1 Giới thiệu về xử lý ngôn ngữ tự nhiên

Dave Bowman: Open the pod bay doors, HAL.
HAL: I'm sorry Dave, I'm afraid I can't do that.

Stanley Kubrick và Arthur C. Clarke,
cảnh trong phim *2001: A Space Odyssey*

Ý tưởng về việc máy tính có khả năng xử lý ngôn ngữ tự nhiên đã đến từ rất lâu ngay từ khi máy tính mới xuất hiện. Quyển sách này nói về các phương pháp và sự thực hiện ý tưởng thú vị đó. Chúng tôi giới thiệu về một lĩnh vực nghiên cứu sôi động, liên ngành với rất nhiều tên gọi khác nhau, như **xử lý ngôn ngữ và tiếng nói**, **công nghệ ngôn ngữ con người**, **xử lý ngôn ngữ tự nhiên**, **ngôn ngữ học tính toán**, và **tổng hợp và nhận dạng tiếng nói**. Mục tiêu của lĩnh vực này là giúp máy tính có thể thực hiện các nhiệm vụ liên quan đến ngôn ngữ của con người, những nhiệm vụ có sự tương tác giữa máy tính-con người, cải thiện giao tiếp con người-con người, hoặc đơn giản là xử lý văn bản và tiếng nói.

Một ví dụ cho lĩnh vực này là các **máy giao tiếp**. Chiếc máy tính HAL 9000 trong phim *2001: A space Odsyssey* của Stanley Kubrick là một trong những nhân vật đáng chú ý trong điện ảnh thế kỷ hai mươi. HAL là một chiếc máy thông minh có khả năng xử lý ngôn ngữ tiên tiến như hiểu và nói thành thạo tiếng Anh, thậm chí còn đọc được cử chỉ môi. Đến bây giờ, có thể thấy tác giả Clarke đã hơi lạc quan khi dự đoán rằng một chương trình thông minh như HAL có thể thành hiện thực. Nhưng trí tưởng tượng của ông đã đi xa như thế nào? Chúng ta gọi những chương trình như HAL là các **máy giao tiếp** hay các **hệ thống hội thoại**. Trong quyển sách này, chúng ta cùng tìm hiểu về các thành phần để làm nên một chương trình hội thoại hiện đại, gồm có xử lý ngôn ngữ đầu vào (**nhận dạng tiếng nói** hay **lĩnh vực hiểu ngôn ngữ**) và xử lý ngôn ngữ đầu ra (**sinh ngôn ngữ tự nhiên** hay **tổng hợp tiếng nói**)

Hãy chuyển qua một nhiệm vụ hữu ích khác, đã giúp cho rất nhiều người không nói tiếng anh có thể tiếp cận được với những thông tin khoa học khổng lồ trên các trang web tiếng Anh. Hay dịch cho những người nói tiếng Anh hàng trăm triệu

trang web viết bằng các ngôn ngữ khác như tiếng Trung Quốc. Mục tiêu của **dịch máy** là tự động dịch một văn bản từ ngôn ngữ này sang ngôn ngữ khác. Chúng tôi sẽ giới thiệu các thuật toán và các công cụ toán học cần thiết để hiểu một chương trình dịch máy hiện đại hoạt động như thế nào. Dịch máy là một vấn đề còn gặp nhiều thách thức, chúng tôi sẽ giới thiệu các thuật toán được sử dụng trong lĩnh vực này, cũng như các thành phần chính.

Rất nhiều nhiệm vụ xử lý ngôn ngữ liên quan đến Web. Một trong đó là **hệ thống hỏi đáp dựa trên nền Web**. Ví dụ như web tìm kiếm, khi người dùng có thể gõ một vài từ khóa để tìm kiếm thông tin, hay có thể hỏi một câu trọn vẹn, từ đơn giản đến phức tạp:

- Từ "bất đồng" có nghĩa là gì?
- Abraham Lincoln sinh năm nào?
- Mỹ có bao nhiêu bang trong năm đó?
- Người Trung Quốc xuất khẩu bao nhiêu lụa sang Anh cho đến hết thế kỷ 18?
- Các nhà khoa học nghĩ gì về vấn đề đạo đức trong nhân bản người?

Một vài câu hỏi ở trên là câu hỏi **rõ ràng**, hoặc đơn giản là câu hỏi **dữ liệu thực tế** như ngày tháng hay địa điểm, có thể trả lời bằng cách sử dụng các máy tìm kiếm. Nhưng trả lời các câu hỏi phức tạp hơn, có thể cần phải trích xuất thông tin trong các văn bản, hay phải thực hiện **suy diễn** (đưa ra kết luận từ các sự kiện), hoặc cần tóm tắt và tổng hợp thông tin từ nhiều nguồn. Trong quyển sách này, chúng ta nghiên cứu nhiều thành phần tạo nên các hệ thống hiểu ngôn ngữ hiện đại, như **trích rút thông tin**, **phân giải nhập nhằng nghĩa từ**.

Mặc dù các lĩnh vực con và các vấn đề chúng tôi vừa nêu còn rất xa mới được giải quyết triệt để, chúng là các lĩnh vực nghiên cứu sôi động và có nhiều kỹ thuật đã có các ứng dụng công nghiệp. Trong phần còn lại của chương này, chúng tôi trình bày tóm tắt những tri thức cần cho các nhiệm vụ này (và những nhiệm vụ khác như **sửa lỗi chính tả**, **kiểm tra ngữ pháp**), cũng như các mô hình toán học sẽ được giới thiệu trong sách.

1.1. TRI THỨC TRONG XỬ LÝ TIẾNG NÓI VÀ NGÔN NGỮ

Sự khác biệt giữa những ứng dụng xử lý ngôn ngữ và những ứng dụng xử lý dữ liệu thông thường là việc sử dụng *tri thức về ngôn ngữ*. Như chương trình Unix `wc`, được sử dụng để đếm số lượng ký tự, từ và dòng trong một file văn bản. Khi dùng để đếm ký từ và dòng, `wc` là một chương trình xử lý dữ liệu bình thường. Tuy nhiên,

khi nó được dùng để đếm số từ, nó cần biết được *tri thức về thế nào là một từ*, từ đó nó trở thành một chương trình xử lý ngôn ngữ.

Tất nhiên, `wc` là một hệ thống rất đơn giản với sự hiểu biết rất hạn chế về tri thức của ngôn ngữ. Những hệ thống giao tiếp phức tạp như HAL, hay hệ thống thống dịch máy, hay các hệ thống hỏi đáp, yêu cầu tri thức về ngôn ngữ rộng và sâu hơn rất nhiều. Để mường tượng được phạm vi và loại tri thức cần thiết, hãy xem xét những gì mà HAL cần biết để tham gia vào cuộc hội thoại như ở ví dụ đầu chương, hoặc trong trường hợp hệ hỏi đáp, hệ thống có thể trả lời được một trong những câu hỏi kể trên.

HAL cần phải nhận ra các từ từ tín hiệu âm thanh, sau đó sinh ra âm thanh từ một chuỗi từ. Các nhiệm vụ như nhận dạng tiếng nói và tổng hợp tiếng nói yêu cầu hiểu biết về ngữ âm và âm vị học; các từ được phát âm thế nào, và các âm được nhận ra như thế nào.

Chú ý rằng không giống như Star Trek's Commander Data, HAL có khả năng nói các câu phủ định như *can't*. Để có thể nói và tiếp nhận những điều này hay những điều khác (như nhận ra *doors* là số nhiều) đòi hỏi hiểu biết về **ngôn ngữ hình thái học**, các từ được tách thành các thành phần mang nghĩa như *số ít* hay *số nhiều*.

Chuyển đến các từ độc lập, HAL cần sử dụng một cấu trúc thích hợp để đưa ra câu trả lời. Ví dụ, HAL cần phải hiểu chuỗi từ này là vô nghĩa đối với Dave, mặc dù nó vẫn bao gồm chính xác số từ như trong câu gốc.

(1.1) I'm I do, sorry that afraid Dave I'm can't.

Tri thức cần thiết để sắp xếp và tổ chức các từ để trở thành câu được gọi là **cú pháp**. Giờ hãy xem xét một câu hỏi trong hệ thống hỏi đáp:

(1.2) Người Trung Quốc xuất khẩu bao nhiêu lụa sang Tây Âu cho đến hết thế kỷ 18?

Để trả lời câu hỏi này chúng ta cần biết về nghĩa từ, nghĩa của tất cả các từ (như xuất khẩu hay lụa) cũng như tổng hợp nghĩa (cụm từ Tây Âu so với Đông hoặc Nam Âu), từ *hết* có nghĩa là gì khi kết hợp với *thế kỷ 18*. Chúng ta cũng cần biết về liên hệ giữa từ với cấu trúc ngữ pháp. Ví dụ chúng ta cần biết cụm từ *cho đến hết thế kỷ 18* là một cụm từ chỉ thời gian, chứ không phải một đơn vị, như trong ví dụ dưới đây:

(1.3) Người Trung Quốc xuất khẩu bao nhiêu lụa sang Tây Âu thông qua qua các thương nhân miền nam?

HAL cũng cần phân biệt câu nói của Dave là một yêu cầu, thay vì là một câu mô tả hay câu hỏi, như các cách nói khác nhau dưới đây

YÊU CẦU: *HAL, open the pod bay door.*
MÔ TẢ: *HAL, the pood bay door is open.*
CÂU HỎI: *HAL, is the pod bay door open?*

Mặc dù HAL không nghe lời. Nhưng HAL hiểu đủ để trả lời một cách lịch sự với Dave. Nó có thể trả lời trống không, như *No* hay *No, I won't open the door*. Thay vào đó, đầu tiên nó nói xin lỗi *I'm sorry*, sau đó nó thêm vào cụm từ *I'm afraid* thay vì từ chối một cách trực tiếp *I can't*. Tri thức cần có cho loại câu này là tri thức về **ngữ dụng** hay **hội thoại**.

(1.4) Nước Mỹ có bao nhiêu bang trong *năm đó*?

Năm đó là năm nào? Để có thể hiểu được những từ như *năm đó*, hệ thống hỏi đáp cần xem xét các câu hỏi trước đó; trong trường hợp này, câu trước đó hỏi về năm Lincoln ra đời. Nhiệm vụ này gọi là **phân giải đồng tham chiếu**, sử dụng các từ như *đó* hay các đại từ như nó hoặc cô ấy được đề cập trong những câu trước đó.

Tổng kết lại, để có sử dụng thành thạo ngôn ngữ cần có các hiểu biết về các loại tri thức:

- Ngữ âm và âm vị học - tri thức về âm thanh trong tiếng nói
- Ngôn ngữ hình thái học - tri thức về ý nghĩa của các thành phần trong từ
- Ngữ pháp - tri thức về cấu trúc của các từ
- Ngữ nghĩa - tri thức về nghĩa
- Ngữ dụng - tri thức về sự liên hệ giữa mục tiếu và ý định của người nói
- Diễn ngôn - tri thức về các đơn vị ngôn ngữ nhiều hơn một câu

1.2. NHẬP NHẰNG

Một điều đáng chú ý về việc phân chia các tri thức về ngôn ngữ là phần lớn các nhiệm vụ trong xử lý ngôn ngữ và tiếng nói có thể xem như việc giải quyết **nhập nhằng** ở một cấp độ nhất định. Một đầu vào được gọi là **nhập nhằng** nếu có nhiều cấu trúc ngôn ngữ có thể chỉ đến nó. Xem xét câu nói *I made her duck*. Có năm cách hiểu câu trên (bạn có thể nghĩ thêm), mỗi một cách hiểu ví dụ cho sự nhập nhằng ở từng cấp độ:
(1.5) Tôi nấu món vịt cho cô ấy ăn.
(1.6) Tôi nấu con vịt của cô ấy
(1.7) Tôi tạo ra con vịt đồ chơi cho cô ấy.
(1.8) Tôi khiến cô ấy trở nên ngốc nghếch (như con vịt).
(1.9) Tôi biến cô ấy thành con vịt.

Các nghĩa khác nhau xảy ra bởi những sự nhập nhằng. Đầu tiên, từ *duck* và *her* là nhập nhằng về hình thái từ và cú pháp. Từ *duck* có thể là động từ hoặc một danh

từ, trong khi *her* có thể là đại từ tân ngữ hoặc đại từ sở hữu. Thứ hai, từ *make* cũng có sự nhập nhằng ngữ nghĩa, có thể có nghĩa là *make* (tạo) hoặc *cook* (nấu ăn). Cuối cùng, động từ *make* cũng dẫn đến sự nhập nhằng về cú pháp. *Make* có thể là ngoại động từ một tân ngữ (1.6), hoặc có thể là ngoại động từ hai tân ngữ (1.9), trong đó tân ngữ thứ nhất (*her*) trở nên giống như tân ngữ thứ hai (*duck*). Cuối cùng, *make* có thể nhận một tân ngữ trực tiếp và một động từ (1.8), tân ngữ thứ nhất (*her*) bị biến thành tân ngữ thứ hai (*duck*). Hơn nữa, trong một câu nói, thậm chí còn có một mức độ nhập nhằng sâu hơn; từ thứ nhất có thể là từ *eye* và từ thứ hai có thể là từ *maid*.

Chúng tôi thường xuyên giới thiệu các mô hình và giải thuật trong quyển sách này là cách để **khử** hay **phân giải** nhập nhằng. Ví dụ để quyết định *duck* là động từ hoặc danh từ có thể được giải quyết qua bài toán **gán nhãn từ loại**. Quyết định từ *make* có nghĩa là "create" (tạo) hay "cook" (nấu ăn) có thể giải quyết qua bài toán **khử nhập nhằng nghĩa từ**. Phân giải nhập nhằng từ loại và nghĩa từ là hai bài toán quan trọng trong **khử nhập nhằng từ vựng**. Rất nhiều các bài toán khác có thể cũng được coi là bài toán khử nhập nhằng từ vựng. Như một hệ thống tổng hợp tiếng nói đọc từ *lead* cần quyết định nó phát âm như trong *lead pipe* hay *lead me on*. Mặt khác, quyết định *her* và *duck* là một thành phần của một thực thể (như 1.5 hay 1.8) hay ở các thực thể riêng biệt (như 1.6) là một ví dụ của nhiệm vụ **khử nhập nhằng cú pháp** và có thể được giải quyết bằng phương pháp **probabilistic parsing** (phân tích câu dựa xác suất). Chúng ta cũng sẽ xem xét các sự nhập nhằng không xuất hiện trong ví dụ cụ thể này, như xác định một câu là câu trần thuật hay câu hỏi (có thể được giải quyết bằng việc **diễn giải hành động lời nói**).

1.3. CÁC MÔ HÌNH VÀ GIẢI THUẬT

Một trong những điểm nhấn trong 50 năm nghiên cứu về xử lý ngôn ngữ tự nhiên là rất nhiều kiến thức được trình bày ở phần trước có thể xử lý bằng một vài mô hình và giả thuyết. May mắn thay, các mô hình và giả thuyết này được xuất hiện như những công cụ cơ bản cho khoa học máy tính, toán học, ngôn ngữ học và có thể được áp dụng giống nhau trong các ngành này. Bên cạnh những mô hình quan trọng nhất như **state machines**, **rule systems**, **logic**, **probabilistic models** và **vector-space models**. Bản thân các mô hình này, lại dựa vào một vài giải thuật cơ bản, như các giải thuật tìm kiếm trong gian trạng thái như **dynamic programming**, hay các giải thuật về học máy như các phân lớp và **Expected-Maximization (EM)** và các giải thuật khác.

Ở dạng đơn giản nhất, state machines là các mô hình hình thức bao gồm trạng thái, chuyển giữa các trạng thái, và một biểu diễn đầu vào. Một vài biến thể của mô

hình cơ bản này như **deterministic** và **non-deterministic finite-state automata** và **finite-state transducers**.

Liên quan chặt chẽ đến các mô hình này là các hệ thống formal rule. Chúng ta sẽ xem xét những mô hình quan trọng nhất như **regular grammars** và **regular relations**, **context-free grammars**, và **feature-augmented grammars**. Các hệ thống máy trạng thái và formal rule là các công cụ chính để xử lý các tri thức về âm vị học, ngôn ngữ hình thái học và cú pháp.

Lớp mô hình thứ ba đóng vai trò quan trọng trong việc mô hình hóa tri thức là logic. Chúng ta sẽ tìm hiểu về **first order logic**, hay còn gọi là **predicate calculus**, cũng như các mô hình hình thức liên quan như lambda-calculus, feature-structues hay semantic primitives. Các biểu diễn logic này thường được sử dụng để mô hình hóa ngữ nghĩa và ngữ dụng, mặc dù các kỹ thuật gần đây thường tập trung vào các phương pháp hiệu quả hơn từ các ngữ liệu phi logic.

Mô hình xác suất đóng vai trò quan trọng trong việc mô hình hóa tất cả các cấp độ về ngôn ngữ. Các mô hình khác (state machines, formal rule systems hay logic) có thể tăng cường với xác suất. Ví dụ, state machine có thể tăng cường với xác suất để trở thành **weighted automaton** hay **Markov model**. Chúng ta cũng sẽ dành một thời lượng nhất định cho **hidden Markov models** hay **HMMs**, được sử dụng ở mọi nơi, trong gán nhãn từ loại, nhận dạng tiếng nói, hiểu hội thoại, tổng hợp tiếng nói và dịch máy. Một điểm quan trọng của mô hình xác xuất là khả năng có thể giải quyết được nhiều loại nhập nhằng đã đề cập trước đó; hầu hết các bài toán về xử lý ngôn ngữ và tiếng nói có thể coi như "cho N lựa chọn từ một đầu vào nhập nhằng, chọn lựa chọn có khả năng xảy ra cao nhất".

Cuối cùng, vector-space models, dựa vào đại số tuyến tính, được sử dụng trong bài toán trích rút thông tin và nghĩa từ.

Xử lý ngôn ngữ sử dụng những mô hình này thường liên quan đến bài toán tìm kiếm trong một không gian trạng thái biểu diễn các giả định từ một đầu vào. Trong nhận dạng tiếng nói, chúng ta cần tìm kiếm trong một không gian các chuỗi âm cho một từ đầu ra. Trong bài toán phân tích cú pháp, chúng ta cần tìm kiếm trong không gian các cây biểu diễn cú pháp cho một câu đầu vào. Trong dịch máy, chúng ta cần tìm kiếm trong không gian các giả định đầu ra cho một ngôn ngữ cho một câu đầu vào ở ngôn ngữ khác. Trong các bài toán không dựa xác suất, thường liên quan đến các máy trạng thái, chúng ta sử dụng các thuật toán phổ biến về đồ thị như **depth-first search**. Đối với các bài toán dựa xác suất, chúng ta thường sử dụng nhiều giải thuật heuristic như **best-first** hay **A* search**, và dựa vào các giải thuật quy hoạch động để giảm khối lượng tính toán.

Các công cụ học máy như **các bộ phân lớp** và **các mô hình chuỗi** đóng một vị trí quan trọng trong nhiều bài toán xử lý ngôn ngữ. Dựa vào tính chất của từng đối tượng, một bộ phân lớp thường gán mỗi đối tượng cho một lớp, trong khi một mô hình chuỗi cố gắng gán các chuỗi đối tượng thành chuỗi các nhãn tương ứng.

Ví dụ, trong bài toán xác định một từ có đúng chính tả hay không, các bộ phân lớp như **cây quyết định**, **support vector machines**, **Gaussian Mixture Models**, và **logistic regression** có thể sử dụng để làm các bộ phân lớp nhị phân (đúng hay không đúng) cho một từ ở một thời điểm. Các mô hình chuỗi như **hidden Markov models**, **maximum entropy Markov models**, và **conditional random fields** có thể gán nhãn đúng/không đúng cho tất cả các từ trong một câu cùng một lúc.

Cuối cùng, các nhà nghiên cứu xử lý ngôn ngữ sử dụng rất nhiều công cụ từ các nhà nghiên cứu trong lĩnh vực học máy - như sử dụng tập huấn luyện và tập test, các kỹ thuật thống kê như **cross-validation**, và việc đánh giá hệ thống đã huấn luyện một cách kỹ lưỡng.

1.4. NGÔN NGỮ, SUY NGHĨ VÀ SỰ HIỂU BIẾT

Với nhiều người, khả năng máy tính có thể xử lý được ngôn ngữ một cách thành thạo như con người là dấu hiệu của trí thông minh thực sự. Nguyên nhân của niềm tin này đến từ việc sử dụng ngôn ngữ hiệu quả gắn liền với các khả năng nhận thức của chúng ta. Một trong những người đầu tiên coi tác động của sự liên hệ đó là Alan Turing (1950). Trong bài báo nổi tiếng này, Turing giới thiệu một thử nghiệm mà sau này được biết với tên Turing test. Turing bắt đầu với một luận điểm, rằng câu hỏi "Liệu máy tính có thể suy nghĩ hay không?" là một câu hỏi rất khó trả lời chính xác vì sự nhập nhằng vốn có trong các cụm từ *máy tính* và *nghĩ*. Thay vào đó, ông đề xuất một phép thử thực tế, một trò chơi, trong đó việc sử dụng ngôn ngữ của máy tính cho thấy khả năng suy nghĩ của nó. Nếu máy tính có thể thắng trò chơi này, nó sẽ được đánh giá là thông minh.

Trong trò chơi của Turing, có ba bên tham gia: hai người và một máy tính. Một người đóng vai trò là người hỏi. Để chiến thắng, người hỏi cần phải xác định xem hai thành viên còn lại ai là máy tính bằng cách hỏi một chuỗi câu hỏi thông qua việc gõ bàn phím. Nhiệm vụ của máy tính là lừa người hỏi tin rằng nó là con người bằng cách trả lời các câu hỏi. Nhiệm vụ của người chơi thứ hai là thuyết phục người hỏi, người chơi kia là máy tính, mình mới là con người.

Đoạn hội thoại dưới đấy được lấy từ bài báo của Turing minh họa cho trò chơi. Rõ rằng, việc giả làm con người không yêu cầu kiến thức của tất cả các lĩnh vực

Q: Please write me a sonnet on the topic of Forth Bridge	Hãy viết một bài thơ về cầu Forth
A: Count me out on this one. I never could write poetry	Bỏ qua cho tôi câu này. Tôi chưa bao giờ làm thơ
Q: Add 34957 to 70764	Cộng 34957 với 70764
A: (Pause about 30 seconds and the give answer as) 105621	(Dừng 30 giây sau đó trả lời) 105621

Turing dự đoán đến hết thế kỷ hai mươi, một chiếc máy tính với 10 gigabytes bộ nhớ có khoảng 30% cơ hội lừa một người trong 5 phút hỏi đáp. Có một thực tế

rằng bạn luôn có thể lừa một vài người trong tất cả mọi lần, không rõ tiêu chuẩn này sẽ nghiêm ngặt đến đâu. Mặt khác, vấn đề quan trọng của Turing là việc sử dụng ngôn ngữ thành thạo như con người, là một cách thử trí thông minh.

Rất ngạc nhiên, nhiều vấn đề được đề xuất bởi Turing liên hệ nhiều đến chương trình **ELIZA** xuất hiện vào năm 1966 (Weizenbaum, 1966). ELIZA là hệ thống xử lý ngôn ngữ tự nhiên có khả năng thực hiện một số đoạn hội thoại với người dùng. Cùng xem xét đoạn hội thoại sau đây với một phiên bản của ELIZA

User_1 :	You are like my father in some ways. *Bạn rất giống bố tôi*
ELIZA_1 :	WHAT RESEMBLANCE DO YOU SEE *BẠN THẤY GIỐNG Ở ĐIỂM NÀO*
User_2 :	You are not very aggressive but I think you don't want me to notice that. *Bạn không quá hung hăng nhưng tôi nghĩ bạn không muốn tôi để ý điều đó*
ELIZA_2 :	WHAT MAKES YOU THINK I AM NOT AGGRESSIVE *ĐIỀU GÌ KHIẾN BẠN NGHĨ TÔI KHÔNG HUNG HĂNG*
User_3 :	You don't argue with me *Bạn không cãi nhau với tôi*
ELIZA_3 :	WHY DO YOU THINK I DON'T ARGUE WITH YOU *TẠI SAO BẠN NGHĨ TÔI KHÔNG CÃI NHAU VỚI BẠN*
User_4 :	You are afraid of me. *Bạn sợ tôi*
ELIZE_4 :	DOES IT PLEASE YOU TO BELIEVE I AM AFRAID OF YOU *BẠN NGHĨ RẰNG TÔI SỢ BẠN À?*

Như chúng tôi sẽ thảo luận chi tiết trong chương 2, ELIZA là một chương trình rất đơn giản sử dụng pattern-matching để xử lý đầu vào và chuyển thành đầu ra tương ứng. Sự thành công của kỹ thuật đơn giản trong lĩnh vực này dựa vào việc ELIZA không thực sự cần biết điều gì để giả người. Như trong ghi chép của Weizanbaum, đây là một trong số ít các loại hội thoại trong đó người nghe có thể phản hồi mà không cần biết điều gì về thế giới thực.

ELIZA liên hệ nhiều đến ý tưởng của Turing khi có nhiều người tương tác với ELIZE đã tin nó thực sự hiểu họ và vấn đề của họ. Hơn nữa, Weizenbaum (1976) cho thấy nhiều người tiếp tục tin vào khả năng của ELIZA thậm chí sau khi cách thức hoạt động của nó được giải thích cho họ. Trong những năm sau, thí nghiệm của Weizenbaum được lặp lại trong các môi trường khác nhau. Năm 1991, một sự kiện được biết đến như cuộc thi Loebner Prize được tổ chức, cố gắng đưa nhiều chương trình máy tính vào thí nghiệm Turing. Mặc dù những cuộc thi như vậy không thu hút nhiều giới khoa học, những kết quả qua từng năm cho thấy ngay cả các chương trình đơn giản nhất cũng có lúc có thể lừa các giám khảo (Shieber, 1994a). Không ngạc nhiên, các kết quả đó không thể dập tắt được được cuộc tranh luận về tính thích hợp của phép thử Turing cho việc kiểm tra trí thông minh giữa các nhà triết học và các nhà nghiên cứu trí tuệ nhân tạo (Searle, 1980).

May thay, với mục đích của cuốn sách, các kết quả đó không liên quan đến việc

khả năng của máy tính trong việc trở nên thông minh, hay hiểu ngôn ngữ tự nhiên. Điều quan trọng hơn nữa là các kết quả nghiên cứu trong lĩnh vực xã hội học đã khẳng định một dự báo của Turing trong cùng bài báo đó.

> Tôi tin rằng cuối thế kỷ này, việc sử dụng ngôn và các ý kiến được giáo dục sẽ thay đổi nhiều đến mức chúng ta có thể trò chuyện với các máy tính suy nghĩ mà không cảm thấy bị mâu thuẫn.

Ngày nay, rõ ràng là bất kể mọi người tin hay biết gì về cách thức hoạt động của máy tính, mọi người thường nói về chúng và tương tác với chúng như là những thực thể trong xã hội. Mọi người thường nói chuyện với máy tính nếu họ nghĩ máy tính là người; họ sẽ lịch sự, đối xử như những thành viên, hi vọng rằng máy tính có thể hiểu nhu cầu của họ, và có thể tương tác với họ một cách tự hiên. Ví dụ, Reeves và Nass (1966) phát hiện rằng khi một máy tính hỏi một người đánh giá máy tính như thế nào, người đó đưa ra nhiều phản hồi tính cực khi một máy tính khác hỏi các câu tương tự. Con người thường sợ trở nên bất lịch sự. Trong một thử nghiệm khác, Reeves và Nass phát hiện rằng con người thường cho điểm cao hơn nếu máy tính biết tâng bốc. Từ các thiên hướng này, các hệ thống dựa vào ngôn ngữ và tiếng nói cần đưa cho người dùng những giao diện tự nhiên nhất. Điều này dẫn đến sự tập trung dài hạn vào thiết kế cho các **máy giao tiếp** và các hệ thống thông minh có thể giao tiếp.

1.5. CÁC KẾT QUẢ TỐT NHẤT

> *Chúng ta chỉ nhìn thấy một quãng đường ngắn trước mắt,*
> *nhưng chúng ta có thể thấy rất nhiều việc cần làm.*
>
> Alan Turing.

Hiện tại là thời điểm sôi động lĩnh vực xử lý ngôn ngữ và tiếng nói. Bắt đầu từ việc có nhiều tài nguyên tính toán hơn, việc phát triển của Web đưa đến một lượng lớn thông tin có thể thu thập được, hỗ trợ việc phát triển mọi khía cạnh của các ứng dụng xử lý ngôn ngữ và tiếng nói. Các ví dụ dưới đây đưa ra các hệ thống đã được triển khai trong xu hướng này:

- Các công ty du lịch như Amtrak, United Airlines hay các nhà cung cấp du lịch khác đưa ra các chương trình hội thoại có thể hướng dẫn trong quá trình đặt vé và đưa ra các thông tin về các chuyến đi.
- Các công ty xe cao cấp như Mercedes-Benz đưa ra các hệ thống tổng hợp và nhận diện tiếng nói có thể cho người lái điều kiển các hệ thống môi trường,

giải trí và định vị trong xe bằng giọng nói. Hệ thống thoại tương tự cũng được triển khai cho các phi hành gia ở Interational Space Station.

- Blinkx và các công ty trong lĩnh vực tìm kiếm video khác cung cấp các dịch vụ tìm kiếm cho hàng triệu giờ video trên Web sử dụng công nghệ nhận diện tiếng nói để xác định các từ trong video.
- Google cung cấp các dịch vụ dịch và trích rút thông tin đa ngôn ngữ khi người dùng có thể gõ câu hỏi tìm kiếm bằng ngôn ngữ mẹ đẻ trong các ngôn ngữ khác. Google sẽ tự động dịch câu hỏi, tìm kiếm những trang web thích hợp và sau đó tự động dịch chúng lại về tiếng bản ngữ của người dùng.
- Các tổ chức xuất bản lớn như Pearson, hay các dịch vụ kiểm tra như ETS, sử dụng các hệ thông tự động để phân tích, chấm điểm và đánh giá hàng ngàn bài luận của sinh viên.
- Các hệ thống hướng dẫn tự động, dựa vào các nhân vật hoạt hình, dạy trẻ em đọc tiếng Anh (Wise et al., 2017).
- Các công ty phân tích văn bản như Nielsen, Buzzmetrics, Umbria và Collective Intellect cung cấp các dịch vụ tiếp thị thông minh dựa vào việc tự động đánh giá ý định, sở thích, tính cách và thái độ của người dùng từ các trang web và forum.

1.6. MỘT VÀI LỊCH SỬ NGẮN GỌN

Trong lịch sử, xử lý ngôn ngữ và tiếng nói được nghiên cứu rất khác nhau trong nhiều ngành như khoa học máy tính, kỹ thuật điện, ngôn ngữ học, tâm lý học. Bởi tính đa dạng đó, xử lý ngôn ngữ và tiếng nói bao gồm rất nhiều lĩnh vực riêng biệt từ các ngành khác nhau: **ngôn ngữ học tính toán** trong ngôn ngữ học, **xử lý ngôn ngữ tự nhiên** trong khoa học máy tính, **xử lý tiếng nói** trong kỹ thuật điện, **tâm lý học tính toán** trong tâm lý học. Phần này sẽ trình bày sơ lược quá trình phát triển của các chủ đề nghiên cứu này trong lịch sử, các chủ đề này sẽ được thảo luận chi tiết hơn trong các chương sau.

1.6.1. NỀN MÓNG: 1940 - 1950

Cội nguồn của lĩnh vực này được bắt đầu từ ngay sau chiến tranh thế giới thứ hai cùng với sự phát triển của máy tính. Giai đoạn này từ những năm 1940 đến hết những năm 1950 với rất nhiều công trình nghiên cứu về hai mô hình cơ bản: **automaton** và **xác suất** hay **information-theoretic models**.

Lĩnh vực automaton phát triển trong những năm 1950 xuất phát từ mô hình của Turing (1936) về các giải thuật tính toán, được nhiều nhà nghiên cứu đánh giá là nền tảng của ngành khoa học máy tính hiện đại. Các công trình của Turing dẫn đến mô hình **McCulloch-Pitts neuron** (McCulloch và Pitts, 1943), một mô hình đơn giản của neuron như là một đơn vị tính toán được mô tả dưới dạng logic mệnh đề, và sau đó là các nghiên cứu của Kleene (1951) và (1956) trong finite automata và regular expressions. Shannon (1948) áp dụng mô hình xác suất của các quá trình Markov rời rạc để automata cho ngôn ngữ. Lấy ý tưởng của một quá trình Markov hữu hạn từ công trình của Shannon, Chomsky (1956) đầu tiên coi các máy trạng thái hữu hạn là một cách để biểu diễn ngữ pháp, và định nghĩa các ngôn ngữ trạng thái hữu hạn như một phương thức sinh ngôn ngữ dựa vào ngữ pháp trạng thái hữu hạn. Các mô hình ban đầu này dẫn đến lĩnh vực **lý thuyết ngôn ngữ hình thức**, sử dụng đại số và lý thuyết tập để định nghĩa ngôn ngữ hình thức như là một chuỗi của các biểu tượng. Bao gồm context-free grammar, định nghĩa bởi Chomsky (1956) cho ngôn ngữ tự nhiên, đồng thời được phát hiện độc lập bởi Backus (1959) và Naur et al. (1960) trong phần mô tả của ngôn ngữ lập trình ALGOL.

Điểm đáng chú ý thứ hai trong giai đoạn này là sự phát triển của các mô hình xác suất trong xử lý ngôn ngữ và tiếng nói, dẫn đến một đóng góp khác của Shannon: metaphor của **noisy channel** và **decoding** cho việc truyền tải ngôn ngữ thông qua các kênh giao tiếp và âm thanh tiếng nói. Shannon mượn ý tưởng từ khái niệm **entropy** của nhiệt động lực học như một cách đo kích thước của thông tin qua một kênh, hay nội dung thông tin của một ngôn ngữ, và phát triển một độ đo của entropy của tiếng anh sử dụng kỹ thuật xác suất.

Đó cũng là giai đoạn đầu của phổ âm được phát triển (Koenig et al., 1946), và các nghiên cứu cơ bản được thực hiện trong lĩnh vực âm học tạo nền tảng cho các nghiên cứu sau đó trong lĩnh vực nhận dạng tiếng nói. Dẫn đến hệ thống nhận diện tiếng nói trong những năm đầu 1950. Năm 1952, các nhà nghiên cứu ở Bell Labs xây dựng một hệ thống thống kê có thể nhận diện được 10 chữ số từ một người nói (Davis et al., 1952). Hệ thống có 10 mẫu nhận dạng thể hiện hai nguyên âm đầu của các số. Họ đạt độ chính xác 97-99% bằng cách chọn mẫu có correlation coefficient cao nhất với đầu vào.

1.6.2. HAI TRƯỜNG PHÁI: 1957 - 1970

Cuối những năm 1950, đầu 1960, xử lý ngôn ngữ và tiếng nói chia ra thành hai trường phái riêng rẽ: symbolic và stochastic.

Trường phái symbolic được phát triển từ hai hướng nghiên cứu. Đầu tiên là các công trình của Chomsky và những tác giả khác về lý thuyết ngôn ngữ hình thức và sinh cú pháp trong giai đoạn cuối 1950 và giữa những năm 1960, và các công trình từ rất nhiều nhà ngôn ngữ học, khoa học máy tính trong các thuật toán phân tích,

đầu tiên theo top-down và bottom-up và sau đó thông qua quy hoạch động. Một trong những hệ thống parsing hoàn thiện đầu tiên là Zelig Harris's Transformations and Discourse Analysis Project (TDAP), được cài đặt trong khoảng từ tháng 6 năm 1968 đến tháng 7 năm 1959 tại đại học Pennsylvania (Harris, 1962). Hướng nghiên cứu thứ hai là sự xuất hiện của một ngành mới - trí tuệ nhân tạo. Trong mùa hè 1956, John McCarthy, Marvin Minsky, Claude Shannon và Nathaniel Rochester lập thành một nhóm nghiên cứu trong một workshop hai tháng trong một lĩnh vực mà họ quyết định đặt tên là trí tuệ nhân tạo (AI). Mặc dù AI luôn có những nghiên cứu tập trung vào các thuật toán hỗn độn và thống kê (bao gồm các mô hình xác suất và neural net), vấn đề tập trung trong lĩnh vực mới này là các công trình về suy diễn và logic của Newell và Simon trong Logic Theorist and the General Problem Solver. Tại thời điểm này, các hệ thống hiểu ngôn ngữ tự nhiên đầu tiên được xây dựng. Những hệ thống đơn giản này hoạt động trong một miền đơn bằng cách kết hợp các luật và tìm kiếm theo từ khoá với phương pháp heuristic trong việc suy diễn và trả lời câu hỏi. Cuối những năm 1960, các hệ thống logic hình thức được phát triển.

Trường phái stochastic được phát triển chính trong lĩnh vực thống kê và kỹ thuật điện. Cuối năm 1950, phương pháp Bayesian được áp dụng trong lĩnh vực nhận diện ký tự quang học. Bledsoe và Browning (1959) xây dựng một hệ thống Bayesian cho nhận dạng văn bản sử dụng một từ điển lớn và tính likelihood của mỗi chuỗi ký tự quan sát được từ mỗi từ trong từ điển bằng cách nhân các likelihood của từng ký tự. Mosteller và Wallace (1964) áp dụng các phương pháp Bayesian để giải quyết vấn đề về quyền tác giả của các báo cáo khoa học trong The Federalist.

Những năm 1960 cho thấy sự phát triển của các mô hình tâm lý học kiểm chứng được đầu tiên của xử lý ngôn ngữ tự nhiên dựa vào ngữ pháp chuyển đổi, cũng như các kho ngữ liệu đầu tiên: tập dữ liệu Brown của tiếng Anh Mỹ, một triệu từ được thu thập từ 500 văn bản từ nhiều nguồn (báo chí, tiểu thuyết, giả tưởng, học thuật,...), được thu thập bởi đại học Brown vào năm 1963-64 (Kucera và Francis, 1967; Francis, 1979; Francis và Kucera, 1982), và một bộ từ điển tiếng địa phương Trung Quốc của William S. Y. Wang năm 1967.

1.6.3. BỐN MÔ HÌNH: 1970-1983

Gia đoạn tiếp theo chứng kiến sự bùng nổ trong nghiên cứu về xử lý ngôn ngữ và tiếng nói và sự phát triển của các mô hình vẫn thống trị các hướng nghiên cứu trong ngành.

Mô hình **stochastic** đóng vai trò to lớn trong sự phát triển các giải thuật nhận dạng tiếng nói trong giai đoạn này, đặc biệt là việc sử dụng mô hình Hidden Markov Model và metaphor trong noisy channel và decoding, phát triển độc lập bởi Jelinek, Bahl, Mercer và các đồng nghiệp tại trung tâm nghiên cứu IBM's Thomas J. Wat-

son, Baker tại đại học Carnegie Mellon University, người chịu ảnh hưởng bởi các công trình của Baum và đồng nghiệp tại Institute for Defense Analysis tại Princeton. AT&T's Bell Laboratories cũng đóng vai trò trung tâm trong lĩnh vực xử lý và tổng hợp tiếng nói, đọc Rabiner và Juang (1993) để biết thêm chi tiết về công trình này.

Mô hình **hướng logic** được bắt đầu từ công trình của Colmerauer và đồng nghiệp trong Q-systems và metamorphosis grammars (Colmerauer, 1970, 1975), tiền thân của Prolog và Define Clause Grammars (Pereira và Warren, 1980). Một cách độc lập, các công trình của Kay (1979) về ngữ pháp hàm, Bresnan và Kaplan (1982) về LFG, cho thấy sự quan trọng của cấu trúc đặc trưng thống nhất.

Lĩnh vực **hiểu ngôn ngữ** cũng bắt đầu trong giai đoạn này, với hệ thống SHRDLU của Terry Winograd, giả lập một con robot trong thế giới của các viên gạch đồ chơi (Winograd, 1972a). Chương trình có thể hiểu được các lệnh bằng ngôn ngữ tự nhiên khá phức tạp (đặt viên gạch màu đỏ lên trên viên gạch nhỏ màu xanh). Hệ thống này còn có khả năng xây dựng một tập ngữ pháp với tiếng Anh, dựa vào hệ thống ngữ pháp của Halliday. Mô hình của Winograd cho thấy vấn đề parsing đã được hiểu đầy đủ để bắt đầu tập trung vào các mô hình về ngữ nghĩa và diễn ngôn. Roger Schank cùng các đồng nghiệp và sinh viên xây dựng các chương trình hiểu ngôn ngữ tập trung vào các tri thức về kịch bản, kế hoạch và mục tập và tổ chức bộ nhớ con người (Schank và Albelson, 1977; Schank và Riesbeck, 1981; Cullingford, 1981; Wilensky, 1983; Lehnert, 1977). Các công trình này sử dụng các mô hình ngữ nghĩa dựa mạng lưới (Quillian, 1968; Norman và Rumelhart, 1975; Schank, 1972; Wilks, 1975c, 1975b; Kínch, 1974) và bắt đầu kết hợp với các lưu ý của Fillmore về case roles (Fillmore, 1968) trong biểu diễn của họ (Simmons, 1973).

Các mô hình hướng logic và hiểu ngôn ngữ tự nhiên được kết hợp trong các hệ thống sử dụng logic vị từ như phương pháp biểu diễn ngữ nghĩa, như hệ thống hỏi đáp LUNAR (Woods, 1967, 1973).

Mô hình **diễn ngôn** tập trung vào bốn lĩnh vực trong diễn ngôn. Grosz và các cộng sự giới thiệu các nghiên cứu về kết cấu con trong diễn ngôn và tập trung diễn ngôn (Grosz, 1977a; Sidner, 1983), một số nhà nghiên bắt đầu nghiên cứu về tự động phân giải tham chiếu (Hobbs, 1978), và mô hình **BID** (Belief-Desire-Intention) cho các công trình dựa logic trong xử lý tiếng nói (Perrault và Allen, 1980; Cohen và Perrault, 1979).

1.6.4. CHỦ NGHĨA KINH NGHIỆM VÀ MÔ HÌNH HỮU HẠN TRẠNG THÁI REDUX: 1983-1993

Thập kỷ tiếp theo chứng kiến hai lớp mô hình đã không còn phổ biến từ cuối những năm 1950 hoặc đầu những năm 1960, đặc biệt là do các lập luận lý thuyết chống lại

chúng như những đánh giá của Chomsky về Skinner's Verbal Behavior (Chomsky, 1959b). Mô hình đầu tiên là các mô hình hữu hạn trạng thái, bắt đầu nhận được sự chú ý trở lại từ các công trình về âm vị học và ngôn ngữ hình thái học hữu hạn trạng thái của Kaplan và Kay (1981) và các mô hình hữu hạn trạng thái của Church (1980). Quyển sách này sẽ dành một thời lượng lớn cho các mô hình hữu hạn trạng thái sẽ được trình bày ở các chương sau.

Xu hướng thứ hai trong giai đoạn này là cái mà được gọi là "sự quay lại của chủ nghĩa kinh nghiệm", với điểm nhấn đáng chú ý nhất là sự phát triển của các mô hình xác suất trong xử lý ngôn ngữ và tiếng nói, ảnh hưởng mạnh mẽ bởi các công trình tại IBM Thomas J. Watson Research Center với các mô hình xác suất trong bài toán nhận dạng tiếng nói. Các phương pháp dựa xác suất và các cách tiếp cận hướng dữ liệu lan tỏa từ tiếng nói đến các bài toán gán nhãn từ loại, parsing và ngữ nghĩa. Hướng đi thực dụng này cũng đi cùng bởi một hướng tiếp cận mới trong đánh giá mô hình, dựa vào việc lựa chọn dữ liệu, phát triển các độ đo định tính cho việc đánh giá, và việc chú ý so sánh chất lượng qua các độ đo so với các báo cáo đã xuất bản trước đó.

Giai đoạn này cũng xuất hiện các công trình đáng chú ý trong lĩnh vực sinh ngôn ngữ.

1.6.5. KẾT HỢP CÁC LĨNH VỰC: 1994-1999

Năm năm cuối của thiên niên trải qua sự thay đổi to lớn trong lĩnh vực này. Đầu tiên, các mô hình xác suất và hướng dữ liệu đã trở thành tiêu chuẩn trong việc xử lý ngôn ngữ tự nhiên. Các thuật toán cho parsing, gán nhãn từ loại, phân giải đồng tham chiếu và diễn ngôn tất cả đều được kết hợp với xác suất, và sự phát triển các phương pháp đánh giá lấy ý tưởng từ nhận dạng tiếng nói và truy hồi thông tin. Thứ hai, việc tăng tốc độ và bộ nhớ của máy tính cho phép tiếp cận với các lĩnh vực con của xử lý ngôn ngữ và tiếng nói, như tổng hợp tiếng nói hay kiểm tra lỗi chính tả và ngữ pháp. Các thuật toán xử lý ngôn ngữ và tiếng nói bắt đầu được áp dụng cho Augmentative và Alternative Communication (AAC). Cuối cùng, sự phát triển của Web nhấn mạnh sự cần thiết của các hệ thống truy hồi và trích rút thông tin dựa ngôn ngữ.

1.6.6. SỰ PHÁT TRIỂN CỦA HỌC MÁY: 2000-2007

Xu hướng theo chủ nghĩa kinh nghiệm bắt đầu vào cuối những năm 1990 đã bắt đầu tăng tốc với tốc độ đáng kinh ngạc trong thế kỷ mới. Sự tăng tốc này xuất phát từ ba xu hướng. Đầu tiên, một lượng lớn các dữ liệu tiếng nói và văn bản ra đời với sự hỗ trợ của Linguistic Data Consortium (LDA) và các tổ chức tương tự. Trong các tài nguyên này có các tập dữ liệu được gán nhãn như Penn Treebank (Marcus et al., 1993), Prague Dependency Treebank (Hajic, 1998), PropBank (Palmer et al.,

2005), Penn Discourse Treebank (Miltsakaki et al., 2004b), RSTBank (Carlson et al., 2001) và TimeBank (Pustejovsky et al., 2003b), tất cả các nguồn dữ liệu này đã hình thành các tầng ngữ liệu tiêu chuẩn cho các bài toán ngữ pháp, ngữ nghĩa và ngữ dụng. Sự xuất hiện của các tài nguyên này thúc đẩy xu hướng chuyển các bài toán phức tạp truyền thống, như parsing hay phân tích cảm xúc, trở thành các bài toán học máy có giám sát. Các nguồn tài nguyên này cũng thúc đẩy việc ra đời của các cuộc thi đánh giá cho các bài toán parsing (Dejean và Tjong Kim Sang, 2001), trích rút thông tin (NIST, 2007a; Sang, 2002; Sang và De Meulder, 2003), khử nhập nhằng nghĩa từ (Palmer et al., 2001a; Kilgarriff và Palmer, 2000), hỏi đáp (Voorhees và Tice, 1999), và tóm tắt văn bản (Dang 2006).

Thứ hai, việc tập trung hơn vào quá trình học dẫn đến sự tương tác sâu hơn với cộng đồng học máy thống kê. Các kỹ thuật như support vector machines (Boser et al., 1992; Vapnik, 1995), maximum entropy và các thuật toán tương tự như multinomial logistic regression (Berger et al., 1996), graphical Bayesian models (Pearl, 1988) trở thành tiêu chuẩn trong cộng đồng ngôn ngữ học tính toán. Thứ ba, việc phát triển rộng khắp của các hệ thống máy tính hiệu năng cao giúp cải tiến quá trình huấn luyện và triển khai mà không thể xuất hiện ở thập kỷ trước.

Cuối cùng, ở cuối giai đoạn này, các phương pháp tiếp cận không giám sát cũng nhận được sự quan tâm. Từ các cách tiếp cận thống kê đến dịch máy (Brown et al., 1990; Och và Ney, 2003) và topic modeling (Blei et al., 2003) thể hiện các hệ thống cũng có thể được huấn luyện hiệu quả từ các dữ liệu chưa gán nhãn. Thêm vào đó, giá cả và sự khó khăn trong việc đảm bảo chất lượng cho các ngữ liệu có nhãn là một nhược điểm của các phương pháp học giám sát trong rất nhiều bài toán. Xu hướng sử dụng các kỹ thuật không giám sát sẽ ngày càng tăng.

1.6.7. CÁC PHÁT HIỆN ĐỒNG THỜI

Mặc dù trong phần lược sử này, chúng tôi đã nêu ra nhiều trường hợp về các phát hiện đồng thời cho cùng một ý tưởng. Chỉ một vài trường hợp "đồng thời" này sẽ được thảo luận trong cuốn sách như ứng dụng của quy hoạch động đối với so sánh chuỗi bởi Viterbi, Vintsyuk, Needleman và Wunsch, Sakoe và Chiba, Sankoff, Reichert e al., và Wanger và Fischer (chương 3, 5 và 6), HMM/noisy channel trong nhận dạng tiếng nói bởi Baker vaf cuar Jelinek, Bahl và Mercer (chương 6, 9 và 10); sự phát triển của context-free grammars bởi Chomsky và bởi Backus và Naur (chương 12); chứng minh Swiss-German có một cú pháp không phụ thuộc ngữ cảnh bởi Huybregts và bởi Shieber (chương 15); ứng dụng của sự thống nhất trong xử lý ngôn ngữ bởi Colmerauer et al. và bởi Kay (chương 16)

Những trường hợp này có phải là sự trùng hợp đáng kinh ngạc hay không? Một giả thuyết nổi tiếng bởi nhà xã hội học Robert K. Merton (1961) lập luận, hoàn toàn ngược lại

> Tất cả các khám phá khoa về cơ bản đều được phát hiện đồng thời, kể cả là những khám phá xuất hiện độc lập.

Có những trường hợp rất nổi tiếng về những phát hiện đồng thời; sau đây là một vài ví dụ lấy từ danh sách của Ogburn và Thomas (1922), gồm có sự đồng phát minh ra giải tích của Leibnitz và Newton, sự phát triển đồng thời lý thuyết về chọn lọc tự nhiên của Wallace và Darwin, và sự phát minh ra điện thoại đồng thời bởi Gray và Bell. Nhưng Merton đã chỉ ra những bằng chứng cho giả thiết rằng các phát hiện đồng thời là quy luật chứ không phải trường hợp ngoại lệ, trong đó có những phát minh tưởng chừng do một người, sau đó được chứng minh rằng là sự khám phá lại các kết quả chưa được công bố hay các công trình đã bị lãng quên. Một chứng cứ còn mạnh hơn trong điểm nghiên cứu của ông ấy rằng các nhà khoa học bởi thân họ đã hoạt động dưới giả thiết rằng các phát minh đồng thời là điều bình thường. Do đó, nhiều khía cạnh trong cuộc sống khoa học được thiết kế để giúp các nhà khoa học tránh trùng lập, ngày xuất bản trong các tạp chí, sự lựa chọn thời gian cẩn thận trong các kết quả nghiên cứu, việc điều tra sơ bộ và các báo cáo kỹ thuật.

1.6.8. GHI CHÚ NGẮN CUỐI CÙNG VỀ TÂM LÝ HỌC

Rất nhiều chương trong quyển sách này có phần ghi chú ngắn gọn về những nghiên cứu tâm lý trong việc xử lý ngôn ngữ. Tất nhiên, việc hiểu ngôn ngữ tự nhiên là một mục tiêu khoa học quan trọng và là một phần của lĩnh vực chung là khoa học nhận thức. Tuy nhiên, hiểu về xử lý ngôn ngữ tự nhiên cũng thường xuyên giúp xây dựng các mô hình ngôn ngữ tốt hơn. Điều này dường như trái ngược lại với quan điểm cố hữu, cho rằng việc hiểu sự mô phỏng trực tiếp quy luật của tự nhiên hiếm khi hữu ích trong các ứng dụng kỹ thuật. Ví dụ, một luận điểm thường được đưa ra rằng nếu chúng ta cố gắng sao chép tự nhiên, thì các máy bay sẽ cần vỗ cánh; trong khi các máy bay với các cánh cố định lại là giải pháp kỹ thuật thành công hơn nhiều. Nhưng ngôn ngữ không phải là ngành hàng không. Lấy ý tưởng từ tự nhiên đôi khi hữu ích cho ngành hàng không (dù sao đi nữa thì các máy bay đều có cánh), nhưng nó lại cực kỹ hữu ích khi chúng ta đang xử lý các bài toán liên quan đến con người. Máy bay bay có các mục tiêu khác so với các con chim bay; nhưng mục tiêu của các hệ thống nhận dạng tiếng nói, ví dụ, để thực hiện chính xác những gì mà con người vẫn làm hàng ngày: ghi lại các cuộc nói chuyện. Do con người đã làm việc này rất tốt, chúng ta có thể học từ các giải pháp sẵn có. Do sự quan trọng của các hệ thống xử lý ngôn ngữ và tiếng nói trong việc giao tiếp người máy, việc sao chép các giải pháp mà con người vẫn thường làm cũng là điều hợp lý.

1.7. KẾT LUẬN

Chương này đã giới thiệu lĩnh vực xử lý ngôn ngữ và tiếng nói. Danh sách dưới đây là các điểm đáng chú ý trong chương này.

- Một cách đơn giản để hiểu các chủ đề trong xử lý ngôn ngữ và tiếng nói là xem xét những điều cần thực hiện để tạo ra một robot thông minh như HAL trong phim 2001: A Space Odyssey, hay việc xây dựng một hệ thống hỏi đáp trên nền Web, hay hệ thống dịch máy tự động.

- Các công nghệ tiếng nói và ngôn ngữ dựa trên các mô hình hình thức hay biểu diễn, của tri thức về các tầng của ngôn ngữ như âm vị học và ngữ âm, ngôn ngữ hình thái học, ngữ pháp, ngữ nghĩa, ngữ dụng và diễn ngôn. Các mô hình hình thức như máy trạng thái, hệ thống luật hình thức, logic, mô hình xác suất được sử dụng để biểu diễn các tri thức này.

- Nền tảng của công nghệ ngôn ngữ và tiếng nói dựa vào khoa học máy tính, ngôn ngữ học, toán học, kỹ thuật điện và tâm lý học. Các thuật toán từ các nền tảng tiêu chuẩn này được sử dụng trong các quá trình xử lý ngôn ngữ và tiếng nói.

- Sự liên hệ mật thiết giữa ngôn ngữ và suy nghĩ trong công nghệ xử lý ngôn ngữ và tiếng nói là trung tâm tranh luận về các máy thông minh. Hơn nữa, nghiên cứu về cách con người tương tác với các nội dung số phức tạp chỉ ra rằng công nghệ xử lý ngôn ngữ và tiếng nói sẽ đóng vai trò quan trọng trong sự phát triển các công nghệ tương lai.

- Các ứng dụng tân tiến ứng dụng xử lý ngôn ngữ và tiếng nói đã xuất hiện ở khắp mọi nơi. Việc phát triển của wweb, cũng như những sự cải tiến đáng kể trong nhận dạng và tổng hợp tiếng nói, sẽ dẫn đến nhiều ứng dụng hơn nữa trong tương lai.

1.8. TÀI LIỆU THAM KHẢO

Nghiên cứu trong các lĩnh vực con của xử lý ngôn ngữ và tiếng nói xuất hiện tại khắp các hội nghị và tạp chí lớn. Các hội nghị và tạp chí tập trung vào lĩnh vực xử lý ngôn ngữ và ngôn ngữ học tính toán gán liền với tổ chức ngôn ngữ học tính toán - Association for Computational Linguistics (ACL), hay ở châu Âu (EACL) và hội nghị quốc tế về ngôn ngữ học tính toán - International Conference

on Computational Linguistics (COLING). Các kỷ yếu hội nghị từ ACL, NAACL, EACL và COLING là các diễn đàn chính cho các công trình nghiên cứu trong ngành. Ngoài ra còn có các hội nghị liên quan như ACL Special Interest Groups (SIGs), Conference on Natural Language Processing (CoNLL), cũng như hội nghị Empirical Methods in Natural Language Processing (EMNLP).

Các nghiên cứu trong lĩnh vực tổng hợp và nhận dạng tiếng nói được trình bày hàng năm ở hội nghị INTERSPEECH, International Conference on Spoken Language Processing (ICSLP), European Conference on Speech Communication and Technology (EUROSPEECH), hay hội nghị hàng năm IEEE International Conference on Acoustics, Speech, and Signal Processing (IEEE ICASSP). Các nghiên cứu trong lĩnh vực hội thoại tiếng nói thường được trình bày ở các hội nghị này hay ở các workshops như SIGDial.

Các tạp chí gồm có *Conputational Linguistics*, *Natural Language Engineering*, *Speech Communication*, *Computer Speech and Language*, *IEEE Transactions on Audio*, *Speech Language Processing* và *ACM Transactions on Speech and Language Processing*.

Các công trình trong xử lý ngôn ngữ từ quan điểm trí tuệ nhân tạo có thể tìm thấy trong các hội nghị thường niên như American Association for Artificial Intelligence (AAAI), International Joint Conference on Artifical Intelligence (IJCAI). Các tập chí thường xuyên có các bài nghiên cứu về lĩnh vực xử lý ngôn ngữ và tiếng nói gồm có Machine Learning, Journal of Machine Learning Research, và Journal of Artificial Intelligence Research.

Có rất nhiều quyển sách giáo khoa đề cập đến nhiều khía cạnh của xử lý ngôn ngữ và tiếng nói. Manning và Schutze (1999) (*Foundations of Statistical Language Processing*) tập trung vào các mô hình thống kê trong các bài toán tagging, parsing, khử nhập nhằng, đồng tham chiếu và các bài toán khác. Charniak (1993) (*Statistical Language Learning*) là một nguồn có thể truy cập, đề cập đến các vấn đề cũ hơn và nhẹ nhàng hơn. Manning et al. (2008) tập trung vào bài toán truy hồi thông tin, phân loại văn bản và phân cụm. NLTK, the Natural Language Toolkit (Bird và Loper, 2004), là một tập hợp các module và dữ liệu cho việc xử lý ngôn ngữ, với quyển sách Natural Language Processing dựa trên NLTK. Allen (1995) (Natural Language Understanding) đề cập đến các chủ đề sâu về xử lý ngôn ngữ từ quan điểm AI. Gazdar vaf Mellish (1989) (*Natural Language Processing in Lisp/Prolog*) đề cập đến automata, parsing, các đặc trưng và unification và có sẵn trên mạng. Pereira và Shieber (1987) giới thiệu các chương trình parsing và dịch dựa vào Prolog. Russell và Norvig (2002) là một quyển sách giới thiệu về trí tuệ nhân tạo trong đó có các chương giới thiệu về xử lý ngôn ngữ. Partee et al. (1990) có một phần lớn đề cập đến các kiến thức toán cho ngôn ngữ học tính toán. Một tập hợp danh sách các báo cáo khoa học đáng chú ý có thể tìm được Grosz et al. (1986) (*Readings in Natural Language Processing*).