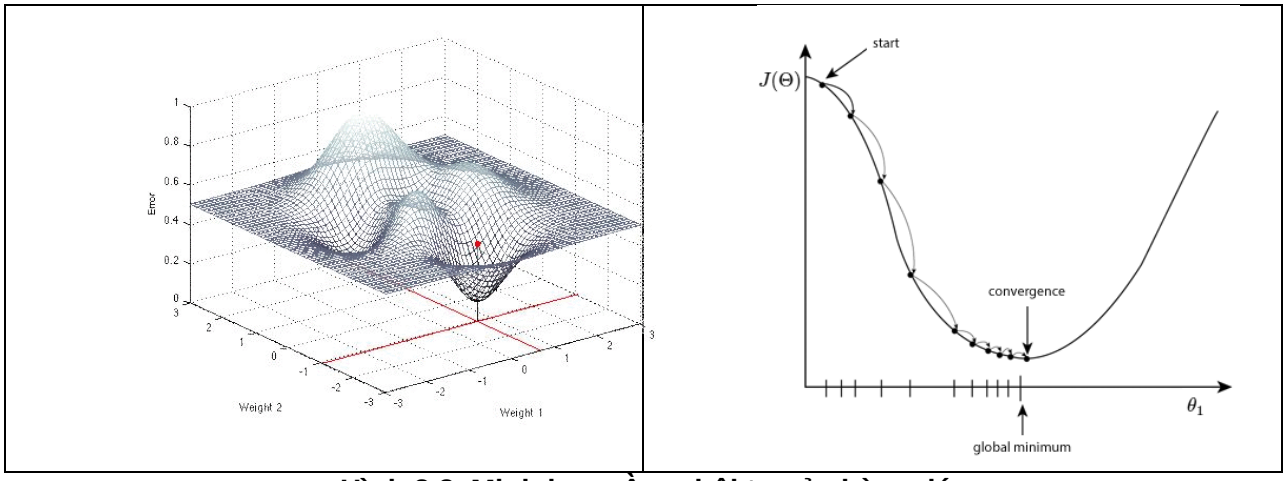
2.3. PHƯƠNG PHÁP GRADIENT DESCENT VÀ STOCHASTIC GRADIENT DESCENT

Mạng nơron nhân tạo là sự liên kết giữa các nơron một cách có thứ tự, giá trị tính toán của nơron sau sẽ phụ thuộc vào giá trị của nơron trước đó. Phương pháp lan truyền ngược được thực hiện dựa trên ý tưởng các sai số của nơron phía sau sẽ là cơ sở để cập nhật trọng số liên kết và xác định giá trị sai số của nơron trước đó. Có rất nhiều phương pháp để thực hiện lan truyền ngược sai số nhưng điển hình vẫn là phương pháp Gradient descent. Ý tưởng của phương pháp này là hiệu chỉnh các trọng số dựa trên vi phân hàm giá. Cho đến khi giá trị của sai số hội tụ, thuật toán sẽ lặp đi lặp lại công thức sau:

$$\theta_j \leftarrow \theta_j - \propto \frac{\partial}{\partial\theta_j} J(\theta_1, \theta_2, \dots, \theta_n)$$

Trong đó, $\theta_1, \theta_2, \dots, \theta_n$ là tập các trọng số của mạng nơron, $\propto$ là hệ số học (learning rate) của mạng nơron, $\frac{\partial}{\partial\theta_j} J(\theta_1, \theta_2, \dots, \theta_n)$ là vi phân của hàm giá theo trọng số

θ_j. Với việc lặp lại sự cập nhật này, hàm giá sẽ hội tụ và sai số của hàm giả thiết sẽ đạt giá trị cực tiểu. Hình 3 mô phỏng sự hội tụ của hàm giá với phương pháp Gradient descent



Hình 2.3. Minh họa về sự hội tụ của hàm giá

Mặc dù vậy, đối với tập dữ liệu có lực lượng lớn, phương pháp Gradient descent tỏ ra không hiệu quả vì chi phí tính toán hàm giá lớn dẫn đến thời gian hội tụ lâu. Giả sử với tập dữ liệu với 100.000.000 phần tử, mỗi lần cập nhật 1 giá trị trọng số, máy tính sẽ phải tính toán $\frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$ với m=100.000.000. Đó sẽ là một chi phí rất lớn khi thực tế là một mạng nơron có đến hàng trăm, hàng nghìn trọng số. Theo Andrew Ng, giảng viên tại Stanford hiện đang giảng dạy Machine Learning trên website học tập uy tín coursera.org, trung bình sự hội tụ của hàm giá sẽ diễn ra trong 1.000 lần lặp cuối cùng của quá trình học, như vậy thời gian để máy tính thiết lập được mạng nơron hoạt động tốt với bài toán này sẽ rất lớn và bất khả thi trên phương diện cài đặt.

Phương pháp Stochastic Gradient Descent ra đời và hạn chế nhược điểm trên của Gradient Descent. Với phương pháp Gradient Descent máy tính trong mỗi lần lặp sẽ tính tất cả tổng các sai số rồi mới cập nhật giá trị trọng số còn đối với Stochastic Gradient Descent, mỗi lần lặp, máy tính sẽ cập nhật ngay trọng số dựa trên sai số của một cặp đầu ra và kết quả kiểm tra bất kỳ. Hàm giá của Stochastic Gradient Descent được biểu diễn như sau:

$$cost(\theta, (x^{(i)}, y^{(i)})) = \frac{1}{2}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

$$J_{train}(\theta) = \frac{1}{2m}\sum_{i=1}^{m} cost(\theta, (x^{(i)}, y^{(i)}))$$

Với Stochastic Gradient Descent, tốc độ hội tụ diễn ra nhanh hơn nhiều lần so với Gradient Descent nhưng nó có những yêu cầu về kỹ thuật để đảm bảo kết quả chính xác như Gradient Descent:

- Tập dữ liệu phải được xáo trộn trước mỗi lần lặp để đảm bảo tính đồng xác suất của mỗi ví dụ huấn luyện.

- Trước mỗi lần cập nhật trọng số cần có thao tác kiểm tra tính hội tụ của ví dụ vừa huấn luyện.

- Phải có cơ chế kiểm soát sự bùng nổ của giá trị trọng số để tránh trường hợp hội tụ giả do đặc tính của hàm kích hoạt.

- Cần có chiến thuật chọn hệ số học $\propto$ thích hợp để cân bằng giữa tốc độ hội tụ và chất lượng hội tụ, tối thiểu hóa các điểm hội tụ tại cực tiểu địa phương.

2.4. ĐẶC ĐIỂM CỦA NGÔN NGỮ TIẾNG VIỆT

Đề tài có đối tượng nghiên cứu là các quy phạm pháp luật về giao thông Việt Nam, đó là các văn bản được viết bằng tiếng Việt, do vậy việc hiểu được đặc điểm của ngôn ngữ là một công việc hết sức quan trọng. Đề tài quan tâm đến 3 đặc điểm chính của tiếng Việt khiến việc áp dụng nguyên văn những nghiên cứu đã có trên thế giới về xử lý ngôn ngữ tự nhiên cho ngôn ngữ của chúng ta là bất khả thi.

Một là vấn đề tách từ trong tiếng Việt. Do tiếng Việt đa dạng về từ loại (từ đơn, từ phức, từ ghép, thành ngữ...) nên việc xác định ranh giới của một từ không thể dựa vào hình thức của văn bản. Ở một số ngôn ngữ khác, việc xác định ranh giới của một từ đơn giản là sử dụng các dấu câu, dấu cách, ký tự xuống dòng trong một văn bản còn trong tiếng Việt, để tách được một từ cần phải hiểu được ý nghĩa của từ trong ngữ cảnh. Ví dụ như câu: “I am a teacher” trong tiếng Anh, gồm 4 từ được cách nhau bằng các dấu cách, cùng ý nghĩa như vậy, câu “Tôi là giáo viên” được tạo thành bởi 3 từ “Tôi”, “là”, “giáo viên” và việc xác định từ không đơn thuần là việc sử dụng dấu cách để ngắt văn bản.

Hai là sự đa nghĩa của từ trong tiếng Việt. Tiếng Việt là một ngôn ngữ phong phú, trong tập hợp từ vựng tiếng Việt có rất nhiều các từ mà cách viết giống nhau nhưng mang ý nghĩa khác nhau. Hiện tượng đa nghĩa không phải chỉ tiếng Việt mới có, cuốn *“Nhập môn ngôn ngữ học”* của tác giả Lê Đình Tư và Vũ Ngọc Cân đã khẳng định hiện tượng đa

nghĩa là hiện tượng phổ biến trong mọi ngôn ngữ. Mặc dù vậy, việc xử lý ngôn ngữ tiếng Việt khó khăn ở chỗ có các từ viết giống hệt nhau nhưng lại thuộc các loại từ khác nhau và động từ trong tiếng việt thì không hề thay đổi hình thái trong mọi ngữ cảnh. Ví dụ như từ "Ứng dụng" có thể hiểu là động trong câu "Các nhà khoa học *ứng dụng* công nghệ tế bào gốc trong chữa bệnh" nhưng lại là danh từ trong câu "Viettel vừa cho ra mắt *ứng dụng* lọc tin nhắn rác trên điện thoại di động".

Thứ ba, chữ viết của tiếng Việt là chữ ghi âm, loại chữ không biểu hiện ý nghĩa của từ mà tái hiện chuỗi âm thanh tiếp nối của từ. Ngữ nghĩa của một câu tiếng Việt đôi khi phụ thuộc vào cách ngắt nghỉ, âm điệu trầm bổng của người nói vì thế tồn tại những câu mà ngay cả một người thạo tiếng Việt cũng không thể hiểu nếu không được nghe tác giả đọc câu đó lên. Ví dụ như câu: "Ông già đi nhanh quá" hay "Học sinh học rất vui".

CHƯƠNG 3. ĐỀ XUẤT MẠNG NƠRON ĐÁNH GIÁ NGỮ NGHĨA

3.1. BÀI TOÁN

Nhiệm vụ của đề tài là khái quát cơ sở lý thuyết, kế thừa những nghiên cứu đã có, đề xuất giải pháp và xây dựng được một chương trình có khả năng phát hiện ra những cặp câu luật giao thông Việt Nam được thể hiện khác nhau nhưng mang ý nghĩa giống nhau. Đề tài sử dụng luật giao thông làm đối tượng áp dụng nghiên cứu nhằm giới hạn ngữ cảnh, giới hạn kích thước bộ từ vựng, tăng tốc thời gian huấn luyện các hệ nơron phù hợp với phạm vi khóa luận tốt nghiệp của sinh viên.

Để tiện cho việc trình bày các kết quả nghiên cứu, bài toán thực nghiệm được mô tả như sau:

Đầu vào của hệ thống là một tập các câu phát biểu về các chế định trong luật giao thông đường bộ Việt Nam chứa trong đó các câu có cùng ý nghĩa, được xáo trộn trật tự từ nhưng vẫn đảm bảo đúng chính tả và bảo tồn ý nghĩa câu.

Đầu ra của hệ thống: Với mỗi câu trong tập đầu vào, hệ thống cần tìm ra được tập các câu có ý nghĩa gần với nó nhất.

$$INPUT: D = \{s_1, s_2, \dots, s_n\}$$
$$OUTPUT: D_i = \{s_j \,|\, mean(s_j) \cong mean(s_j)\} \; \forall s_i$$

Ví dụ: Trong tập các câu nói về luật giao thông được đưa vào làm đầu vào của hệ thống. Tập đầu ra trong điều kiện lý tưởng ứng với câu "*cấm lạng lách đánh võng trên đường*" bao gồm:

1. "người lái xe không được lạng lách đánh võng"
2. "nghiêm cấm đánh võng đối với người điều khiển xe máy"
3. "lạng lách đánh võng là hành vi trái pháp luật"
4. "người điều khiển phương tiện không được lạng lách đánh võng"

3.2. CÁC NGHIÊN CỨU ĐƯỢC KẾ THỪA

Mô hình được đưa ra trong cách tiếp cận này có sự tham khảo, học tập từ những nghiên cứu về Xử lý ngôn ngữ tự nhiên, Học máy với một số lượng lớn các công việc liên quan khác. Ý tưởng chính của giải pháp này là sử dụng các vector để biểu thị ngữ nghĩa của một từ và sự kết hợp của chúng trong câu luật giao thông.

Như đã phân tích ở phần cơ sở lý thuyết về đặc điểm của tiếng Việt, để đạt được mục tiêu có được một hệ thống hiệu quả làm việc với dữ liệu tiếng Việt, cần sử dụng cơ chế giúp giảm thiểu sự nhập nhằng trong tiếng Việt gây ra bởi tính đồng nghĩa của các từ khác nhau. Theo cuốn *"Nhập môn ngôn ngữ học"* của tác giả Lê Đình Tư và Vũ Ngọc Cân *"Ngữ cảnh, nói một cách đơn giản, là tình huống, bối cảnh ngôn ngữ, trong đó từ xuất hiện với một ý nghĩa cụ thể của nó. Thông qua ngữ cảnh, ta có thể xác định được những yếu tố hạn chế phạm vi ý nghĩa của từ, làm cho nghĩa được sử dụng nổi rõ lên."*[3]. Chú ý đến yếu tố ngữ cảnh khi làm việc với các từ tiếng Việt, trong công trình nghiên cứu, không gian vector mô tả ngữ nghĩa của từ được xây dựng dựa trên ý tưởng của Eric H. Huang, Richard Socher, Christopher D. Manning và Andrew Y. Ng trong bài báo *"Improving Word Representations via Global Context and Multiple Word Prototypes"* (2012)[4], đó là mô hình học có giám sát có thể học ngữ nghĩa của vector từ cả ngữ cảnh cục bộ và ngữ cảnh toàn cục.

Trong mối quan hệ về ngữ nghĩa, từ là đơn vị nhỏ nhất cấu tạo nên câu. Trên cơ sở của các vector từ, mạng nơron phát hiện sự đồng nghĩa của các câu được xây dựng theo ý tưởng được đề xuất trong bài báo *"Grounded Compositional Semantics for Finding and Describing Images with Sentences"* của Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, Andrew Y. Ng (2013) [5]. Mạng nơron được đặt tên là *Mạng nơron hồi quy dựa trên cây phụ thuộc* (DT-RNN) sử dụng một mạng nơron hồi quy (Recursive Neural Network) được triển khai trên nền của cây phụ thuộc (Dependency tree) khi khai triển các câu. Cây phụ thuộc là một trong những hướng nghiên cứu lớn của xử lý ngôn ngữ tự nhiên, công trình này sử dụng kết quả nghiên cứu của Dat Quoc Nguyen, Dai Quoc Nguyen, Son Bao Pham, Phuong-Thai Nguyen và Minh Le Nguyen

[3] Nhập môn ngôn ngữ học. Hà Nội, 2009. Lê Đình Tư & Vũ Ngọc Cẩn

[4] H. Huang, R. Socher, C. D. Manning, and A. Y. Ng.2012. Improving Word Representations via Global Context and Multiple Word Prototypes. In ACL

[5] Richard Socher, Andrej Karpathy, Quoc V. Le*, Christopher D. Manning, Andrew Y. Ng. Grounded Compositional Semantics for Finding and Describing Images with Sentence

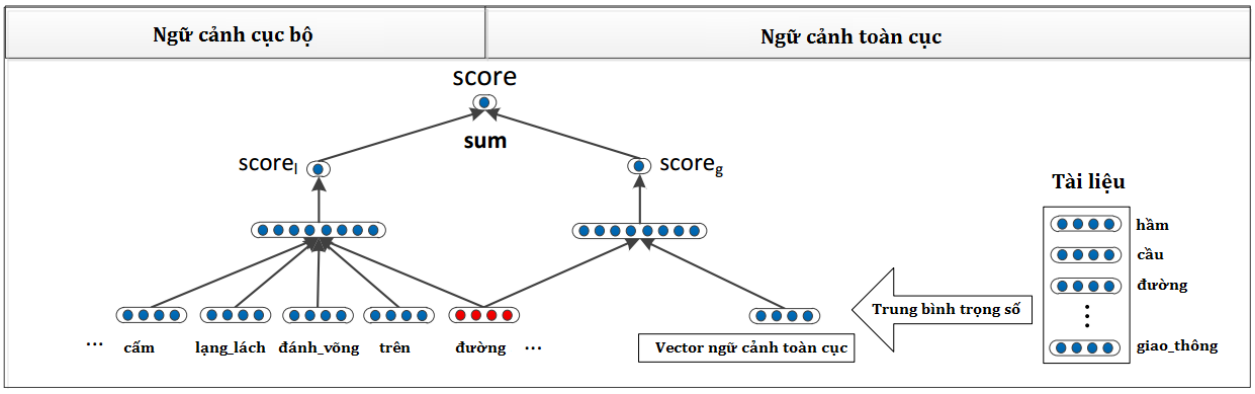
trong đề tài *"From Treebank Conversion to Automatic Dependency Parsing for Vietnamese"* (2014), một công trình có đóng góp lớn khi xây dựng được cây Phụ thuộc từ Treebank tiếng Việt[6].

3.3. XÂY DỰNG KHÔNG GIAN VECTOR TỪ

3.3.1. Tổng quan cách tiếp cận

Mô hình không gian vector mô tả nghĩa của từ sẽ thể hiện được cả những thông tin về ngữ nghĩa và hình thức của từ. Những mô tả đó có thể được dùng cho việc đo lường tính tương đồng về mặt ý nghĩa bằng cách đo khoảng cách giữa các vector thể hiện các từ. Đây là nghiên cứu gốc của nhiều ứng dụng hữu ích, nó cũng là tiền đề để xây dựng *Mạng nơron hồi quy dựa trên cây phụ thuộc* (DT-RNN) biểu thị ngữ nghĩa của câu trong công trình này.

Trong kỹ thuật sử dụng không gian vector mô tả ngữ nghĩa các từ tính đến thời điểm hiện tại, thì ý tưởng của Eric H. Huang và các đồng tác giả (2012) có ưu điểm hơn cả. Mô hình được đề xuất sử dụng cả ngữ cảnh cục bộ và ngữ cảnh toàn cục kết hợp trong mục tiêu huấn luyện. Vector được huấn luyện sẽ đảm bảo thể hiện tốt hơn ngữ nghĩa của từ mà vẫn giữ được hình thức của nó, các hiện tượng đồng âm, đồng nghĩa sẽ được giải quyết.



Hình 3.1 Mô hình đánh giá ngữ cảnh do Eric H. Huang và các đồng tác giả đề xuất 2012

[6] Đề tài KC01.01/06-10 "Nghiên cứu phát triển một số sản phẩm thiết yếu về xử lí tiếng nói và văn bản tiếng Việt" (VLSP)

Cho một chuỗi từ *s* và văn bản *d* chứa chuỗi đó, mục tiêu là phân biệt được chính xác từ cuối cùng trong *s* đối với các từ ngẫu nhiên khác. Theo đó, *g(s,d)* và $g(s^w,d)$ được tính toán, với s^w là chuỗi *s* được thay từ cuối bằng từ *w*, *g(.,.)* là hàm tính điểm mà mạng nơron sử dụng. Chúng ta mong muốn *g(s,d)* sẽ lớn hơn $g(s^w,d)$ với biên tối đa là 1. Do đó mục tiêu huấn luyện là tối thiểu hóa hàm giá:

$$C_{s,d} = \sum_{w \in V} max(0, 1 - g(s,d) + g(s^w,d))$$

3.3.2. Thu thập và tiền xử lý dữ liệu

Trong thời gian và giới hạn của khóa luận tốt nghiệp đại học, để hạn chế kích thước của corpus nhưng vẫn đảm bảo tính phủ của tập các vector từ đối với các câu về luật giao thông, corpus huấn luyện được lấy từ nguồn của Luật giao thông bao gồm Bộ luật giao thông (2008), các nghị định của chính phủ, các thông tư của các bộ và các website chuyên đưa tin giao thông[7].

Như đã trình bày ở phần cơ sở lý thuyết, do đặc trưng về việc phân tách từ của tiếng Việt, muốn xây dựng được một bộ vector biểu thị ý nghĩa cho các từ trong các câu mô tả luật giao thông Việt Nam, cần tiến hành tách từ cho dữ liệu đầu vào.Công cụ tách từ tiếng Việt sử dụng trong đề tài là vnTokenizer [8] của tác giả Lê Hồng Phương. Công cụ này sử dụng kết hợp từ điển và ngram, trong đó mô hình ngram được huấn luyện sử dụng treebank tiếng Việt (70,000 câu đã được tách từ) với độ chính xác trên 97%[9].

Tệp dữ liệu sau khi được thu thập và tách từ có dung lượng 10,9MB, chứa 4,290 từ, trong đó các cụm ký tự chứa số được chuyển đổi chung thành từ "NUMBER" để tránh ảnh hưởng tới độ chính xác của vector từ được sinh ra.

3.3.3. Xây dựng mạng nơ ron

Mạng nơron tính điểm cho một chuỗi từ (có thể hiểu là một câu) thông qua hai bước là tính trên ngữ cảnh cục bộ và ngữ cảnh toàn cục, sau đó điểm số cuối cùng cho

[7] http://www.vovgiaothong.vn, http:// www.gttm.go.vn,http://www.mt.gov.vn, http://www.baogiaothong.vn
[8] http://mim.hus.vnu.edu.vn/phuonglh/softwares/vnTokenizer
[9] Đây là công cụ thuộc Đề tài KC01.01/06-10 "Nghiên cứu phát triển một số sản phẩm thiết yếu về xử lí tiếng nói và văn bản tiếng Việt" (VLSP)

mỗi chuỗi từ là tổng điểm của hai bước tính toán trên. Điểm ứng với ngữ cảnh cục bộ sử dụng chuỗi từ cục bộ *s*. Chuỗi *s* được mô tả dưới dạng một danh sách được sắp xếp gồm các vector x = (x_1, $x_{2,...,}$ $x_{m)}$ với x_i là vector biểu thị từ thứ i trong chuỗi. Tất cả các vector biểu thị các từ trong Tập từ vựng tạo thành ma trận L.

$$L \in R^{n \times |V|}$$

Trong đó n là số chiều của vector từ, $|V|$ là lực lượng của tập từ vựng

Để tính toán điểm cục bộ, một mạng nơron gồm một lớp ẩn được sử dụng:

$$a_1 = f(W_1[x_1; x_2; \dots; x_m] + b_1$$

$$score_l = W_2 a_1 + b_2$$

Trong đó, $[x_1; x_2; \dots; x_m]$ là sự ghép nối các vector từ nằm trong chuỗi s, hàm f là một hàm kích hoạt, trong công trình sử dụng hàm $tanh$, W_1 và W_2 là ma trận trọng số, b_1 và b_2 là $bias$ mỗi lớp.

Nếu ngữ cảnh cục bộ được tính dựa trên sự ghép nối của *m* từ trong một chuỗi thì ngữ cảnh toàn cục được tính dựa trên sự ghép nối của *giá trị trọng số trung bình* của văn bản và từ cuối cùng của chuỗi. Giá trị trọng số trung bình của văn bản được tính theo công thức:

$$c = \frac{\sum_{i=1}^{k} w(t_i) d_i}{\sum_{i=1}^{k} w(t_i)}$$

Trong đó d_i là các vector từ trong d $w(.)$ là bất kỳ một hàm đánh trọng số nào, để giảm độ phức tạp tính toán, công trình sử dụng hàm đánh trọng số idf.

$$idf(t, D) = log \frac{|D|}{1 + |\{d \in D: t \in d\}|}$$

Trong đó, $|D|$ là tổng số văn bản trong corpus, thông thường $|\{d \in D: t \in d\}|$ thường được đặt dưới mẫu thức là số văn bản chứa từ t. Nếu từ đó không xuất hiện ở bất cứ văn bản nào trong tập thì mẫu số sẽ bằng 0 dẫn tới phép chia cho không không hợp lệ, vì thế người ta thường thay bằng $1+|\{d \in D: t \in d\}|$. Trọng số này trong xử lý ngôn ngữ tự nhiên thường được dùng để loại bỏ các từ ít ý nghĩa (stop-word), những từ xuất hiện

nhiều mà lại đóng ít ý nghĩa cho câu. Nhưng cụ thể trong công việc này, nó chỉ đóng vai trò biểu thị mối quan hệ xác suất giữa từ và ngữ cảnh toàn cục trong câu.

Sau khi đã tính được *giá trị trọng số trung bình* của văn bản, ta tiến hành tính điểm chuỗi s trong ngữ cảnh toàn cục, quá trình tính điểm toàn cục được thể hiện bởi công thức:

$$a_1{}^{(g)} = f(W_1{}^{(g)}[c; x_m]) + b_1{}^{(g)}$$

$$score_g = W_2{}^{(g)} a_1{}^{(g)} + b_2{}^{(g)}$$

Trong đó a_1 là giá trị hàm kích hoạt của vector lớp ẩn, $[c; x_m]$ là sự ghép nối giữa vector *giá trị trọng số trung bình c và* vector từ cuối cùng trong cầu x_m để tạo ra một vector nhiều phần tử hơn, làm đầu vào của hàm kích hoạt. $score_g$ là điểm toàn cục được tính bởi tổng các giá trị của nơron lớp ẩn được trọng số hóa. Cuối cùng, điểm của của cặp (s,d) được tính bởi công thức:

$$g(s, d) = score_l + score_g$$

Trong đó $g(s, d)$ là hàm tính điểm tổng hợp, $score_l$ và $score_g$ lần lượt là hàm tính điểm cục bộ và toàn cục của câu đã được trình bày ở phía trước.

3.3.4. Huấn luyện mạng và hiệu chỉnh vector từ

Khác với các mạng nơron được tạo ra với mục tiêu dự đoán kết quả đầu ra khi biết đầu vào, mạng nơron này sử dụng kết quả đầu ra như cơ sở để thực hiện quá trình truyền ngược thông số đến vector từ và hiệu chỉnh chúng. Điểm khác biệt thứ hai đó là trong các mạng nơron thông thường, yếu tố được cập nhật chỉ là trọng số của mạng nơron, còn trong mạng nơron này yếu tố được cập nhật bao gồm cả trọng số của mạng và vector từ tương ứng. Do vậy việc tính giá trị cập nhật của thuật toán Gradient descent được tính trên cả đạo hàm của sai số đối với trọng số mạng và đạo hàm đối với các nơron đầu vào (ở đây là các thành phần của các vector từ)

Cho một chuỗi từ *s* và văn bản *d* chứa chuỗi đó, mục tiêu là phân biệt được chính xác từ cuối cùng trong *s* đối với các từ ngẫu nhiên khác. Khi mạng nơron có thể làm được như vậy, các vector từ sẽ được hiệu chỉnh trong mối quan hệ của ngữ cảnh toàn cục và ngữ cảnh cục bộ. Theo đó, *g(s,d)* và *g(s^w,d)* được tính toán, với s^w là chuỗi *s* được thay từ

cuối bằng từ *w*, *g(.,.)* là hàm tính điểm mà mạng nơron sử dụng. Để huấn luyện mạng, chúng ta sử dụng hàm giá theo dạng mô hình máy vector hỗ trợ (SVM), tạo ra một siêu phẳng phân tách được những câu thực tế tồn tại trong ngữ pháp tiếng Việt trong ngữ cảnh đoạn văn cụ thể và những câu bị chỉnh sửa do có từ cuối bị thay đổi. Ta mong muốn *g(s,d)* sẽ lớn hơn $g(s^w,d)$ với biên tối đa là 1. Do đó mục tiêu huấn luyện là tối thiểu hóa hàm giá:

$$C_{s,d} = \sum_{w \in V} max(0, 1 - g(s,d) + g(s^w, d))$$

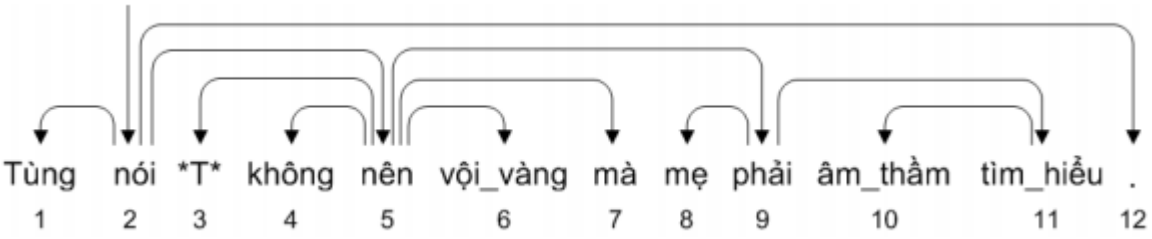
Trong một văn bản là ngữ cảnh toàn cục, đối với ngữ cảnh cục bộ, mỗi cửa sổ 10-từ được đưa vào huấn luyện. Các cặp ví dụ và phản ví dụ được tạo ra bằng cách đổi từ cuối cùng trong cửa sổ bằng một từ bất kỳ khác. Đối với mỗi cặp ví dụ và phản ví dụ, ta chạy một bước Stochastic gradient descent để tối ưu dần các tham số bao gồm các thông số của mạng nơron và các vector từ trong ví dụ. Các vector từ được huấn luyện trong phần này là đầu vào của mạng nơron phân biệt ngữ nghĩa câu được trình bày ở phần tiếp theo của khóa luận.

3.4. MẠNG NƠRON ĐÁNH GIÁ NGỮ NGHĨA

3.4.1. Phân tách cú pháp bằng cây phụ thuộc

Cây phụ thuộc là một trong những hướng nghiên cứu lớn của xử lý ngôn ngữ tự nhiên, nó được xây dựng dựa trên mối quan hệ phụ thuộc của các từ trong một câu. Các từ trong cây phụ thuộc được nối bằng liên kết có hướng, trong đó động từ là trung tâm về mặt cấu trúc của câu, tất cả các thành phần còn lại được nối trực tiếp hoặc gián tiếp tới động từ, cấu trúc này tạo điều kiện thuận lợi để xây dựng một mạng nơron có các trọng số được cập nhật để phân loại được ý nghĩa của câu. Xử lý ngôn ngữ tự nhiên tiếng Việt là một hướng nghiên cứu mới, vì vậy đến năm 2014 mô hình cây phụ thuộc cho tiếng Việt mới được đề xuất bởi Dat Quoc Nguyen, Dai Quoc Nguyen, Son Bao Pham,

Phuong-Thai Nguyen và Minh Le Nguyen với đề tài xây dựng cây Phụ thuộc cho tiếng Việt từ Treebank tiếng Việt[10].



Hình 3.2 Ví dụ câu phụ thuộc của Dat Quoc Nguyen và các đồng tác giả.

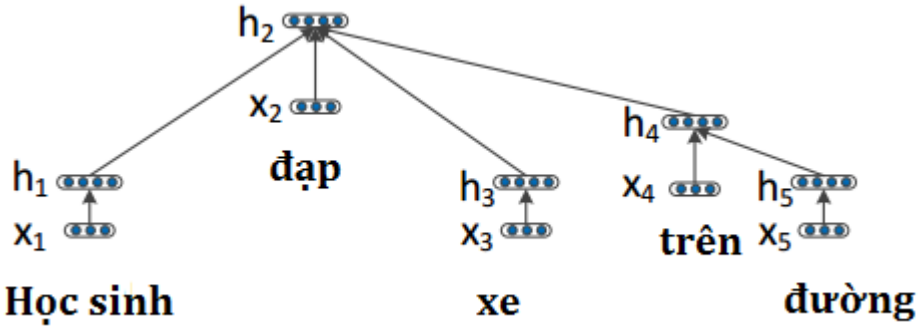
Công trình này sử dụng kết quả nghiên cứu của Dat Quoc Nguyen và các đồng tác giả trong việc tiền xử lý các phát biểu luật giao thông ở tập huấn luyện (D_{learn}) và tập kiểm thử (D_{test}) để chuyển sang dạng cây phụ thuộc sau khi được tách từ.

3.4.2. Xây dựng mạng nơron đánh giá ngữ nghĩa trên cây phụ thuộc

Ý tưởng này dựa trên bài báo của Richard Socher và các đồng tác giả (2013). Khi đã tách được đầu vào sử dụng cây phụ thuộc, ta có thể biểu diễn mỗi câu s dưới dạng danh sách các chỉ số $d(s) = \{(i; j)\}$ trong đó $i = 1; \ldots; m$ đại diện cho các từ có một từ nào đó có chỉ số $j \in \{1; \ldots; m\} \cup \{0\}$ là nút cha của nó. Từ nằm ở gốc có chỉ số cha bằng 0.

Lớp ẩn của mỗi nút được tính từ đầu vào (vector từ) của chính nút đó và lớp ẩn của các từ là con của nút đó.

[10] Đề tài KC01.01/06-10 "Nghiên cứu phát triển một số sản phẩm thiết yếu về xử lí tiếng nói và văn bản tiếng Việt" (VLSP)



Hình 3.3 Mạng nơron tính điểm dựa trên cây phụ thuộc

Ví dụ trên giúp hiểu rõ hơn ý tưởng của mạng nơron này, với câu *Học_sinh*$_1$ *đạp*$_2$ *xe*$_3$ *trên*$_4$ *đường*$_5$, ta có cây phụ thuộc như ở Hình 3.3 Mạng nơron tính điểm dựa trên cây phụ thuộc. Cây phụ thuộc có thể được thể hiện bằng các cạnh (cha,con) như sau: $d = \{(1,2),(2,0),(3,2),(4,2),(5,4)\}$. Mô hình mạng nơron đề quy trên cây phụ thuộc sẽ tính toán vector của nút cha tại mỗi từ thông qua vector biểu diễn từ đó và các vector thuộc lớp ẩn của nút con trực tiếp của nó. Chương trình sẽ tính đệ quy cho đến ghi nó gặp được nút không có nút con nào hoặc đã nút được tính toán từ trước. Sau đó các giá trị được tính ngược lên trên và cuối cùng cho ra được vector biểu thị câu.

Ví dụ trên có thể được tổng quát hóa bằng công thức:

$$h_4 = g_\theta(x_4, h_5) = f(W_v x_4 + W_{r1} h_5)$$
$$h_2 = g_\theta(x_2, h_1, h_3, h_4) = f(W_v x_2 + W_{l1} h_1 + W_{r1} h_3 + W_{r2} h_4)$$

Trong đó g_θ là hàm kết hợp (tính tổng) được tham số hóa bằng các trọng số của mạng nơron, f được sử dụng là hàm kích hoạt của mạng nơron (trong đề tài sử dụng hàm $tanh$), W_v là ma trận trọng số đối với các vector từ, các ma trận $W_{l1}, W_{l2} \dots W_{r1}, W_{r2} \dots$ là các ma trận trọng số đối với vector lớp ẩn của các nút con trực tiếp của nút hiện tại. Việc xác định ma trận nào sẽ sử dụng để trọng số hóa vector lớp ẩn nào phụ thuộc vào vị trí tương đối của nút cha và nút con.

Khi tất cả các câu đã được vector hóa (vector tại lớp ẩn của nút gốc của cây phụ thuộc), sự đồng nghĩa của hai câu được thể hiện bởi giá trị tích vô hướng của hai vector. Hai vector càng có tích vô hướng lớn thì chúng ta hiểu rằng hai câu có nghĩa càng gần nhau.

3.4.3. Huấn luyện mạng nơron tính điểm

Để huấn luyện mạng nơron, ta thiết lập các bộ, mỗi bộ gồm 1 câu thể hiện nội dung của một chế định luật giao thông và các biến thể của nó được đưa vào huấn luyện. Ta gọi S là toàn bộ các câu đưa vào huấn luyện, $S(i)$ là tập các biến thể của câu luật giao thông i(bao gồm chính nó). Hàm mục tiêu cần tối thiểu là:

$$J_{train}(\theta) = \sum_{i,j \in S(i)\ k \in S \setminus S(i)} max(0, \Delta - h(i).h(j) + h(i).h(k))$$

Hàm mục tiêu được sử dụng ở đây cũng có dạng hàm của mô hình máy vector hỗ trợ (SVM), các tham số θ được tối ưu hóa qua mỗi bước của thuật toán Stochastic gradient descent biến mạng nơron của chúng ta có dạng một siêu phẳng có khả năng tách được tập các câu có cùng ý nghĩa và các cặp câu không cùng ý nghĩa. Khi hàm mục tiêu được tối thiểu, các câu cùng ý nghĩa sẽ có tích vô hướng được đánh giá cao hơn so với các cặp câu xa nghĩa.

3.5. PHÂN TÍCH ĐIỂM MẠNH, HẠN CHẾ PHƯƠNG PHÁP

3.5.1. Điểm mạnh của phương pháp

Đánh giá một cách khách quan, có thể thấy phương pháp được đề xuất trong chương này có một số điểm mạnh sau đây:

- Đưa ra được một cách tiếp cận phù hợp, giải quyết được yêu cầu của bài toán nghiên cứu.
- Giải quyết được các vấn đề phát sinh khi thao tác với ngôn ngữ tiếng Việt như đã phân tích trong phần cơ sở lý luận. Đặc biệt là giải quyết một cách tốt nhất vấn đề đa nghĩa trong tiếng Việt bằng cách sử dụng vector từ được huấn luyện gắn với ngữ cảnh và sử dụng cây phụ thuộc.
- Kế thừa những ý tưởng từ những nghiên cứu mới nhất trong nước và quốc tế về lĩnh vực xử lý ngôn ngữ tự nhiên và xử lý ngôn ngữ tự nhiên tiếng Việt của các nhà khoa học có uy tín.

3.5.2. Hạn chế của phương pháp

Bên cạnh những điểm mạnh, phương pháp còn tồn tại một số hạn chế sau đây:

- Hạn chế của phương pháp chứa trong đó là hạn chế của mô hình mạng nơron nhân tạo, để có thể hoạt động chính xác, mạng nơron nhân tạo cần nhiều tài nguyên như dữ luyện huấn luyện, chi phí về mặt thời gian, tính toán và bộ nhớ cho hoạt động huấn luyện.
- Độ chính xác của phương pháp bị giới hạn bởi độ chính xác của chương trình tách từ tiếng Việt (~98%) và chương trình sinh cây phụ thuộc (~80%).
- Việc phân tích các đặc điểm của tiếng Việt để áp dụng các nghiên cứu của nước ngoài đối với dữ liệu tiếng Anh còn chưa có nhiều cơ sở khoa học chắc chắn.

CHƯƠNG 4. THỰC NGHIỆM, KẾT QUẢ, VÀ SO SÁNH ĐÁNH GIÁ

4.1. CÔNG CỤ VÀ MÔI TRƯỜNG THỰC NGHIỆM

4.1.1. Win web crawler - chương trình lấy nội dung của các trang web

Như đã phân tích ở các chương trước, mạng nơron sẽ chỉ là một chương trình máy tính không có giá trị nếu như thiếu đi các dữ liệu huấn luyện. Mạng nơron được đề xuất trong đề tài hoạt động trực tiếp với dữ liệu là từ vựng tiếng Việt nên yêu cầu đối với dữ liệu thu thập được cần thỏa mãn những tiêu chí về độ lớn, tính bao phủ, tính ngẫu nhiên và khách quan. Để đạt được những tiêu chí về dữ liệu như vậy, nguồn dữ liệu về từ vựng được sử dụng trong đề tài được lấy từ những trang web chuyên đưa tin về giao thông[11] và các trang web chứa các văn bản pháp luật[12].

Công trình sử dụng chương trình Win web crawler được cung cấp miễn phí trên trang web http://www.winwebcrawler.com/ để làm công cụ thu thập dữ liệu từ vựng. Chương trình có thiết kế đơn giản và đủ chức năng giúp thu thập từ vựng một cách nhanh chóng. Tệp từ vựng được lấy từ các website thông qua chương trình chứa hơn 4.000 từ, đạt yêu cầu về độ bao phủ, tính ngẫu nhiên và tính khách quan. Tệp có dung lượng 10.9MB sau khi được xử lý tách từ.

4.1.2. vnTokenizer - công cụ tách từ tiếng Việt

Do đặc trưng của tiếng Việt, bộ công cụ tách từ là yêu cầu không thể thiếu nếu muốn thao tác với dữ liệu tiếng Việt. Công cụ tách từ tiếng Việt sử dụng trong đề tài là vnTokenizer [13] của tác giả Lê Hồng Phương. Công cụ này sử dụng kết hợp từ điển và ngram, trong đó mô hình ngram được huấn luyện sử dụng treebank tiếng Việt (70,000 câu đã được tách từ) với độ chính xác trên 97%[14].

Tác giả của chương trình cung cấp cả file nhị phân và mã nguồn của công cụ cho mục đích nghiên cứu khoa học. Công trình sử dụng trực tiếp mã nguồn của công cụ để có

---

[11] http://www.vovgiaothong.vn, http:// www.gttm.go.vn,http://www.mt.gov.vn, http://www.baogiaothong.vn

[12] http:// thuvienphapluat.vn, http://www.vanban.chinhphu.vn

[13] http://mim.hus.vnu.edu.vn/phuonglh/softwares/vnTokenizer

[14] Đây là công cụ thuộc Đề tài KC01.01/06-10 "Nghiên cứu phát triển một số sản phẩm thiết yếu về xử lí tiếng nói và văn bản tiếng Việt" (VLSP)

thể tùy biến các chức năng, phục vụ cho các nhu cầu của công trình như chuẩn hóa file đầu vào của mạng nơron, chuẩn hóa tự động đầu vào của cây phụ thuộc hay sinh ra các bộ kiểm thử đơn giản cho mục đích gỡ rối chương trình.

Công cụ được đánh giá là có độ chính xác 97%, mặc dù có độ chính xác rất cao xong trong quá trình được sử dụng và áp dụng vào thực nghiệm của đề tài, công cụ này có lúc đã rơi vào 3% còn lại ví dụ như các từ "cấm_vận", "đường_không", "kéo_theo", "tính_từ". Sai sót của công cụ tuyệt vời này xảy ra là do đặc trưng của các câu luật giao thông, ví dụ như 2 tiếng *"tính"* và "*từ"* đứng cạnh nhau thì thường có xác suất cao tạo thành từ "tính_từ" nhưng rõ ràng trong câu *"không dừng xe, đỗ xe nơi đường giao nhau và trong phạm vi 5 mét* ***tính từ*** *đường giao nhau"* thì chúng không mang ý nghĩa như vậy. Để khắc phục những tồn tại nhỏ này, chương trình thực nghiệm của đề tài có bước chuẩn hóa lại dữ liệu đã tách để đảm bảo không xảy ra lỗi trong quá trình huấn luyện mạng nơron.

4.1.3. vndp - công cụ khai triển cây phụ thuộc tiếng Việt[15]

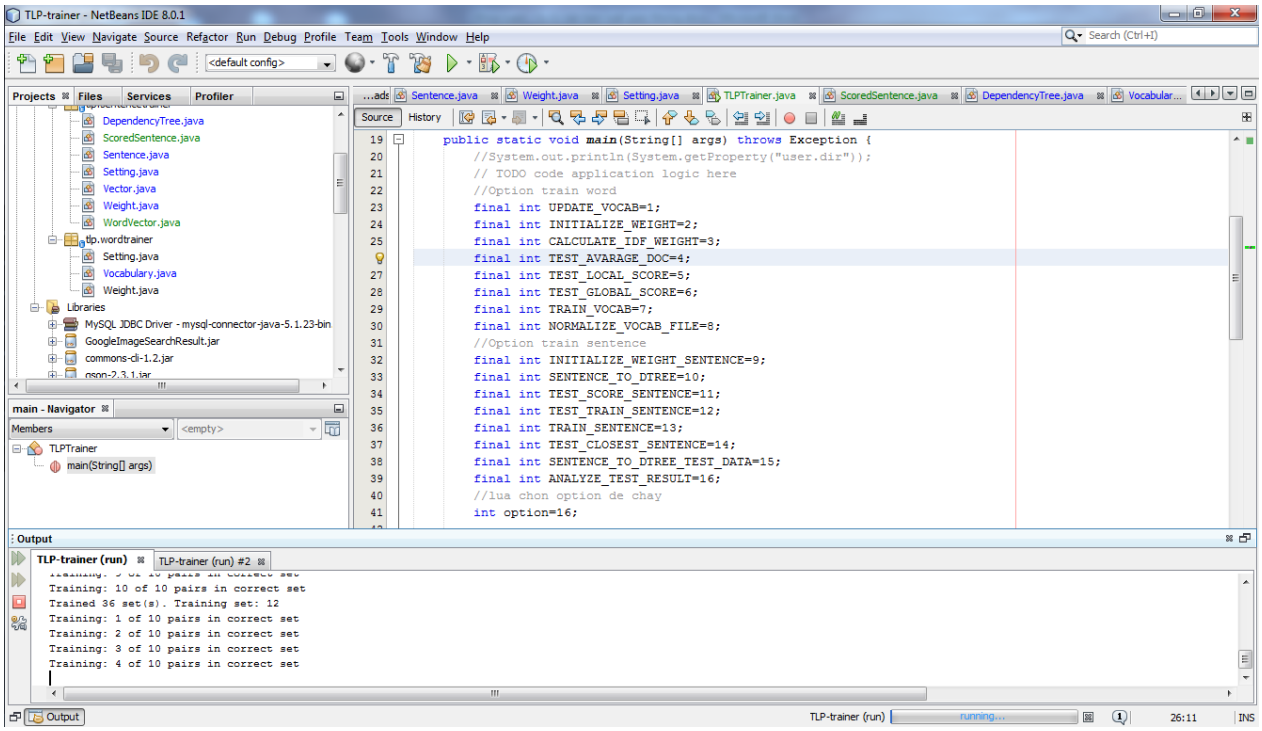
Đây là chương trình của các tác giả Dat Quoc Nguyen, Dai Quoc Nguyen, Son Bao Pham, Phuong-Thai Nguyen và Minh Le Nguyen thuộc đề tài *"From Treebank Conversion to Automatic Dependency Parsing for Vietnamese"*. Mã nguồn của chương trình được cung cấp miễn phí cho mục đích học tập và nghiên cứu khoa học, việc sử dụng công cụ này trong đề tài đã được thông qua sự đồng ý của tác giả.

Đặc điểm của công cụ này là mỗi lần hoạt động, công cụ mất một khoảng thời gian khá lâu để nạp bộ dữ liệu có dung lượng 160MB rồi mới có thể hoạt động được, bên cạnh đó, công cụ chỉ hỗ trợ việc đọc và ghi từ các tệp trên đĩa. Do đó việc sử dụng mã nguồn của công cụ như một phần của chương trình có nhiệm vụ tách câu sẽ làm giảm tốc độ của chương trình vì thời gian nạp dữ liệu và độ trễ do phải giao tiếp với đĩa cứng. Với dữ liệu khổng lồ mà mạng nơron phải tính toán thì việc giảm tốc độ cho mỗi câu khi đưa vào xử lý như vậy sẽ dẫn đến tổng thời gian vô ích của chương trình vô cùng lớn. Điều đó đồng nghĩa với việc ta phải chờ đợi lâu hơn hàng trăm lần để có một mạng nơron hoạt động tốt với các hàm giá hội tụ.

[15] http://vndp.sourceforge.net/

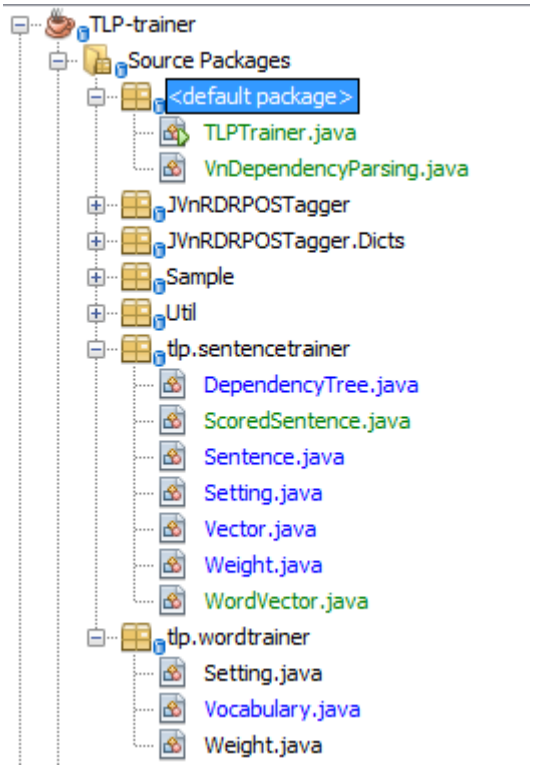
Chú ý đặc điểm đó của công cụ khai triển cây phụ thuộc tiếng Việt, công trình đề xuất cách giải quyết là tiến hành khai triển tất cả các câu trong tập huấn luyện và tập kiểm thử, sau đó tiến hành phân tích tệp kết quả rồi lưu vào cơ sở dữ liệu như một thuộc tính nằm kế bên giá trị của câu tương ứng. Mỗi lần câu được đưa vào huấn luyện hay tính điểm, chương trình chỉ cần thực hiện việc truy vấn cây phụ thuộc từ cơ sở dữ liệu chứ không cần phải khai triển cây phụ thuộc nữa.

4.1.4. Chương trình thực nghiệm tự thiết kế và triển khai



Hình 4.1. Bố trí mã nguồn của chương trình

Do mục đích của chương trình thực nghiệm là triển khai các phương pháp đã được đề xuất trong công trình, nhằm chứng minh tính đúng đắn của chúng, chương trình được thiết kế rất đơn giản và chỉ đủ cho các chức năng cần thiết tối thiểu để triển khai được các thuật toán, không có thiết kế giao diện người dùng. Chương trình được viết bằng java với IDE là Netbean, về cấu trúc mã nguồn, chương trình gồm 2 gói (package) chính, mỗi gói phụ trách hoạt động của một mạng nơron. Các lớp được thiết kế theo các thực thể được nhắc tới trong phương pháp đã đề xuất (bao gồm câu, từ, cây phụ thuộc, trọng số, vector) và các lớp hỗ trợ khác.



Hình 4.2. Các lớp của chương trình thực nghiệm

Tại hàm *main()* của chương trình, 16 lựa chọn được liệt kê, mỗi lựa chọn sẽ triển khai chương trình vào nhánh thực hiện một nhiệm vụ cụ thể. 16 lựa chọn đó bao gồm:

Bảng 4.1. Danh sách các tham số chạy chương trình

STT	Lựa chọn	Ý nghĩa
Đối với mạng nơron vector hóa từ vựng		
1	NORMALIZE_VOCAB_FILE	Tách từ, chuẩn hóa tệp từ vựng
2	UPDATE_VOCAB	Bổ sung từ vựng từ tệp vào cơ sở dữ liệu
3	INITIALIZE_WEIGHT	Khởi tạo trọng số của mạng nơron, lưu vào CSDL
4	CALCULATE_IDF_WEIGHT	Tính trọng số IDF để tính giá trị trọng số trung bình
5	TEST_AVARAGE_DOC	Tính giá trị trọng số trung bình phục vụ cho việc tính điểm câu trong ngữ cảnh toán cục.
6	TEST_LOCAL_SCORE	Kiểm thử việc tính điểm câu trong ngữ cảnh cục bộ
7	TEST_GLOBAL_SCORE	Kiểm thử việc tính điểm câu trong ngữ cảnh toàn cục
7	TRAIN_VOCAB	Huấn luyện mạng để vector hóa từ vựng

Đối với mạng nơron tính điểm sự đồng nghĩa của các câu		
9	INITIALIZE_WEIGHT_SENTENCE	Khởi tạo trọng số của mạng nơron, lưu vào CSDL
10	SENTENCE_TO_DTREE	Chuyển các câu thành dạng cây phụ thuộc và lưu vào cơ sở dữ liệu
11	TEST_SCORE_SENTENCE	Kiểm thử việc tính điểm sự tương đồng ý nghĩa
12	TEST_TRAIN_SENTENCE	Kiểm thử việc huấn luyện nơron (kiểm tra hướng hội tụ của hàm giá thông qua 1 ví dụ cụ thể)
13	TRAIN_SENTENCE	Tiến hành huấn luyện mạng dựa trên tập huấn luyện thật
14	TEST_CLOSEST_SENTENCE	Kiểm thử việc tìm ra câu sát nghĩa đối với câu truy vấn
15	SENTENCE_TO_DTREE_TEST_DATA	Chuyển các câu trong tập kiểm thử thành dạng cây phụ thuộc và lưu vào cơ sở dữ liệu
16	ANALYZE_TEST_RESULT	Đánh giá kết quả đầu ra khi cho mạng nơron chạy với dữ liệu kiểm thử

4.1.5. Môi trường thực nghiệm

Chương trình thực nghiệm được chạy hệ điều hành Windows 7 (64 bit), trên máy tính xách tay cá nhân có cấu hình như sau:

- Bộ vi xử lý: Intel(R) Core(TM)i5-2410M
- CPU: 2.30 GHz
- Ram: 4GB
- Hệ thống: 64 bit

4.2. DỮ LIỆU DÙNG CHO THỰC NGHIỆM

Như đã trình bày ở phần trước, dữ liệu là yếu tố hết sức quan trọng đối với tính hữu dụng của một mạng nơron. Để đảm bảo thỏa mãn những tiêu chí về độ lớn, tính bao phủ, tính ngẫu nhiên và khách quan cho mạng nơron huấn luyện vector từ, dữ liệu được lấy từ các website về giao thông[16] và các văn bản pháp luật[17].

[16]http://www.vovgiaothong.vn, http:// www.gttm.go.vn,http://www.mt.gov.vn, http://www.baogiaothong.vn
[17] http:// thuvienphapluat.vn, http://www.vanban.chinhphu.vn

Đối với mạng nơron phát hiện sự đồng nghĩa của câu, do ngữ cảnh của bài toán đã được giới hạn trong khuôn khổ luật giao thông Việt Nam, dữ liệu huấn luyện được sử dụng trong đề tài nhỏ hơn so với dữ liệu thực nghiệm được sử dụng trong công trình của Richard Socher và các đồng tác giả. Việc giới hạn ngữ cảnh đóng góp rất lớn cho tính khả thi của công trình tại điều kiện hiện tại. Dữ liệu huấn luyện gồm 110 bộ gồm 5 câu cùng nghĩa được xáo trộn về trật tự từ nhưng vẫn đảm bảo đúng ngữ pháp tiếng Việt.

Dữ liệu huấn luyện được chia làm 2 phần, 75 bộ đóng vai trò là dữ liệu học được sử dụng cho việc huấn luyện mạng nơron và 35 bộ sử dụng cho việc kiểm thử chất lượng của mạng. Từ 75 bộ của tập dữ liệu học ta sinh ra được 550.000 bộ ba các ví dụ chứa 2 câu cùng nghĩa và 1 câu khác nghĩa với hai câu còn lại, các bộ ba được tạo ra là đầu vào của mạng nơron đã được trình bày ở chương trước. Các bộ được sử dụng trong kiểm thử sẽ được trộn ngẫu nhiên và nhiệm vụ của chương trình là nhặt ra được trong đó các câu cùng bộ với nhau.

4.3. CÁCH THỨC TỔ CHỨC THỰC NGHIỆM

Quá trình thực nghiệm được tổ chức theo 4 bước: Thu thập dữ liệu, Tiền xử lý dữ liệu, Huấn luyện mạng nơron và Đo độ chính xác của mạng nơron.

Tại bước đầu tiên, các dữ liệu về từ vựng được thu thập một cách tự động bằng chương trình Win web crawler, tệp dữ liệu ban đầu có kích thước 9.2MB. Dữ liệu về các bộ của các câu thể hiện luật giao thông được thu thập từ các cộng tác viên thông qua một website được xây dựng tạm, phục vụ riêng cho mục đích nhập liệu cho khóa luận.

Sau khi được thu thập, các dữ liệu lưu trong tệp và CSDL được tiền xử lý, quá trình này bao gồm việc tách từ, chuyển các câu sang dạng cây phụ thuộc, chuyển các thông tin dạng số về cùng một từ thể hiện. Đây là bước rất quan trọng trước khi tiến hành huấn luyện dữ liệu.

Bước huấn luyện mạng nơron gồm hai quá trình con đó là vector hóa từ vựng sử dụng mạng nơron đầu tiên và huấn luyện mạng nơron thứ 2 để nó có khả năng phát hiện được những câu đồng nghĩa. Quá trình đầu tiên sử dụng mạng nơ ron có một lớp ấn chứa 20 nơron, lớp vào gồm 10 vector từ xếp cạnh nhau, mỗi vector từ chứa 10 phần tử và lớp ra có 1 nơron mang giá trị điểm của chuỗi 10 từ này. Quá trình thứ 2 sử dụng mạng nơ ron đệ quy có lớp ẩn và lớp vào đều có 10 phần tử. Cả hai quá trình đều sử dụng thuật

toán Stochastic Gradient Descent với hệ số học 10^{-5}, quá trình đầu tiên được chạy liên tiếp trong vòng 2 ngày, quá trình thứ 2 được chạy liên tiếp trong vòng 7 ngày trước khi hệ thống được đem ra đánh giá.

4.4. KẾT QUẢ THỰC NGHIỆM

Để đánh giá được mô hình, 175 câu (từ 35 bộ) được dùng làm đối tượng kiểm thử. Với mỗi câu, hệ thống sẽ tìm ra những câu gần nghĩa với nó nhất dựa vào tích vô hướng của các vector để rồi xếp hạng từ trên xuống dưới về mức độ gần nghĩa, sau đó hệ thống ghi lại thứ hạng của câu gần nó nhất thuộc cùng một bộ (gọi là mean rank). Giá trị mean rank càng nhỏ chứng tỏ mạng nơron hoạt động với ví dụ đó càng chính xác. Đây là cơ sở để đánh giá chất lượng của mạng nơron đã được huấn luyện. Bảng 4.2 liệt kê một số ví dụ của kết quả đầu ra

Bảng 4.2. Một số kết quả đầu ra ví dụ

Câu phát biểu về luật giao thông	Mean Rank
luật giao thông đường bộ quy định không được dừng xe, đỗ xe trong phạm vi an toàn của đường sắt.	0
không được để phương tiện giao thông ở lòng đường, hè phố trái quy định.	0
luật giao thông đường bộ không cho phép sử dụng lòng đường, lề đường, hè phố trái phép.	0
không kéo lê hàng hóa trên mặt đường.	1
nghiêm cấm dừng xe, đỗ xe nơi dừng của xe buýt	11
luật giao thông đường bộ quy định người điều khiển xe mô tô hai bánh, xe mô tô ba bánh, xe gắn máy không được đi xe vào phần đường dành cho người đi bộ và phương tiện khác.	71
nghiêm cấm dừng xe, đỗ xe trước cổng và trong phạm vi 5 mét hai bên cổng trụ sở cơ quan, tổ chức	141

Để đánh giá được chất lượng mạng nơ ron một cách định lượng, đề tài đề xuất cách đánh giá là ghi lại và thống kê tỷ lệ các câu trong tập dữ liệu kiểm thử có số mean rank nhỏ (cụ thể là 3 trường hợp mean rank = 0, mean rank < 5 và mean rank < 10). Bảng dưới đây ghi lại kết quả thống kê sơ bộ khi sử dụng mạng nơron phân tích dữ liệu kiểm thử

Bảng 4.3. Bảng thống kê mean rank

Trường hợp	Số trường hợp	Tỷ lệ %
Câu cùng bộ nằm trên cùng của bảng xếp hạng	48/175	27,43%
Câu cùng bộ nằm trong top 5 của bảng xếp hạng	86/175	49,14%
Câu cùng bộ nằm trong top 10 của bảng xếp hạng	115/175	65,71%

Xét trên cả tập dữ liệu kiểm thử, giá trị mean rank trung bình là 14,1. Nhìn vào kết quả đầu ra của mạng nơ ron và quá trình khảo sát gỡ rối khi cài đặt mạng, có thể thấy mạng hoạt động tốt với những câu có độ dài trung bình (từ 6-10 từ). Mạng nơ ron thể hiện kết quả kém đối với câu ngắn hơn hoặc dài hơn phạm vi đó, với những câu ngắn, cây phụ thuộc được sinh ra có độ sâu thấp và trở nên không linh hoạt trong việc cập nhật trọng số mạng trong quá trình huấn luyện, ngược lại, cây phụ thuộc sinh ra bởi câu dài có độ sâu lớn nên thường xảy ra trường hợp tràn bộ nhớ đối với các trọng số, dẫn đến tình trạng hội tụ ảo. Ngoài ra, kết quả phân loại của mạng đối với những câu rút gọn (khuyết chủ ngữ) kém hơn so với những câu có cấu trúc đầy đủ.

4.5. PHÂN TÍCH, ĐÁNH GIÁ KẾT QUẢ THỰC NGHIỆM

Dựa vào kết quả thực nghiệm, ta có thể thấy mạng nơron được cài đặt chính xác và hoạt động tương đối tốt với tập dữ liệu hiện tại. Tỷ lệ các câu cùng nhóm đứng đầu bảng xếp hạng, trong top 5 và top 10 lần lượt là 27,43% ; 49,14% và 65,71%, điều này chứng tỏ các trọng số của mạng nơron được tối ưu đúng hướng và hàm mục tiêu đang hội tụ dần về điểm cực trị, tỷ lệ này lớn hơn rất nhiều lần so với việc xếp hạng ngẫu nhiên đối với các câu.

Công trình sử dụng cùng một cách đánh giá với bài báo *"Grounded Compositional Semantics for Finding and Describing Images with Sentences"* của Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, Andrew Y. Ng. (2013) Mặc dù xử lý dựa trên ngôn ngữ tiếng Việt, với thời gian, dữ liệu hạn chế và độ chính xác bị giới hạn bởi độ chính xác của các công cụ xử lý ngôn ngữ tiếng Việt đã sử dụng nhưng công trình vẫn có giá trị mean rank ấn tượng là 14.1. Giá trị mean rank này chưa tốt bằng kết quả do mạng nơron được đề xuất trong công trình của Richard Socher và các đồng tác giả

sinh ra nhưng đã trội hơn một số phương pháp khác đối với dữ liệu tiếng Anh. Bảng dưới đây cho ta thấy rõ hơn điều đó[18]. Giá trị mean rank càng nhỏ chứng tỏ mô hình càng tốt.

Bảng 4.4 Bảng các giá trị mean rank của các phương pháp được khảo sát bởi Richard Socher và các đồng tác giả.

Mô hình	Giải thích	Mean Rank
Random	Mô hình ngẫu nhiên	101.1
BoW	Mô hình Bag of word	11.8
CT-RNN	Mô hình sử dụng mạng nơron đệ quy với cây bỏ phiếu	15.8
Recurrent NN	Mô hình mạng nơron hồi quy vòng	18.5
kCCA	Mô hình Kernel Canonical Correlation Analysis	10.7
DT-RNN	Mô hình mạng nơron đệ quy với cây phụ thuộc	11.1
SDT-RNN	Mô hình mạng nơron đệ quy với cây phụ thuộc ngữ nghĩa	10.5

[18] Bảng lấy từ số liệu trong công trình *"Grounded Compositional Semantics for Finding and Describing Images with Sentences"* của Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, Andrew Y. Ng. (2013)

CHƯƠNG 5. KẾT LUẬN

Trong khuôn khổ của một khóa luận tốt nghiệp đại học, đề tài đã giải quyết phần nào bài toán đặt ra là xây dựng chương trình có khả năng phát hiện các câu thể hiện luật giao thông có hình thái khác nhau nhưng biểu hiện ý nghĩa giống nhau. Bài toán nghiên cứu là cơ sở để phát triển các ứng dụng công nghệ cao trong lĩnh vực pháp lý sau này, giúp phát hiện, loại bỏ sự chồng chéo trong hệ thống pháp luật Việt Nam và hỗ trợ hữu ích cho những người đang hoạt động trong lĩnh vực pháp lý.

Hướng tiếp cận chính để giải quyết vấn đề là sử dụng kỹ thuật nơron nhân tạo trong học máy. Phương pháp thực nghiệm của đề tài phù hợp và có được những kết quả bước đầu khá ấn tượng. Đóng góp của công trình là đưa ra được cơ sở lý thuyết, đề xuất được một bài toán có ý nghĩa thực tiễn và xây dựng được một hệ thống hoạt động tương đối hiệu quả với dữ liệu là tiếng Việt dựa trên những công cụ, nghiên cứu đã có trước đó và một số cải tiến về kỹ thuật. Kết quả nghiên cứu của đề tài không chỉ có ý nghĩa trong giới hạn các văn bản luật giao thông mà có thể ứng dụng rộng rãi hơn trong các đề tài liên quan đến xử lý ngôn ngữ tự nhiên tiếng Việt khác.

Bên cạnh đó vẫn còn một số điểm có thể hoàn thiện trong các nghiên cứu tiếp theo để có một công trình hoàn chỉnh hơn:

- Thứ nhất, một cơ chế để người dùng có thể tham gia cải thiện việc học của hệ thống sẽ giúp tăng chất lượng phân loại của mạng nơron.
- Thứ hai, dữ liệu huấn luyện còn nghèo nàn, vẫn nhiều khâu phải nhập dữ liệu thủ công.
- Thứ ba, hàm giá đối với mạng nơron phát hiện sự đồng nghĩa chưa tối ưu, chưa tính điểm chính xác được cho câu có độ dài lớn do nơron bị tràn tìn hiệu.
- Thứ tư, ngôn ngữ lập trình Java chưa phải là ngôn ngữ phù hợp nhất để cài đặt mô hình mạng nơron nhân tạo.

Từ những phân tích nêu trên, các hướng đề xuất cải thiện hệ thống bao gồm:

- Thiết kế một ứng dụng web hoàn chỉnh phục vụ người dùng và sử dụng chính các thao tác, đánh giá của người dùng làm yếu tố đầu vào của mạng nơ ron và cải tiến việc học của mạng.
- Thiết kế lại hàm giá đối với mạng nơ ron để có thể giải quyết được cả những câu có độ dài lớn.

- Nghiên cứu xây dựng mạng nơron sử dụng các ngôn ngữ chuyên dụng hơn như Mathlab, Python...

TÀI LIỆU THAM KHẢO

Tiếng Anh

[1] Andrew Ng, "*Machine learning course - Stanford University"*, https://class.coursera.org/ml-005/lecture. Last visited: April 2015

[2] Collobert, Ronan, and Jason Weston. *"A unified architecture for natural language processing: Deep neural networks with multitask learning."* InProceedings of the 25th international conference on Machine learning, pp. 160-167. ACM, 2008.

[3] Huang, Eric H., Richard Socher, Christopher D. Manning, and Andrew Y. Ng. *"Improving word representations via global context and multiple word prototypes."* In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, pp. 873-882. Association for Computational Linguistics, 2012.

[4] Nguyen, Dat Quoc, Dai Quoc Nguyen, Son Bao Pham, Phuong-Thai Nguyen, and Minh Le Nguyen. *"From treebank conversion to automatic dependency parsing for vietnamese."* In Natural Language Processing and Information Systems, pp. 196-207. Springer International Publishing, 2014.

[5] Richard Socher, Andrej Karpathy, Quoc V. Le*, Christopher D. Manning, Andrew Y. Ng., *"Grounded Compositional Semantics for Finding and Describing Images with Sentences"* Transactions Of The Association For Computational Linguistics, 2, 207-218.

[6] Socher, Richard, Cliff C. Lin, Chris Manning, and Andrew Y. Ng. *"Parsing natural scenes and natural language with recursive neural networks."* InProceedings of the 28th international conference on machine learning (ICML-11), pp. 129-136. 2011.

[7] Socher, Richard, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. *"Recursive deep models for semantic compositionality over a sentiment treebank."* In Proceedings of the conference on empirical methods in natural language processing (EMNLP), vol. 1631, p. 1642. 2013.

[8] Socher, Richard, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. *"Semantic compositionality through recursive matrix-vector spaces."* InProceedings of

the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 1201-1211. Association for Computational Linguistics, 2012.

[9] Thi, Luong Nguyen, Hung Nguyen Viet, Huyen Nguyen Thi Minh, and Phuong Le Hong. *"Building a treebank for Vietnamese dependency parsing."* InComputing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), 2013 IEEE RIVF International Conference on, pp. 147-151. IEEE, 2013.

Tiếng Việt

[10] Lưu Tuấn Anh, *Xử lý ngôn ngữ tiếng Việt*, http://viet.jnlp.org/. Truy cập lần cuối Th 11, 2014

[11] *Hiến Pháp Nước Cộng Hòa Xã Hội Chủ Nghĩa Việt Nam***,** NXB Chính trị Quốc gia, 2013

[12] Trương Thị Diễm, "*Một số đặc trưng ngôn ngữ việt ảnh hưởng đến việc nhận thức tiếng việt của sinh viên nước ngoài"*, http://www.ued.edu.vn/khoavan/mod/resource/view.php?inpopup=true&id=59. Truy cập lần cuối: Th4, 2015

[13] GS. TS. Nguyễn Đăng Dung & TS. Nguyễn Minh Tuấn, *Giáo trình Luật hiến pháp Việt Nam, NXB Đại học Quốc gia Hà Nội*, 2014

[14] *Hệ thống Văn bản Quy phạm pháp luật*, NXB Hồng Đức, 2013

[15] Đào Kiến Quốc & Trương Ninh Thuận, *Giáo trình tin học cơ sở*, Đại học Quốc gia Hà Nội, Hà Nội, 2006 tr.7.

[16] Lê Đình Tư & Vũ Ngọc Cẩn, *Nhập môn ngôn ngữ học*, NXB Khoa học xã hội, Hà Nội, 2009.

[17] Vũ Xuân Tiền, *"Ma trận văn bản pháp luật về thuế"*, http://www.thesaigontimes.vn/125339/Ma-tran-van-ban-phap-luat-ve-thue.html, Truy cập lần cuối: Th4, 2015

[18] Đề tài KC01.01/06-10 *"Nghiên cứu phát triển một số sản phẩm thiết yếu về xử lí tiếng nói và văn bản tiếng Việt"* (VLSP) thuộc Chương trình Khoa học Công nghệ cấp Nhà nước KC01/06-10.