CÁCH TIẾP CẬN THỐNG KÊ CHO HỆ DỊCH TỰ ĐỘNG VIỆT – ANH

Đinh Điền, Nguyễn Thống Nhất, Nguyễn Thái Ngọc Duy
Khoa Công Nghệ Thông Tin, trường ĐH Khoa Học Tự Nhiên – ĐHQG-HCM
(Bài nhận ngày 14 tháng 11 năm 2002)

TÓM TẮT: *Lịch sử dịch máy đã có một cuộc cách mạng về phương pháp luận. Đó là: thay vì phải xây dựng các từ điển, các luật chuyển đổi bằng tay bởi các chuyên gia ngôn ngữ như trong các hệ dịch thông thường khác, thì nay đã có một hệ dịch chỉ dựa trên việc thống kê từ kho ngữ liệu song ngữ để xây dựng nên các từ điển và các luật chuyển ngữ một cách hoàn toàn tự động. Máy tính sẽ thống kê và rút ra xác suất dịch tương ứng về từ / ngữ hay cấu trúc giữa 2 ngôn ngữ cũng như xác suất chuyển dịch vị trí giữa 2 ngôn ngữ, và xác suất xuất hiện của từ/ngữ đó trong một ngữ cảnh nhất định nào đó. Đây chính là cách tiếp cận của hệ dịch máy kiểu thống kê (Statistical Machine Translation)[i].*

Trong bài báo này, chúng tôi sẽ trình bày cách tiếp cận thống kê nói trên cho hệ dịch tự động Việt-Anh, một hệ dịch mà rất khó thực hiện bằng phương pháp luật chuyển đổi thông thường khác vì nó đòi hỏi phải phân tích hình thái, cú pháp và ngữ nghĩa tiếng Việt.

Chúng tôi đã cài đặt thử nghiệm hệ dịch Việt-Anh bằng cách tiếp cận thống kê nói trên, trên kho ngữ liệu song ngữ Anh-Việt 5.000.000 từ và bước đầu đạt được kết quả khả quan.

1. Giới thiệu:

Hiện nay, trên thế giới có rất nhiều cách tiếp cận dịch máy khác nhau, trong đó cách tiếp cận thống kê đang ngày càng tỏ ra hiệu quả hơn. Nó không cần quan tâm nhiều đến lý thuyết phân tích về ngôn ngữ, nó chỉ dựa trên sự xác suất dịch tương ứng của các yếu tố ngôn ngữ được rút ra từ kho ngữ liệu song ngữ dùng để huấn luyện. Kho ngữ liệu song ngữ (tiếng Anh là bilingual corpus, parallel text, hay bitext) này chính là tập hợp một số rất lớn các câu rút từ thực tế và đã được dịch chính xác bởi người sang một ngôn ngữ thứ hai.

Trong bài báo này, chúng tôi sẽ giới thiệu về hệ dịch Việt-Anh dựa trên cách tiếp cận thống kê. Phần ngữ liệu song ngữ dùng để huấn luyện ở đây chính là kho ngữ liệu song ngữ Anh-Việt 5.000.000 từ chuyên về lĩnh vực khoa học tự nhiên và khoa học kỹ thuật và các câu thông thường. Kho ngữ liệu song ngữ này đã được thu thập và xử lý qua công trình [ii], vì vậy trong bài báo này chúng tôi sẽ không đề cập đến các chi tiết đó. Trong phần 2, chúng tôi giới thiệu khái quát về dịch máy thống kê, phần 3 và 4: giới thiệu về mô hình ngôn ngữ và mô hình dịch. Phần 5: mô hình tìm kiếm. Hai phần cuối chúng tôi sẽ đánh giá kết quả mà hệ thống của chúng tôi đạt được và cuối cùng là kết luận.

2. Lý thuyết dịch thống kê:

Chúng ta hãy hình dung một máy tính có thể dịch một câu tiếng Việt v sang một câu tiếng Anh e như cách mô tả sau [iii]:

$$\exists e \in E: \ \arg\max_e P(e|v) \qquad (2.1)$$

với e là câu tiếng Anh, v là câu tiếng Việt, E là tập hợp tất cả các câu có thể có trong tiếng Anh và $P(e|v)$ là xác suất có điều kiện của câu e khi đã cho một câu v. Ký hiệu $\arg\max_e$ có nghĩa là làm cho $P(e|v)$ lớn nhất.

Đặc tả (2.1) có thể được giải thích như sau: giả sử có một người phiên dịch từ tiếng Việt sang tiếng Anh, khi nhận được một câu v từ một người Việt thì người đó sẽ nhẩm trong đầu những câu e có thể để dịch và phải tìm ra một câu e tốt nhất để người bản địa Anh có thể hiểu được.

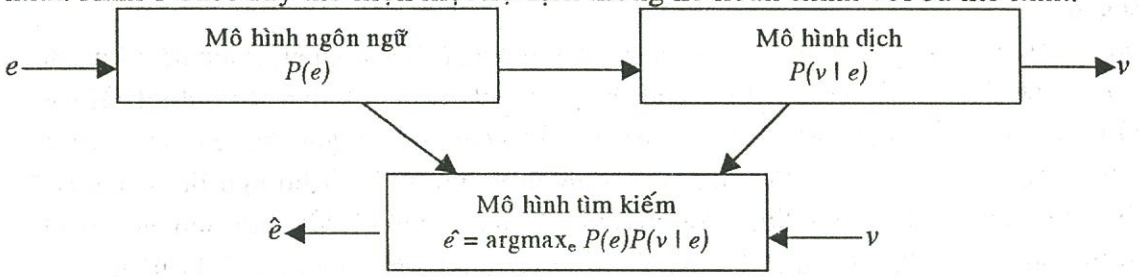
Đối với con người thì chúng ta thấy vấn đề này không khó, vì chúng ta có thể giới hạn lại các câu e trong một tập hợp nhỏ mà trong đầu chúng ta nhẩm được, nhưng ngược lại đối với máy tính thì không vì nó phải duyệt qua toàn bộ tất cả tập hợp các câu e có thể trong thế giới thực, mà điều này thì không thể thực hiện được do số lượng các câu trong thế giới thực là vô cùng lớn. Nhưng thật may mắn, chúng ta có thể giải quyết vấn đề này thông qua việc sử dụng định lý Bayes:

$$P(e \mid v) = \frac{P(e) * P(v \mid e)}{P(v)} \qquad (2.2)$$

Vì mẫu thức trong phương trình (2.2) độc lập với e nên chúng ta có thể tìm một câu tiếng Anh $\hat{e}$ gần đúng nhất bằng cách làm tích $P(e) * P(v \mid e)$ là lớn nhất có thể. Như vậy, chúng tôi có thể viết phương trình căn bản nhất trong dịch máy thống kê như sau:

$$\hat{e} = \arg\max_{e} P(e) * P(v \mid e) \qquad (2.3)$$

Qua phương trình (2.3) thì chúng ta thấy công việc trở nên rất đơn giản và nó tổng kết ba vấn đề cho việc xây dựng một hệ thống dịch máy thống kê hoàn chỉnh, đó là: vấn đề ước lượng xác suất mô hình ngôn ngữ $P(e)$, vấn đề ước lượng xác suất mô hình dịch $P(v \mid e)$; và vấn đề cuối cùng là tìm kiếm một câu tiếng Anh e tốt nhất sao cho tích $P(e) * P(v \mid e)$ là lớn nhất. Hình 1 dưới đây thể hiện một hệ dịch thống kê hoàn chỉnh với ba mô hình.



Hình 1. Hệ dịch thống kê hoàn chỉnh. v là câu tiếng Việt sẽ cần dịch sang tiếng Anh, e là câu tiếng Anh giả định sẽ được dịch, và $\hat{e}$ là câu tiếng Anh được dịch tốt nhất.

3. Mô hình ngôn ngữ:

Ta xem e là một chuỗi của những từ tiếng Anh $e_1 e_2 ... e_l$. Mô hình ngôn ngữ $P(e)$ sẽ tính xác suất mà e sẽ xuất hiện đúng với ngữ pháp tiếng Anh. Nhưng làm sao chúng ta có thể kiểm tra được e có đúng ngữ pháp hay không? Vấn đề này hơi khó để cho máy tính xử lý. Trong phần này chúng tôi sẽ đưa ra cách giải quyết gần đúng cho vấn đề này bằng cách học từ ngữ liệu tiếng Anh (ngữ liệu có ngữ pháp chuẩn) để rút ra các thông số thống kê cần thiết.

Theo công thức nhân của xác suất, chúng tôi viết lại như sau:

$$P(e) = P(e_1 e_2 ... e_l) = P(e_1) * P(e_2 \mid e_1) * P(e_3 \mid e_1 e_2) * ... * P(e_l \mid e_1 e_2 ... e_{l-1}) \qquad (3.1)$$

Nhiệm vụ của mô hình ngôn ngữ là phải ước lượng được mỗi số hạng trong l số hạng ở phía bên phải của phương trình (3.1).

Nếu chúng ta giả sử rằng $|\varepsilon|$ là kích cỡ của tất cả những từ vựng trong từ điển tiếng Anh, thì số trường hợp khác nhau từng đôi một của $e_1 e_2 ... e_{k-1}$ ở xác suất điều kiện thứ k sẽ là $|\varepsilon|^{k-1}$. Đây là một con số rất lớn và do đó chúng ta không thể tính một cách chính xác cho một số

hạng $P(e_k \mid e_1 e_2 ... e_{k-1})$ được. Vì lý do đó, chúng tôi sử dụng mô hình trigram để tính xấp xỉ cho từng số hạng trong (3.1) theo cách sau:

$$P(e_k \mid e_1 e_2 ... e_{k-1}) \approx P(e_k \mid e_{k-2} e_{k-1}) \qquad (3.2)$$

Có nghĩa là thay vì xét k-1 từ vựng $e_1 e_2 ... e_{k-1}$ đã xảy ra ở phía trước từ vựng e_k thì chúng tôi giới hạn lại chỉ xét 2 từ vựng ở phía trước đã xảy ra thôi. Mỗi bộ ba $(e_{k-2} e_{k-1} e_k)$ được gọi là trigram. Như vậy, từ (3.1) chúng ta có thể viết lại như sau:

$$P(e) = P(e_1 e_2 ... e_l) = P(e_1) * P(e_2 \mid e_1) \prod_{i=3}^{l} P(e_i \mid e_{i-2} e_{i-1}) \qquad (3.3)$$

Chúng ta thấy rằng có thể có $|\varepsilon|^3$ trigram khác nhau, và có những trường hợp trigram không bao giờ xuất hiện trong ngữ liệu huấn luyện của chúng ta. Lúc đó sẽ tồn tại một $P(e_k \mid e_{k-2} e_{k-1})$ bằng không, và điều này sẽ dẫn đến P(e) cũng bằng không. Vì lý do đó, chúng tôi mô tả $P(e_k \mid e_{k-2} e_{k-1})$ bằng cách tính gần đúng như sau:

$$P(e_k \mid e_{k-2} e_{k-1}) = \theta_1 * T(e_k \mid e_{k-2} e_{k-1}) + \theta_2 * B(e_k \mid e_{k-1}) + \theta_3 * U(e_k) + \theta_4 \qquad (3.4)$$

θ_1, θ_2, θ_3, θ_4: là các hệ số (hằng số) với $\theta_1 << \theta_2 << \theta_3 << \theta_4$

T: xác suất trigram (ba từ liên tiếp)

B: xác suất bigram (hai từ liên tiếp)

U: xác suất unigram (duy nhất một từ) và bằng $1/|\varepsilon|$

Phương trình (3.4) được gọi là mô hình trigram đã được làm trơn (smoothed trigram model). Theo cách tính này thì chúng ta không cần quan tâm đến ngữ nghĩa và cú pháp giữa những từ với nhau. Để có ước lượng $P(e)$ một cách chính xác hơn khi chúng tôi phải kết hợp với mô hình ngữ pháp liên kết [4](link grammars model). Trong bài báo này chúng tôi sẽ không đề cập đến mô hình ngữ pháp liên kết, chúng tôi hy vọng sẽ nói kỹ về mô hình này ở các bài báo sau. Đây là một mô hình huấn luyện dùng để tính xác suất ngữ pháp trong từng trigram một mà nó sẽ mô tả tất cả thông tin của trigram đó.

4. Mô hình dịch:

Trong phần này chúng tôi sẽ tóm tắt những thành phần của mô hình dịch $P(v \mid e)$ (xin tham khảo thêm chi tiết trong [1]).

Đầu tiên chúng tôi cần quan tâm đến việc liên kết từ giữa một cặp câu (v, e). Vì giữa một cặp câu (v, e) thì có rất nhiều cách liên kết từ khác nhau, do đó để đánh giá độ chính xác của từng liên kết thì chúng tôi phải quan tâm đến xác suất liên kết $P(A = a \mid V = v, E = e)$ gồm một bộ ba biến ngẫu nhiên (V, A, E). Trong đó, V là chuỗi tiếng Việt ngẫu nhiên, E là chuỗi tiếng Anh ngẫu nhiên và A là sự liên kết từ ngẫu nhiên giữa chúng. Theo công thức nhân xác suất chúng tôi có thể viết như sau:

$$P(a \mid v, e) = \frac{P(v, a \mid e)}{P(v \mid e)} \qquad (4.1)$$

Theo phương trình (4.1) thì để tính được $P(a \mid v, e)$ thì ta phải tính được $P(v, a \mid e)$ và $P(v \mid e)$. Đối với $P(v \mid e)$ có thể được tính như sau: $P(v \mid e) = \sum_a P(v, a \mid e)$ (4.2)

với $a \in \mathcal{A}(v, e)$ – $\mathcal{A}$ là tập hợp của các liên kết có thể có giữa e và v.

Đối với $P(v, a \mid e)$ được tính như sau:

$$P(v, a \mid e) = P(m \mid e) \prod_{j=1}^{m} P(a_j \mid a_1^{j-1}, v_1^{j-1}, m, e) * P(v_j \mid a_1^{j}, v_1^{j-1}, m, e) \qquad (4.3)$$

với $e = e_1^j \equiv e_1 e_2 ... e_j$ là chuỗi tiếng Anh có j từ; $v = v_1^j \equiv v_1 v_2 ... v_j$ là chuỗi tiếng Việt có j từ; $a = a_1^j \equiv a_1 a_2 ... a_j : a_j \in [0, l]$, nếu từ tiếng Việt tại vị trí thứ j được kết nối với một từ tiếng Anh ở vị trí thứ i thì khi đó a_j = i, và nếu từ tiếng Việt không kết nối với từ tiếng Anh nào thì $a_j = 0$; $P(m \mid e)$ là xác suất chuỗi v sẽ có m từ khi đã có chuỗi e.

Để chọn được một liên kết từ tốt nhất thì chúng tôi phải chạy các mô hình dịch (từ 1 đến 5) thông qua một số các vòng lặp của thuật toán ước lượng giá trị cực đại (Estimation – Maximization Algorithm) cho từng thông số của mỗi mô hình. Thuật toán này được mô tả chi tiết trong [1].

4.1. Mô hình 1 (Tính xác suất dịch từ):

Đây là mô hình đơn giản nhất của chúng tôi, mô hình dùng để tính toán xác suất dịch của từng từ riêng biệt. Thông số của mô hình này là xác suất dịch từ $t(v_j \mid e_i)$. Mỗi thông số này được gán giá trị ban đầu là $1/|V|$, $|V|$ là tập từ vựng của tiếng Việt. Qua một số vòng lặp trong mô hình sẽ tìm ra được xác suất tương ứng.

$$P(v, a \mid e) = \frac{\varepsilon}{(l+1)^m} \prod_{j=1}^{m} t(v_j \mid e_i) \qquad (4.1.1)$$

với $\varepsilon \equiv P(m \mid e)$ là một hằng số rất nhỏ so với 1; $t(v_j \mid e_i)$ là xác suất của từ tiếng Việt v_j được dịch sang từ tiếng Anh e_i tại vị trí thứ i trong câu e.

4.2. Mô hình 2 (Tính xác suất liên kết từ cục bộ):

Để làm cho mô hình của chúng tôi thực tế hơn, chúng tôi giới thiệu một thông số liên kết từ a_j cho mỗi vị trí j của v_j. a_j là vị trí trong e tương ứng với vị trí của từ thứ j trong v đã được liên kết. Chú ý rằng những từ tiếng Việt mà tự phát sinh (không phát sinh từ những từ tiếng Anh) thì chúng tôi qui ước rằng chúng được liên kết với những *từ rỗng* (*null word*) ở vị trí 0 trong e. Trong mô hình này chúng tôi sẽ tính các thông số liên kết từ cục bộ a cho từng cặp câu trong ngữ liệu song ngữ.

$$P(v \mid e) = \varepsilon \sum_{a_1=0}^{l} ... \sum_{a_m=0}^{l} \prod_{j=1}^{m} t(v_j \mid e_{a_j}) a(a_j \mid j, m, l) \qquad (4.2.1)$$

với ràng buộc là $\sum_{i=0}^{l} a(i \mid j, m, l) = 1$ (4.2.2)

và $a(a_j \mid j, m, l) \equiv P(v_j \mid a_1^j, v_1^{j-1}, m, e)$ (4.2.3)

4.3. Mô hình 3 (Tính xác suất cho giá trị sản sinh):

Một từ tiếng Anh đơn có thể được dịch ra 0, hoặc 1 hoặc nhiều từ tiếng Việt tương ứng. Ví dụ như từ "*child*" có thể được dịch sang tiếng Việt là "*đứa bé*", và khi đó ta gán cho từ "child" một giá trị là 2, và giá trị đó được gọi là *giá trị sản sinh* (*fertility*). Trong mô hình này chúng tôi sẽ tính xác suất cho *giá trị sản sinh* của từng từ tiếng Anh.

Trước tiên đề cập đến mô hình này thì chúng tôi cần định nghĩa các thông số sau:

ϕ_i: là *giá trị sản sinh* của một từ tiếng Anh tại vị trí thứ i, khi $i = 0$ thì ϕ_0 chính là *giá trị sản sinh* của từ tiếng Anh *rỗng* $e_0 = NULL$; $n(\phi_i \mid e_i)$: là xác suất của từ tiếng Anh tại vị trí thứ i trong chuỗi e có thể sản sinh ra ϕ_i từ tiếng Việt; p_1: là xác suất các từ trong chuỗi tiếng Việt được phát sinh từ $e_0 = NULL$; p_0: là xác suất các từ trong chuỗi tiếng Việt được phát

sinh từ $e_i, i \neq 0$; $d(j \mid a_j, l, m)$: là xác suất liên kết từ bởi một từ tiếng Việt (chiều dài là m) tại vị trí thứ j với một từ tiếng Anh tại vị trí a_j.

$$P(v,a \mid e) = n_0(\phi_0 \mid \sum_{i=1}^{l} \phi_i) \prod_{i=1}^{l} n(\phi_i \mid e_i) \phi_i! \prod_{j=1}^{m} t(v_j \mid e_{a_j}) \prod_{j:a_j \neq 0} d(j \mid a_j, l, m) \qquad (4.3.1)$$

với $n_0(\phi_0 \mid \sum_{i=1}^{l} \phi_i) = \binom{m'}{\phi_0} p_0^{m'-\phi_0} p_1^{\phi_0}$: là xác suất phát sinh ra ϕ_0 từ tiếng Việt sau khi đã phát sinh được m' từ bởi những từ tiếng Anh khác rỗng.

4.4. Mô hình 4 (Tính xác suất liên kết dựa trên tập hợp):

Thông thường một chuỗi e cấu tạo bằng những *cụm từ* (*phrase*) như là những đơn vị được dịch sang những cụm từ trong v. Một cụm từ được dịch có thể xuất hiện tại vị trí trong chuỗi v khác với vị trí của cụm từ tương ứng xuất hiện trong chuỗi e. Trong mô hình 3 chỉ quan tâm đến xác suất sản sinh của từng từ tiếng Anh chứ không quan tâm đến vị trí phân phối của từng từ trong những cụm từ của v sẽ được phát sinh. Trong mô hình 4 này, chúng tôi thay đổi cách xử lý của chúng tôi bằng cách thêm $P(\Pi_{ik} = j \mid \pi_{i1}^{k-1}, \pi_{i1}^{i-1}, \tau_0^l, \phi_0^l, \mathrm{e})$ là thông số phân bổ mỗi từ trong mỗi cụm từ ở vị trí j. Những từ mà được kết nối tới từ rỗng trong e thường không là những dạng cụm từ và vì thế chúng tôi giả sử rằng những từ này được trải ra đồng nhất liên tục trên chuỗi v.

Cho $[i] > 0$, ta có:

$$P(\Pi_{[i]k} = j \mid \pi_1^{[i]-1}, \tau_0^l, \phi_0^l, \mathrm{e}) = d_1(j - \Theta_{i-1} \mid A(e_{[i-1]}), B(v_j)) \qquad (4.4.1)$$

Ở đây, A và B là những hàm của từ tiếng Anh và tiếng Việt mà đảm nhận một con số nhỏ của những giá trị khác nhau như là những thông số qua những từ vựng riêng biệt của chúng. Brown et al. mô tả một thuật toán cho việc chia từ vựng vào trong 50 lớp tương ứng với thuật toán này. Chúng tôi gọi $j - \Theta_{i-1}$ là sự dịch chuyển của từ đóng vai trò chính là j. Nó có thể cả âm lẫn dương. Theo kết quả mà chúng tôi thu được thì thấy rằng $d_1(-1 \mid A(e), B(v))$ lớn hơn $d_1(+1 \mid A(e), B(v))$ khi e là một tính từ và v là một danh từ.

Giả định rằng chúng tôi muốn đặt từ thứ k của tại vị trí i, với $[i] > 0$ và $k > 1$ ta có:

$$P(\Pi_{[i]k} = j \mid \pi_{[i]1}^{k-1}, \pi_1^{[i]-1}, \tau_0^l, \phi_0^l, \mathrm{e}) = d_{>1}(j - \pi_{[i]k-1} \mid B(v_j)) \qquad (4.4.2)$$

4.5. Mô hình 5 (Tính xác suất liên kết đầy đủ):

Cả hai mô hình 3 và 4 đều lã mô hình không đầy đủ. Trong cả hai mô hình này không giải quyết cho vấn đề đặt vị trí cho các từ tiếng Việt được phát sinh bởi từ rỗng trong chuỗi tiếng Anh. Như đã đề cập trong mô hình 4 thì những từ mà được kết nối tới từ rỗng trong e thường không là những dạng cụm từ và được giả sử rằng những từ này được trải ra đồng nhất liên tục trên chuỗi v. Trong mô hình 5, sẽ giải quyết vấn đề tính xác suất đặt vị trí cho các từ tiếng Việt được phát sinh bởi *từ rỗng* trong *e*.

Sau khi chúng tôi đã đặt những từ cho $\tau_1^{[i]-1}$ và $\tau_{[i]1}^{k-1}$ là những vị trí còn trống trong chuỗi tiếng Việt. Rõ ràng $\tau_{[i]k}$ nên được đặt ở một trong những vị trí trống này. Đặt $v(j, \tau_1^{[i]-1}, \tau_{[i]1}^{k-1})$ là một con số của những chỗ trống gia tăng và kể cả vị trí j chỉ đặt trước khi chúng tôi đặt $\tau_{[i]k}$. Chúng tôi viết ngắn gọn lại là v_j. Chúng tôi giữ lại hai thông số trong mô hình 4 là d_1 và $d_{>1}$. Cho $[i] > 0$, ta có:

$$P(\Pi_{[i]k} = j \mid \pi_1^{[i]-1}, \tau_0^l, \phi_0^l, \mathrm{e}) = d_1(v_j \mid B(v_j), v_{\Theta_{i-1}}, v_m - \phi_{[i]} + 1)(1 - \delta(v_j, v_{j-1})) \qquad (4.5.1)$$

Trong thông số cuối cùng của d_1, v_m là số chỗ trống còn lại trong chuỗi tiếng Việt. $\tau_{[i]1}$ có thể được đặt trong vị trí bất kỳ nhưng phải là $\phi_{[i]}-1$ của những chỗ trống còn lại. Bởi vì $\tau_{[i]1}$ phải xuất hiện ở vị trí trái nhất của bất kỳ những từ bởi $T_{[i]}$. Như với mô hình 4, chúng tôi cho phép d_1 phụ thuộc vào từ trước và v_j.

Cho $[i]>0$ và $k>1$, ta có:

$$P(\Pi_{[i]k}=j \mid \pi_{[i]1}^{k-1}, \pi_1^{[i]-1}, \tau_0^l, \phi_0^l, \mathrm{e}) = d_{>1}(v_j - v_{[i]k-1} \mid B(v_j), v_m - v_{[i]k-1} - \phi_{[i]} + k)(1-\delta(v_j, v_{j-1})) \quad (4.5.2)$$

5. Mô hình tìm kiếm:

Ý tưởng của mô hình tìm kiếm rất đơn giản. Như trong hình 1, chúng ta cũng dễ nhận thấy rằng sau khi dữ liệu song ngữ qua mô hình ngôn ngữ $P(e)$ và mô hình dịch $P(v \mid e)$ thì chúng ta sẽ thu được các bảng xác suất cho từng thông số tương ứng. Công việc của mô hình tìm kiếm là tìm tích số của $P(e)P(v \mid e)$ là lớn nhất có thể, dựa trên dữ liệu thống kê thu được từ mô hình ngôn ngữ $P(e)$ và mô hình dịch $P(v \mid e)$.

Đến nay, trên thế giới có hai thuật giải và một thuật toán tối ưu cho mô hình tìm kiếm[5]: thuật giải tìm kiếm tham lam (greedy decoding) (Selman et al., 1992; Monasson et al., 1999), thuật giải tìm kiếm dựa trên ngăn xếp (Stack-Based decoding) (Brown et al., 1995; Wang and Waibel, 1997[6]), và thuật toán tìm kiếm dựa trên việc tìm kiếm chu trình *Hamilton* tối ưu trong bài toán tìm chi phí tối thiểu của người bán hàng (Travelling salesman problem – Garey and Johnson, 1979)[**Error! Bookmark not defined.**]. Đối với hai thuật giải thì có thời gian xử lý nhanh hơn thuật toán nhưng bù lại thì kết quả thấp hơn thuật toán.

6. Đánh giá kết quả:

Chúng tôi đã thử nghiệm trên 150 câu tiếng Việt được dịch sang tiếng Anh và ghi lại kết quả thu nhận như bảng 1.

Với kết quả thu nhận được trong bảng 1, chúng tôi nhận thấy rằng khi chiều dài tăng lên, thì tỉ lệ các câu bị lỗi cũng tăng lên.

Bảng 1. Kết quả sau khi chạy thử nghiệm 150 câu tiếng Việt được dịch sang tiếng Anh, với mô hình ngôn ngữ được sử dụng dựa trên mô hình bigram.

Chiều dài câu	*Thời gian trung bình (giây/câu)*	*Số câu đúng*	*Số câu sai có thể chấp nhận được*	*Số câu sai không chấp nhận được*
6	50	30 (20%)	45 (30%)	75 (50%)
8	123	27 (18%)	40 (26.7%)	83 (55.3%)

7. Kết luận:

Trong các phần trên, chúng tôi đã trình bày một cách tổng thể việc ứng dụng cách tiếp cận dịch máy thống kê cho dịch tự động Việt-Anh. Tuy kết quả chưa cao, nhưng đối với tiếng Việt của chúng ta hiện nay, cách tiếp cận dựa trên thống kê này là một trong những con đường khả thi để thực hiện chương trình dịch tự động Việt sang Anh. Vì cách tiếp cận này có ưu điểm là không đòi hỏi phân tích hình thái, cú pháp và ngữ nghĩa tiếng Việt (đây là những điều rất khó thực hiện hoàn thiện ở thời điểm hiện nay cho tiếng Việt). Để nâng cao chất lượng dịch của hệ dịch thống kê, chúng tôi dự kiến sẽ kết hợp với một cách tiếp cận khác (như: luật, cơ sở tri thức, ...). Ngoài ra, chúng tôi sẽ nâng cao số lượng và chất lượng của kho ngữ liệu song ngữ Anh-Việt (ngữ liệu càng nhiều và càng chính xác thì chất lượng của hệ dịch sẽ càng cao).

Cuối cùng, hệ dịch kiểu thống kê nói trên còn cho ra một "sản phẩm phụ" rất quí, đó là kết quả *liên kết từ* (*word alignment*) trong kho ngữ liệu song ngữ Anh-Việt. Kết quả liên kết này sẽ làm tiền đề cho các tất cả các ứng dụng khác dựa trên song ngữ (như: so sánh điểm tương đồng và dị biệt trong ngành ngôn ngữ học so sánh, chuyển đổi trật tự từ, xây dựng từ điển song ngữ cho dịch máy, ...).

A STATISTICAL APPROACH TO VIETNAMESE-ENGLISH MACHINE TRANSLATION

Dinh Dien, Nguyen Thong Nhat, Nguyen Thai Ngoc Duy

ABSTRACT: The history of Machine Translation (MT) has experienced a revolution in its methodology. That is: instead of constructing dictionaries, manual transfer rules by linguists in conventional MT systems, a new MT approach has constructed these dictionaries and transfer rules automatically by calculating probabilities on bilingual corpora. This new MT approach will statistically extract probabilistic figures of corresponding translation of words/phrases or structures as well as transpositions between two languages and appearance probabilistic of those words/phrases in a certain context. This new MT approach is the Statistical MT.

In this paper, we will present the above-mentioned statistical MT approach on the Vietnamese-to-English MT system, which isn't attainable by other conventional approaches due to their requirements of morphological, syntactical and semantic analysis for Vietnamese.

We have trial run this Vietnamese-English MT system on 5,000,000-word English-Vietnamese bilingual corpus with initial encouraging results.

TÀI LIỆU THAM KHẢO

[1] Brown, Peter, Stephen Delta-Pietra, Vincent Dell-Pietra, and Robert Mercer. *The mathematics of statistical machine translation: Parameter estimation.* Computational Linguistics, 19(2): 263-311 (1993).

[2] Đinh Điền, *Bước đầu xây dựng và xử lý ngữ liệu song ngữ Anh-Việt điện tử*, Luận văn thạc sĩ ngôn ngữ học so sánh, ĐH Khoa học Xã hội & Nhân văn – ĐHQG TPHCM, tháng 1-2002.

[3] Brown, Cocke, Delta-Pietra, et. al., *A Statistical Approach to Machine Translation*, Proceedings of 33th - ACL. (1995)

[4] Lafferty, John, Daniel Sleator, Davy Temperly. *Grammatical trigrams: a probabilistics model of link grammar.* Proceedings of the 1992 AAAI Fall Symposium on Probabilistic Approaches to Natural Language (1992).

[5] Ulrich Germann, Michael Jahr, Kevin Knight, Daniel Marcu, and Kenji Yamada. *Fast Decoding and Optimal Decoding for Machine Translation.* In 39th Annual Meeting of the Association for Computational Liguistics. ACL. (2001).

[6] Y. Wang and A. Waibel. *Decoding algorithm in statistical machine translation.* In *Proc. ACL.* (1997).