ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

-------📖-------

KIỀU CÔNG CHÍNH

TỐI ƯU BẢNG CỤM TỪ ĐỂ CẢI TIẾN DỊCH MÁY THỐNG KÊ

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Thái Nguyên - 2015

MỤC LỤC

DANH MỤC HÌNH

DANH MỤC BẢNG

DANH MỤC BIỂU ĐỒ

DANH SÁCH CÁC TỪ VIẾT TẮT

Viết tắt	Đầy đủ
PB-SMT	Cụm từ base Statistical Machine Translation
SMT	Statistical Machine Translation
PR-Enc	Cụm từ Rank Encoding

MỞ ĐẦU

Hiện nay trên thế giới có khoảng 5650 ngôn ngữ khác nhau, với một số lượng ngôn ngữ lớn như vậy đã gây ra rất nhiều khó khăn trong việc trao đổi thông tin, trong giao tiếp, đồng thời ngăn cản sự phát triển của thương mại và mậu dịch quốc tế. Mặt khác, với việc bùng nổ Internet như hiện nay, có một khối lượng văn bản khổng lồ trên Internet mà phần lớn là bằng tiếng Anh. Do tính đa dạng của nó mà việc hiểu các văn bản này hoàn toàn không dễ chút nào. Do đó việc có một hệ dịch tự động Anh-Việt là hết sức cần thiết. Với những khó khăn như vậy người ta đã phải dùng đến một đội ngũ phiên dịch khổng lồ, để dịch các văn bản, tài liệu, lời nói từ tiếng nước này sang tiếng nước khác. Những công việc đó mang tính chất thủ công, nặng nhọc trong khi khối lượng văn bản cần dịch ngày càng nhiều. Để khắc phục những nhược điểm trên hiện nay có rất nhiều những hệ thống tự động dịch miễn phí trên mạng như: systran, google translate, vietgle, vdict... Những hệ thống này cho phép dịch tự động các văn bản với một cặp ngôn ngữ chọn trước (ví dụ dịch từ tiếng Anh sang tiếng Việt) [1]. Điều ấy cho thấy sự phát triển của dịch máy càng ngày càng tiến gần hơn đến ngôn ngữ tự nhiên của con người.

Ngay từ khi xuất hiện chiếc máy tính điện tử đầu tiên người ta đã tiến hành nghiên cứu về dịch máy. Công việc đưa ra mô hình tự động cho việc dịch đã và đang được phát triển, mặc dù chưa giải quyết được triệt để lớp ngôn ngữ tự nhiên. Nhưng sự ra đời của chúng đã khẳng định được ích lợi to lớn về mặt chiến lược và kinh tế, đồng thời các vấn đề liên quan đến dịch máy cũng là những chủ đề quan trọng của ngành khoa học máy tính, bởi chúng liên quan đến vấn đề xử lí ngôn ngữ tự nhiên, một trong những vấn đề có ý nghĩa nhất mà trí tuệ nhân tạo có khả năng giải quyết. Người ta tin rằng việc xử lí ngôn ngữ tự nhiên trong đó có dịch máy sẽ là giải pháp cho việc mở rộng cánh cửa đối thoại người-máy, lúc đó con người không phải tiếp xúc với

máy qua những dòng lệnh cứng nhắc nữa mà có thể giao tiếp một cách trực tiếp với máy.

Với sự phát triển mạnh mẽ của dịch máy tự động thì dịch máy thống kê (Statistical Machine Translation) đã chứng tỏ là một hướng tiếp cận đầy tiềm năng bởi ưu điểm vượt trội so với các phương pháp dịch máy dựa trên cú pháp truyền thống. Kết quả thực tế của hệ thống dịch máy thống kê tốt hơn, ngôn ngữ dịch càng ngày càng gần với ngôn ngữ của người, giúp con người trao đổi thông tin dễ dàng hơn, tốc độ nhanh hơn và cùng với nhiều ngôn ngữ hơn.

Hiện nay, phương pháp dịch thống kê dựa trên cụm từ là phương pháp cho kết quả dịch tốt nhất. Để dịch hiệu quả thì bảng cụm từ phải lớn chính vì vậy việc lưu trữ và tìm kiếm trong bảng cụm từ là rất quan trọng. Chính vì thế, luận văn này tôi lựa chọn và thực hiện đề tài "**Tối ưu bảng cụm từ để cải tiến dịch máy thống kê**".

CHƯƠNG I: DỊCH MÁY THỐNG KÊ TRÊN CƠ SỞ CỤM TỪ

Hiện nay dịch máy thông kê dựa trên cơ sở cụm từ là một trong những hướng phát triển đang được rất nhiều người quan tâm. Dịch máy thống kê dựa trên cụm từ nhằm mục đích dịch một văn bản từ ngôn ngữ nguồn sang ngôn ngữ đích dựa vào bảng ngữ cụm từ sau khi thực hiện việc gióng hàng từ, gióng hàng thống kê, đảo cụm từ… kết hợp với mô hình ngôn ngữ.

1.1 Ngôn ngữ tự nhiên

Ngôn ngữ tự nhiên là những ngôn ngữ được con người sử dụng trong các giao tiếp hàng ngày nghe, nói, đọc, viết. Mặc dù con người có thể dễ dàng hiểu và học các ngôn ngữ tự nhiên, việc làm cho máy hiểu được ngôn ngữ tự nhiên không phải là chuyện dễ dàng. Sở dĩ có khó khăn là do ngôn ngữ tự nhiên có các bộ luật, cấu trúc ngữ pháp phong phú hơn nhiều các ngôn ngữ máy tính, hơn nữa để hiểu đúng nội dung các giao tiếp, văn bản trong ngôn ngữ tự nhiên cần phải nắm được ngữ cảnh của nội dung đó.

Do vậy, để có thể xây dựng được một bộ ngữ pháp, từ vựng hoàn chỉnh, chính xác để máy có thể hiểu ngôn ngữ tự nhiên là một việc rất tốn công sức và đòi hỏi người thực hiện phải có hiểu biết sâu về ngôn ngữ học. Do đó cần phải tìm ra một phương pháp dịch tư động tối ưu để làm giảm công sức trong vấn đề về dịch ngôn ngữ nói chung.

1.2 Dịch máy

Dịch tự động hay còn gọi là dịch máy là một trong những ứng dụng quan trọng của xử lý ngôn ngữ tự nhiên, là sự kết hợp của ngôn ngữ, dịch thuật và khoa học máy tính. Như tên gọi dịch tự động là việc thực hiện dịch một ngôn ngữ đầu vào (ngôn ngữ này gọi là ngôn ngữ nguồn) sang một hoặc nhiều ngôn ngữ khác (gọi là ngôn ngữ đích) bằng các công cụ, phần mềm trên máy tính đã được lập trình sẵn mà không cần có sự can thiệp của con người.