Dịch máy và Hiểu văn bản

Lê Khánh Hùng,

hunglekhanh@gmail.com

Từ tháng 10 năm 2009, Ủy ban tư vấn Khoa học Công nghệ của tổng thống Obama đã đưa ra một danh sách các "Thách thức Lớn" của thế kỷ 21 (*Cam kết mới của Tổng thống về khoa học, công nghệ và sáng tạo sẽ cho phép quốc gia đặt ra và đạt được những mục tiêu đầy tham vọng giúp cải thiện chất lượng cuộc sống của chúng ta và thiết lập nền tảng cho các ngành công nghiệp và việc làm của tương lai*).

Trong số 8 thách thức (8 nhiệm vụ KH&CN khó cần giải quyết) được nêu ra có 3 nội dung về CNTT&TT. Một trong 3 nôi dung đó là:

Biên dịch tự động, chính xác cao và thời gian thực giữa các ngôn ngữ quan trọng trên thế giới - hạ thấp đáng kể các rào cản tới giao thương và hợp tác quốc tế.

CHÂU ÂU : từ năm 2010 dành 26 triệu euro cho nội dung "Tương tác bằng Ngôn ngữ" với MỤC TIÊU: ***Những kiến trúc, mô hình và công cụ mới cho dịch máy tự học với hiệu quả cao về chi phí,*** và TÁC ĐỘNG MONG MUỐN:

- Giảm bớt sự khác nhau về chất lượng giữa bản dịch của con người và bản dịch tự động

- Tăng gấp hai lần tốc độ trung bình của người dịch trong vòng tám năm

- Bản dịch tự động hóa dễ tương tác hơn, dễ thích nghi hơn, có khả năng tự học và dễ sử dụng hơn

Từ những kế hoạch của Mỹ và Châu Âu ta có thể rút ra hai kết luận:

- Chất lượng các hệ dịch máy còn xa với mong muốn.
- Cần có và cần hỗ trợ những nghiên cứu mới, độc đáo, không đi theo dòng chính.

Chúng tôi bắt đầu nghiên cứu về dịch máy từ năm 1989. Năm 1990 đã có bản thử nghiệm dịch Anh-Việt (Tuần lễ Tin học, TP HCM).

Sau đó, 1997, 1999 có phần mềm EVTRAN.

Vào thời gian này chúng tôi cảm thấy bế tắc.

- Chất lượng sẽ bão hòa : Việc bổ sung hiệu chỉnh dữ liệu sẽ không giúp tăng chất lượng
- Phương pháp dựa trên luật
    - Hạn chế ở mức cú pháp.
    - Không rõ dịch là gì? Không có, và không thể có định nghĩa DỊCH. Nguyên nhân : Trong mô hình Chomsky (và các mở rộng về sau) không có ***khái niệm tương đương***. Đã không có khái

niệm tương đương thì không có ***biến đổi tương đương***. Vì vậy không có khái niệm ***dịch***.

- Tam giác Vauquois sai : Sau khi phân tích cú pháp không hề có sự dịch chuyển khỏi ngôn ngữ nguồn.
- Việc gắn các giải thuật phân tích ngữ nghĩa vào mỗi quy tắc văn phạm chỉ càng làm sa lầy vào ngôn ngữ nguồn

- Phương pháp kho ngữ liệu là một bước lùi : dịch không cần hiểu.

Từ năm 2001, chúng tôi bắt đầu xây dựng một cách tiếp cận mới đối với phương pháp dựa trên luật : Đưa ngữ nghĩa vào mô hình văn phạm. Đề xuất một dạng mở rộng của văn phạm Chomsky để xử lý ngữ nghĩa. Phát hiện này đặt ra một số vấn nghiên cứu cần giải quyết. Để có điều kiện tập trung vào bài toán và thu hút nhân tài đi theo hướng này, chúng tôi đã mạnh dạn đăng ký một đề tài cấp nhà nước về Dịch máy vào các năm 2003-2004 nhưng không được duyệt. Để tồn tại, chúng tôi tạm thời nâng cấp và đưa thêm chiều dịch Việt – Anh vào phần mềm với công nghệ cũ. Cho đến bây giờ Bản EV-Shuttle đã cho thu nhập khoảng gần 1 tỷ đồng. Nhờ đó từ năm 2006 chúng tôi đã quay lại tiếp tục nghiên cứu và thử nghiệm giải pháp mới.

Các kết quả chính:

- Với việc tích hợp ngữ nghĩa ngay trong mô hình hình thức văn phạm, lần đầu tiên đưa ra khái niệm tương đương.
- Khái niệm dịch được định nghĩa một cách hình thức
- Tách ngữ nghĩa ra khỏi các ngôn ngữ thành một thành phần độc lập.
- Trong mô hình này, ta có thể tính toán được những câu hỏi như:
  - Câu "*A gửi thư cho B*" tương đương với câu "*B nhận thư từ A*"
  - Từ giả thiết *động vật có chân* và *gà thuộc động vật* kết luận "*gà có chân*"
- Từ điển ngữ nghĩa đa ngôn ngữ : sản phẩm phụ vừa phục vụ dịch máy vừa dùng để tra cứu cho con người
- Mở ra thêm một con đường để giải quyết các bài toán XLNNTN khác như tóm tắt văn bản, text mining, tìm kiếm toàn văn theo ngữ nghĩa : chẳng hạn khi tìm "*dịch tự động*" ta không mong muốn nhận được các kết quả như "*giao dịch tự động*", "*truyền dịch tự động*", "*tiết dịch tự động*", "*chuyển dịch tự động*" hay "*miễn dịch tự động*"... và tìm kiếm đa ngôn ngữ.

Lộ trình:

- 2010 : Từ điển ngữ nghĩa Đa ngôn ngữ gồm Việt, Anh, Nhật.
- 2011 : Dịch máy Đa ngôn ngữ Việt, Anh, Nhật.
- Từ 2011 : Bổ sung các ngôn ngữ dân tộc Việt nam và những ngôn ngữ khác.

Kiến nghị:

- Khối lượng dữ liệu ngôn ngữ rất lớn nên cần có đầu tư
- Khuyến khích các nghiên cứu cơ bản trong lĩnh vực.
- Bỏ quan điểm cho rằng chỉ có làm lại cái của nước ngoài cho tiếng Việt thì mới là nghiên cứu chính thống.