1 Giới thiệu về Xử lý ngôn ngữ tự nhiên

Dave Bowman: Mở cửa khoang ngủ ra, HAL.
HAL: Xin lỗi Dave, tôi e rằng tôi không thể làm thế được.

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 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 được tiếng Anh, và thậm chí còn đọc 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 đấy 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 wweb viết bằng cách 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 đon giản đến phức tạp:

- "Bất đồng" có nghĩa là gì?
- Abraham Lincoln sinh năm nào?
- Ở Mỹ có bao nhiêu bang?
- 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 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 người. 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 giải quyết được 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 loại kiến 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.