



(12) **BẢN MÔ TẢ SÁNG CHẾ THUỘC BẰNG ĐỘC QUYỀN SÁNG CHẾ**

(19) **Cộng hòa xã hội chủ nghĩa Việt nam (VN)**
CỤC SỞ HỮU TRÍ TUỆ

(11) 
1-0019755

(51)⁷ **G06F 17/28**

(13) **B**

(21) 1-2014-02900

(22) 28.08.2014

(45) 25.09.2018 366

(43) 25.03.2016 336

(76) **NGUYỄN DUY THẮNG (VN)**

Thôn Chợ, làng Bình Đà, xã Bình Minh, huyện Thanh Oai, thành phố Hà Nội.

(54) **PHƯƠNG PHÁP DỊCH MÁY KHÔNG LIÊN TỤC**

(57) Sáng chế đề xuất phương pháp dịch máy không liên tục. Với mục đích là đơn giản hóa quá trình dịch và cải thiện chất lượng dịch.

Sáng chế chia quá trình dịch từ ngôn ngữ A sang ngôn ngữ B thành 2 quá trình.

Có thể phát triển độc lập từng quá trình .

Quá trình thứ nhất: Dịch ngôn ngữ A thành các giá trị được lưu trữ theo phương pháp lưu trữ dữ liệu và chuyển đổi ngôn ngữ (dữ liệu trung gian).

Quá trình thứ hai: Chuyển dữ liệu trung gian thành ngôn ngữ bất kỳ B, dữ liệu xuất ra có thể là văn bản, âm thanh, hình ảnh...

Việc tách quá trình dịch thành 2 phần làm giảm độ khó trong việc dịch, độc lập phát triển các ngôn ngữ mới và có thể diễn ra trên các thiết bị khác nhau.

Lĩnh vực kỹ thuật được đề cập

Phương pháp dịch máy không liên tục áp dụng trong kỹ thuật dịch máy (dịch tự động dựa trên CNTT).

Tình trạng kỹ thuật của sáng chế

Trong dịch máy hiện nay có 3 phương pháp là :

- + Dịch máy dựa trên thống kê (phương pháp chủ yếu hiện nay).
- + Dịch máy dựa trên ví dụ.
- + Dịch máy dựa trên luật hay còn gọi là phương pháp tiếp cận dịch thuật trực tiếp.

Việc dịch trong các phương pháp đó là một quá trình liên tục. Nên các phương pháp dịch đó có các hạn chế sau:

- Không thể dịch trực tiếp đa ngôn ngữ. (Theo khái niệm về đẳng hướng. Các chương trình hiện nay giới thiệu khả năng dịch đa ngôn ngữ thực chất là việc thực hiện nhiều lần quá trình dịch song ngữ)
- Không thể diễn ra trên các hệ thống thiết bị khác nhau. (Ví dụ không thể triển khai việc dịch trên một máy chủ và một máy client). Và cũng không thể diễn ra trên hai ứng dụng khác nhau.
- Không thể độc lập phát triển khả năng dịch một ngôn ngữ mới. (Luôn phải tồn tại một cặp song ngữ.)
- Nếu độ khó ngôn ngữ A là x và độ khó ngôn ngữ B là y. Thì quá trình dịch ngôn ngữ A sang ngôn ngữ B sẽ có độ khó là x.y (x nhân y).

Bản chất kỹ thuật của sáng chế :

Mục đích của sáng chế là chuẩn hóa quá trình dịch, nâng cao chất lượng dịch và đơn giản hóa quá trình dịch. Nhằm đạt mục đích trên sáng chế đề xuất phương pháp dịch 2 giai đoạn. Sáng chế sử dụng dữ liệu của phương pháp lưu trữ dữ liệu và chuyển đổi ngôn ngữ để nối kết 2 quá trình đó . Sau đây ta gọi là dữ liệu trung gian. Việc dịch ngôn ngữ A sang ngôn ngữ B được thực hiện như sau :

(2 giai đoạn không cần thực hiện trong cùng 1 hệ thống, hay trong cùng 1 ứng dụng)

Giai đoạn 1 : dịch ngôn ngữ A thành dữ liệu trung gian.

Giai đoạn 2 : chuyển dữ liệu trung gian thành ngôn ngữ B.

Dữ liệu trung gian này đã được chuẩn hóa nên có thể tiến hành chuyển đổi sang các ngôn ngữ khác tương tự như với ngôn ngữ B.

Mô tả vắn tắt hình vẽ

Hình vẽ 1 trong tài liệu mô tả 2 giai đoạn của việc dịch.

Mô tả chi tiết sáng chế :

Dữ liệu trung gian được sử dụng trong sáng chế được đặt tên là DLSC (Đây là kiểu dữ liệu được dùng trong phương pháp lưu trữ dữ liệu và chuyển đổi ngôn ngữ). Mỗi giá trị DLSC có thể tương ứng với 1 từ, 1 cụm từ hoặc thậm chí là cả 1 câu hoàn chỉnh của ngôn ngữ tự nhiên. DLSC có 2 kiểu giá trị và có độ dài là 4 byte. (4 byte = 32 bit = 2^{32} giá trị).

- Dữ liệu cơ sở (kiểu 1): được chia thành 2 phần là Nội dung (21 bit) + Ngữ pháp (11 bit)

Nội dung được chia thành 2 phần nhỏ là Từ loại (5bit) và Giá trị(16 bit).

Ngữ pháp có 3 phần nhỏ là Ngữ pháp chung, Đồng nghĩa và Mở rộng.

Ngữ pháp chung : lưu trữ các thông tin ngữ pháp của các đơn vị ngôn ngữ mang tính chất chung nhất của tất cả các ngôn ngữ.

Đồng nghĩa : Dùng để phân biệt các từ đồng nghĩa. Đây chính là cách hội tụ tất cả các từ đồng nghĩa vào một thể duy nhất. Số lượng các từ đồng nghĩa được hỗ trợ có thể thay đổi tùy vào sự thay đổi của Từ loại và Mở rộng.

Mở rộng: Là phần dùng bổ sung các yếu tố ngữ pháp cho từng ngôn ngữ cụ thể.

- Dữ liệu mở rộng (kiểu 2). chia thành 2 phần: Nội dung (21 bit)+ Thông tin bổ sung (11 bit).

Nội dung được chia thành 2 phần nhỏ là Từ loại (5bit) và Giá trị(16 bit).

Thông tin bổ sung : Dùng để dự trữ. Có thể dùng để hỗ trợ dịch

Thành phần Nội dung trong cả 2 kiểu có giá trị như nhau và tương đương với 1 đơn vị của ngôn ngữ tự nhiên. (Với các ngôn ngữ tự nhiên khác nhau thì giá trị của phần này sẽ không thay đổi. Nó có tác dụng như là cầu nối giữa các ngôn ngữ, và là cầu nối giữa 2 kiểu dữ liệu của DLSC) Dữ liệu mở rộng (kiểu 2) có thể có thể mở rộng thành 64 bit, 128 bit hoặc nhiều hơn (các biến thể). Vì trong quá trình dịch cần phải lưu trữ nhiều thông tin. Tuy nhiên phần 21 bit đầu của vùng dữ liệu này phải giống với 21 bit đầu của dữ liệu cơ sở (kiểu 1).

Phân bố thành phần trong miền giá trị của 4 byte (vị trí các vùng biến có thể thay đổi)

Từ loại lưu trữ từ bit 1 đến bit 5 có 32 giá trị. (Từ loại có tác động đến Giá trị và Ngữ pháp)

Giá trị lưu trữ từ bit 6 đến bit 21 có 65536 giá trị.

Ngữ pháp lưu trữ từ bit 22 đến bit 32 có 2^{11} giá trị. (Nội dung = Từ loại + Giá trị = 2^{21})

Nếu Từ loại có giá trị 0 vùng Giá trị chính là bảng mã Unicode. (Có thể kết hợp với phần Ngữ pháp để tạo ra các bảng mã lớn hơn Unicode, xác định ngôn ngữ được lưu trữ xuất phát từ ngôn ngữ tự nhiên nào...)

Nếu Từ loại có giá trị 1 thì mỗi giá trị của vùng Giá trị tương ứng với 1 phó từ.

Nếu Từ loại có giá trị 2 thì mỗi giá trị của vùng Giá trị tương ứng với 1 tính từ.

Nếu Từ loại có giá trị 3,4,5 thì mỗi giá trị của vùng Giá trị tương ứng với 1 danh từ chỉ động vật.

Nếu Từ loại có giá trị 8,9 thì mỗi giá trị của vùng Giá trị tương ứng với 1 danh từ chỉ thực vật.

Nếu Từ loại có giá trị 12,13 thì mỗi giá trị của vùng Giá trị tương ứng với 1 danh từ chỉ đồ vật.

Nếu Từ loại có giá trị trong miền 16,17 thì mỗi giá trị của vùng Giá trị tương ứng với 1 danh từ chỉ sự việc, hiện tượng...

Nếu Từ loại có giá trị 20 thì mỗi giá trị của vùng Giá trị tương ứng với 1 động từ.

Nếu Từ loại có giá trị 21 thì mỗi giá trị của vùng Giá trị có thể tương ứng với 1 liên từ, giới từ, đại từ, thán từ, mạo từ. (mỗi vùng từ loại có độ lớn là 500 giá trị)

Nếu Từ loại có giá trị 22 thì mỗi giá trị của vùng Giá trị tương ứng với 1 thành ngữ.

Nếu Từ loại có giá trị 23,24 thì mỗi giá trị của vùng Giá trị tương ứng với 1 câu.

Nếu Từ loại có giá trị 6,7,10,11,14,15,18,19 và 25 đến 31 để dành. Vùng Giá trị cũng để dành.

Giá trị của Từ loại cũng ảnh hưởng đến các thành phần của Ngũ pháp.

Nếu Từ loại có giá trị là 1,2 vùng Ngũ pháp sẽ có 3 phần nhỏ là :

Ngũ pháp(2^{11} giá trị) = Ngũ pháp chung(2^3 giá trị) (gồm 3 phần nhỏ Npc.hơn(2^1 giá trị) + Npc.bằng(2^1 giá trị) + Npc.kém(2^1 giá trị)) + Đồng nghĩa(2^5 giá trị) + Mở rộng(2^3 giá trị).

Phần Ngũ pháp chung hỗ trợ xác định các kiểu so sánh (so sánh nhất, so sánh hơn, so sánh bằng, nguyên thể, so sánh kém, so sánh kém nhất).

Nếu Từ loại có giá trị 3,4,5,8,9,12,13,16,17 vùng Ngũ pháp sẽ có 3 phần nhỏ là :

Ngũ pháp(2^{11} giá trị) = Ngũ pháp chung(2^3 giá trị) (gồm 2 phần nhỏ Npc.số (2^1 giá trị) + Npc.giống(2^2 giá trị)) + Đồng nghĩa(2^5 giá trị) + Mở rộng(2^3 giá trị).

Phần Ngũ pháp chung hỗ trợ xác định số, giống của danh từ (số ít, số nhiều, giống đực, giống cái, giống trung, nguyên thể). Việc xác định cách trong một số ngôn ngữ như tiếng Nga thì được xác định thêm ở phần Mở rộng.cách(2^3 giá trị).

Tiếng Anh thì phân biệt danh từ đếm được nên Mở rộng có dạng Mở rộng.đếm 2^1 .

Nếu Từ loại có giá trị là 20. Ngũ pháp sẽ có 3 phần nhỏ là :

Ngũ pháp 2^{11} = Ngũ pháp chung 2^2 (Npc.thời 2^2) + Đồng nghĩa 2^5 + Mở rộng 2^3 .

Phần Ngũ pháp chung hỗ trợ xác định thời của động từ (quá khứ, hiện tại, tương lai, nguyên thể). Với một ngôn ngữ cụ thể thì số lượng từ Đồng nghĩa và Mở rộng sẽ thay đổi. Ví dụ với tiếng Việt thì là Đồng nghĩa 2^5 + Mở rộng 2^3 nhưng nếu là tiếng Anh thì 2 giá trị đó đổi thành Đồng nghĩa 2^2 + Mở rộng 2^7 vì Mở rộng 2^7 = Mở rộng. 2^4 + Mở rộng.Ngôi số 2^3 . Trong tiếng Việt động từ chia không phụ thuộc vào ngôi và số và chỉ có 3 thời đơn giản nên không cần có thêm phần Mở rộng để bổ xung các yếu tố ngữ pháp. Nhưng tiếng Anh thì có phân biệt và cách chia thì phức tạp hơn nên phần Mở rộng phải hỗ trợ để chia hết các thời và phân biệt ngôi và số của chủ ngữ mà động từ phụ thuộc.

Nếu Từ loại có giá trị trong miền từ 21 đến 24 vùng Ngũ pháp có 2 phần nhỏ là :

Đồng nghĩa (2^6 giá trị) và Mở rộng(2^5 giá trị).

Hai giá trị đầu tiên của mảng dữ liệu chứa các giá trị được sử dụng trong sáng chế là giá trị xác định ngôn ngữ nguồn tạo thành dữ liệu đó. 4 byte có giá trị =0 (đánh dấu), 4 byte thứ 2 dùng để xác định ngôn ngữ (được xác lập theo giá trị của mã vùng quốc gia. Việt nam =84.)

Chuyển đổi văn bản sang dạng DLSC. Xác định ngôn ngữ dùng để lưu trữ. Sử dụng các thuật toán tìm kiếm hiện tại (như rẽ nhánh, nút ..) để thực hiện quá trình chuyển đổi (xác định từ vựng, từ loại, ngữ pháp chung và các thông tin thêm). Các ký tự không phải là từ vựng được ghi trong các biến như là một từ vựng bình thường

Chuyển đổi giá trị DLSC thành dạng văn bản hoặc sang các dạng khác như âm thanh, hình ảnh.... Vì DLSC là giá trị và có cấu trúc lên ta không khó để thực hiện việc chuyển đổi này. Chỉ có 1 vài điểm chú ý trong quá trình chuyển đổi. Thành phần ngữ pháp thay đổi qua các ngôn ngữ lên DLSC có thể có các giá trị khác nhau. Nó phụ thuộc vào ngôn ngữ tự nhiên nào đã tạo ra các giá trị DLSC đó. Có 2 trường hợp xảy ra trong quá trình xử lý dữ liệu: (đọc 8 byte đầu tiên của mảng dữ liệu để xác định).

Nếu giá trị DLSC được tạo ra từ ngôn ngữ đúng với ngôn ngữ đang xử lý (trường hợp ngôn ngữ A ánh xạ vào DLSC. và giá trị DLSC ánh xạ lại ngôn ngữ A.) thì các thành phần Đồng nghĩa, Mở rộng được sử dụng toàn bộ để quá trình có thể khôi phục 1 cách chính xác nhất. (Khôi phục giống hoàn toàn)

Nếu giá trị DLSC được tạo ra từ ngôn ngữ khác với ngôn ngữ đang xử lý (trường hợp ngôn ngữ B ánh xạ vào DLSC và giá trị DLSC ánh xạ sang ngôn ngữ A-có tác dụng như là quá trình dịch từ ngôn ngữ B sang ngôn ngữ A). Thì các thành phần Đồng nghĩa, Mở rộng không được sử dụng. 2 giá trị đó sẽ được thay thế bằng các giá trị của ngôn ngữ đang xử lý (ngôn ngữ A). (Khôi phục không hoàn toàn). Ví dụ như giá trị DLSC được tạo ra từ “từ soya”. Giá trị DLSC được xử lý trong tiếng Việt sẽ là

“soya” được lưu trữ dưới dạng Từloại =3, Giá trị=3; Npc.số=0; Npc.giống=0; Đồngnghĩa=0; Npc.đếm=0; => Các giá trị sẽ được sử dụng là : Từloại =3; Giá trị=3; Npc.số=0; Các giá trị Npc.giống, Đồngnghĩa được gán giá trị mặc bằng 0, điều này giúp ta xác định được từ “đậu tương”. Ta có thể sử dụng biến thăm dò để quét hết các giá Đồngnghĩa. Ví dụ nếu ta gán Đồngnghĩa=1 thì từ mà ta sẽ thu được là “đỗ tương”...(Nếu giá trị Đồngnghĩa nhập vào lớn hơn số lượng từ đồng nghĩa có thật thì nó lại trả về từ đầu tiên)

Ví dụ sử dụng cấu trúc DLSC:

Lưu trữ danh từ đậu tương, đỗ tương, đậu nành trong tiếng Việt (3 từ đồng nghĩa)

Từ “đậu tương” được lưu trữ với Từloại =Giá trị=3, Npc.số= Npc.giống=0, Đồngnghĩa=0.

Từ “đỗ tương” được lưu trữ với Từloại =Giá trị=3, Npc.số= Npc.giống=0, Đồngnghĩa=1.

Từ “đậu nành” được lưu trữ với Từloại =Giá trị=3, Npc.số= Npc.giống=0, Đồngnghĩa=2.

tương ứng là soya và soya bean trong tiếng Anh (2 từ đồng nghĩa). Giá trị các thành phần là:

Từ “soya” được lưu với Từloại =Giá trị=3, Npc.số= Npc.giống=0, Đồngnghĩa=0; Npc.đếm=0.

Từ “soya bean” lưu với Từloại =Giá trị=3, Npc.số= Npc.giống=0, Đồngnghĩa=1; Npc.đếm=0.

Như vậy để xác định loài đậu tương trong 2 ngôn ngữ Việt-Anh (hay bất kỳ ngôn ngữ nào). Ta chỉ cần xem xét 2 vùng Từloại=Giá trị=3. Giá trị của vùng Đồngnghĩa chỉ quan trọng với từng ngôn ngữ cụ thể. Triệt tiêu được từ đồng nghĩa.

Đặc điểm của dữ liệu trung gian là khả năng lưu trữ đồng thời từ loại và từ vựng. (không có khả năng lưu trữ vị trí các từ loại trong câu) Nó còn có khả năng triệt tiêu được các từ đồng nghĩa. Mỗi ngôn ngữ thì có một cách sắp xếp các thành phần từ loại trong câu khác nhau, sáng chế quy định một cấu trúc câu duy nhất khi lưu trữ dữ liệu để các ngôn ngữ có thể dễ dàng

chuyển đổi (lấy dữ liệu ra hoặc đưa dữ liệu vào). Chủ ngữ + vị ngữ , vị ngữ = động từ + bổ ngữ , từ bỏ nghĩa đứng sau từ được bỏ nghĩa.

Phương pháp dịch máy được chia thành 2 giai đoạn. Có thể phát triển độc lập từng giai đoạn. Có thể sử dụng từng giai đoạn vào các mục đích khác nhau.

Giai đoạn thứ nhất : Là quá trình dịch ngôn ngữ tự nhiên (A) thành các giá trị được lưu trữ theo phương pháp lưu trữ và chuyển đổi ngôn ngữ. Vì giá trị lưu trữ dưới dạng DLSC tương ứng với một đơn vị ngôn ngữ (từ , cụm từ...) nên quá trình chuyển đổi ngôn ngữ A sang giá trị DLSC cũng dựa trên nền tảng tìm theo đơn vị ngôn ngữ. Có thể sử dụng các thuật toán hiện nay để thực hiện.

Giai đoạn thứ hai : Là quá trình chuyển đổi các giá trị được tạo ra từ quá trình thứ nhất thành các dạng khác nhau của ngôn ngữ tự nhiên bất kỳ. Dữ liệu sử dụng trong sáng chế là những giá trị có cấu trúc lên ngoài việc có khả năng chuyển đổi thành dạng văn bản thì nó có thể chuyển đổi thành các dạng phi văn bản. Các thông tin từ vựng , ngữ pháp đã được lưu trữ dưới dạng các giá trị và vị trí của các giá trị nên ta có thể truyền các giá trị này cho các ứng dụng khác, thiết bị khác để thực hiện việc giai đoạn thứ 2 này.

Với cả hai quá trình ta cần phải chú ý đến vị trí của các giá trị DLSC được sắp xếp theo quy định để xử lý được chính xác.

Bằng việc chia tách thành 2 giai đoạn dịch thì độ khó của việc dịch đã được chuyển từ phép nhân của độ khó ngôn ngữ thành phép cộng của độ khó ngôn ngữ. Việc triệt tiêu các từ đồng nghĩa và có 1 trật tự ngữ pháp cố định cũng làm việc dịch trở lên dễ dàng hơn.

Khi muốn dịch ngôn ngữ A ra ngôn ngữ X ta chỉ cần tập trung vào giai đoạn 1 của quá trình dịch. Khi muốn dịch ngôn ngữ Y thành ngôn ngữ B ta chỉ cần tập trung vào giai đoạn 2 của quá trình dịch. Việc phát triển một ngôn ngữ mới dễ dàng và độc lập hơn.

Hiệu quả đạt được :

Việc sử dụng dữ liệu của phương pháp lưu trữ dữ liệu và chuyển đổi ngôn ngữ đã giúp cho sáng chế có khả năng chia tách quá trình dịch thành 2 phần độc lập, các từ đồng nghĩa được triệt tiêu, các thông tin về từ vựng, ngữ pháp được lưu trữ một cách đầy đủ . Vì quá trình dịch được chia tách thành 2 phần lên sáng chế đạt được những hiệu quả sau:

Hỗ trợ dịch đa ngôn ngữ 1 cách đồng thời không phải thông qua các bước dịch trung gian, chất lượng và thời gian được cải thiện.

Có thể phát triển việc dịch cho một ngôn ngữ mới một cách độc lập .

Giảm được độ khó của quá trình dịch giữa các ngôn ngữ.

Có thể tiến hành việc dịch trên các thiết bị hoặc các ứng dụng khác nhau.

Yêu cầu bảo hộ

1. Phương pháp dịch máy không liên tục, với mục đích giúp cho việc dịch máy đơn giản và đạt chất lượng tốt hơn, khác biệt với các phương pháp thông thường ở điểm: quá trình dịch ngôn ngữ A sang ngôn ngữ B được chia thành 2 phần độc lập, từng phần của quá trình dịch có thể triển khai trên các thiết bị khác nhau, có thể triển khai trên các hệ thống khác nhau, có thể triển khai trên các ứng dụng khác nhau trong cùng hệ thống, phương pháp này gồm 2 phần :

a nhận dữ liệu ngôn ngữ cần dịch A, tiến hành việc chuyển đổi ngôn ngữ A thành các giá trị trung gian, các giá trị này lưu trữ đầy đủ các thông tin của ngôn ngữ A, mọi ngôn ngữ tự nhiên bất kỳ đều có thể ánh xạ 1:1 vào các giá trị trung gian này,

b nhận dữ liệu trung gian, tiến hành quá trình chuyển đổi các giá trị trung gian thành ngôn ngữ B, các từ đồng nghĩa và đa nghĩa của ngôn ngữ A đã bị triệt tiêu trong các giá trị trung gian, trật tự các thành phần câu được xác định trước nên việc chuyển đổi chỉ là việc ánh xạ tới các vùng dữ liệu tương ứng, trong giai đoạn này có 1 trường hợp đặc biệt xảy ra khi ngôn ngữ B chính là ngôn ngữ A,

trong đó giá trị trung gian được dùng để lưu trữ tất cả các thông tin cần thiết của 1 ngôn ngữ, nó được chia làm 5 phần, phần lưu trữ từ loại (phần 1), phần lưu trữ từ vựng (phần 2), phần lưu trữ đặc điểm ngữ pháp chung (phần 3), phần lưu trữ từ đồng nghĩa (phần 4), phần lưu trữ các thông tin bổ sung (phần 5), phần 1 và phần 2 và phần 3 có giá trị giống nhau cho tất cả các ngôn ngữ khi chỉ tới 1 đối tượng cụ thể, giá trị ở phần 4 và phần 5 phụ thuộc vào giá trị ở phần 1 và ngôn ngữ nguồn, các từ vựng được sắp xếp theo từng nhóm nên khi kết hợp với giá trị của phần 1 ta có thể lưu trữ được nhiều thông tin khác mà không tốn thêm dữ liệu, nếu ngôn ngữ A và B là khác nhau thì không cần xử lý dữ liệu phần 4 và phần 5 tức là các từ đồng nghĩa của ngôn ngữ A bị triệt tiêu, các từ đa nghĩa trong 1 ngôn ngữ cụ thể sẽ có các giá trị khác nhau do cách lưu trữ từ vựng theo đối tượng, vị trí các thành phần của 1 câu lưu trữ trong sáng chế được sắp xếp theo 1 trật tự nhất định và không phụ thuộc bởi ngôn ngữ A, hai giá trị đầu tiên của mảng dữ liệu chứa các giá trị trung gian là giá trị xác định ngôn ngữ nguồn tạo thành dữ liệu đó.

Hình vẽ 1

