TẠP CHÍ KHOA HỌC ĐẠI HỌC ĐÀ LẠT Tập 8, Số 2, 2018 3–12

DỊCH TỰ ĐỘNG VIỆT-K'HO SỬ DỤNG PHƯƠNG PHÁP DỰA VÀO THỐNG KÊ

Nguyễn Minh Hiệp[a], Nguyễn Thị Lương[a], Lê Văn Phượng[b], Nguyễn Thị Minh Huyền[b], Đinh Viết Tuấn[a*]

[a]Khoa Công nghệ Thông tin, Trường Đại học Đà Lạt, Lâm Đồng, Việt Nam
[b]Khoa Toán - Cơ - Tin học, Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Hà Nội, Hà Nội, Việt Nam
[*]Tác giả liên hệ: Email: tuandv@dlu.edu.vn

Lịch sử bài báo
Nhận ngày 11 tháng 01 năm 2018
Chỉnh sửa ngày 25 tháng 03 năm 2018 | Chấp nhận đăng ngày 14 tháng 04 năm 2018

Tóm tắt

Một ứng dụng dịch tự động (Machine Translation - MT) từ tiếng Việt sang tiếng dân tộc K'Ho được trình bày. Ứng dụng nhằm mục đích giới thiệu phương pháp dịch tự động dựa vào thống kê (Statistics Machine Translation - STMT). Do tiếng Việt và tiếng dân tộc K'Ho cùng ngữ hệ Nam Á, nhưng lại thuộc nhóm ngôn ngữ khác nhau, nên phần chuyển ngữ thường được xử lý bằng cách sử dụng từ vựng, cụm từ và câu, thay vì bằng quy tắc cú pháp tổng quát. Các nguyên tắc thiết kế của ứng dụng được mô tả chi tiết, cùng với giao diện của hệ thống. Một số kết quả dịch tự động cũng được trình bày để minh họa cho khả năng ứng dụng phương pháp STMT.

Từ khóa: Dịch máy; Dịch tự động dựa vào thống kê; Dịch tự động; MT; STMT.

Mã số định danh bài báo: http://tckh.dlu.edu.vn/index.php/tckhdhdl/article/view/398

Loại bài báo: Bài báo nghiên cứu gốc có bình duyệt

AN APPLICATION TO TRANSLATE FROM VIETNAMESE INTO K'HO USING STMT APPROACH

Nguyen Minh Hiep[a], Nguyen Thi Luong[a], Le Van Phuong[b], Nguyen Thi Minh Huyen[b], Dinh Viet Tuan[a*]

[a]The Faculty of Information Technology, Dalat University, Lamdong, Vietnam
[b]The Faculty of Mathematics - Mechanics - Informatics, VNU University of Science, Hanoi, Vietnam
[*]Corresponding author: Email: tuandv@dlu.edu.vn

Article history
Received: January 11th, 2018
Received in revised form: March 25th, 2018 | Accepted: April 14th, 2018

Abstract

This paper describes the Statistics Machine Translation (STMT) application to translate from Vietnamese into K'Ho. Both Vietnamese and K'Ho are in the same South Asian language family but they belong to different language groups, so the vocabulary, phrases, and sentences are used for language translation instead of the method based on general syntactic rules. The design principles of the application are described in detail, along with the system interface. Several machine translation results are also presented to illustrate the applicability of the STMT method.

Keywords: Machine Translation; MT; Statistics Machine Translation; STMT.

Article identifier: http://tckh.dlu.edu.vn/index.php/tckhdhdl/article/view/398

Article type: (peer-reviewed) Full-length research article

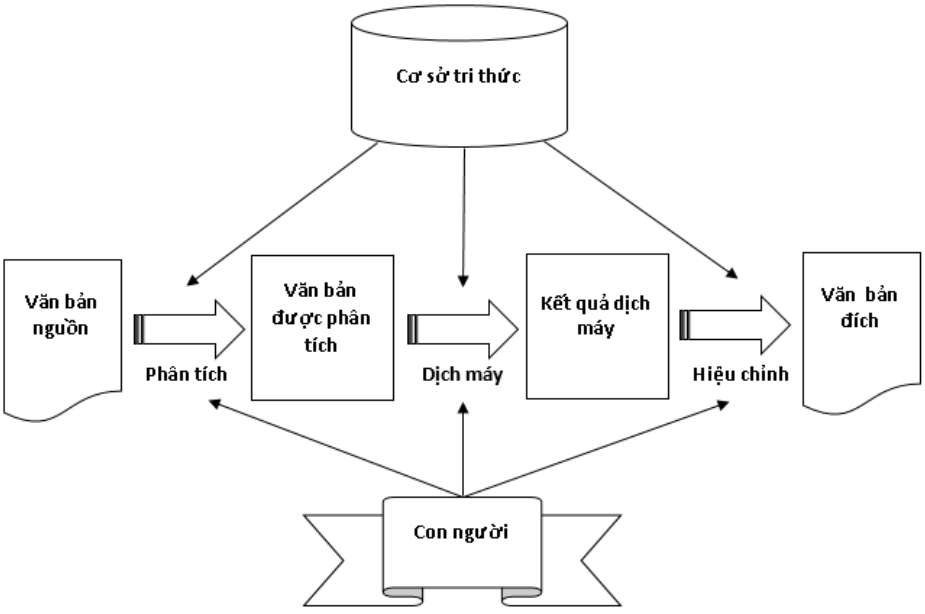
1. GIỚI THIỆU

Hiện nay, trên thế giới có khoảng 5650 ngôn ngữ khác nhau (Đào, 2007). Số lượng ngôn ngữ lớn như vậy đã gây ra rất nhiều khó khăn trong việc trao đổi thông tin. Để có thể trao đổi thông tin phải cần đến một đội ngũ phiên dịch khổng lồ để dịch các văn bản, tài liệu, lời nói từ tiếng này sang tiếng khác. Vì vậy, con người đã nghĩ đến việc thiết kế một hệ thống tự động trong việc dịch.

Hiện nay, khái niệm dịch tự động (dịch máy) đã được nhiều tác giả trong lĩnh vực xử lý ngôn ngữ tự nhiên định nghĩa, tuy có một vài điểm khác biệt nhưng hầu hết đều tương đương với định nghĩa của Pushpak (2006, tr. 1) như sau:

> Dịch máy hay dịch tự động bằng máy tính là tiến trình dịch từ một ngôn ngữ nguồn (ngôn ngữ tự nhiên) sang những ngôn ngữ đích, có hoặc không có sự trợ giúp của con người. Dịch máy thường được thiết kế hoặc cho một cặp ngôn ngữ đặc biệt hoặc cho nhiều hơn hai ngôn ngữ.

Theo Trần (2006) thì quá trình xử lý tài liệu của dịch tự động được mô tả như Hình 1. Đầu vào của một hệ dịch tự động thường là một văn bản được viết bằng ngôn ngữ nguồn và quá trình dịch được chia thành hai giai đoạn: i) Văn bản được phân tích thành các thành phần và ii) Được dịch thành văn bản ở dạng ngôn ngữ đích. Kết quả dịch có thể được con người hiệu chỉnh để trở thành bản dịch tốt hơn.



Hình 1. Quá trình xử lý tài liệu của dịch tự động

Hiện nay, dịch tự động vẫn còn nhiều khó khăn trong việc xử lý các nhập nhằng về ngôn ngữ trong quá trình dịch tự động. Các phương pháp thường dùng trong dịch tự động bao gồm:

- Dịch dựa trên cụm từ (*Phrase Based Machine Translation - PBMT*) (Đào, 2007) là phương pháp xác định nghĩa của câu đích chỉ được thực hiện bởi sự ghép từ và hoán đổi vị trí của từ theo cấu trúc cú pháp của cụm từ. Do thiếu thông tin ngữ cảnh khi xác định xác suất của các từ, nên nghĩa của từ được chọn nhiều lúc không đúng với ngữ cảnh. Đôi khi, nghĩa một từ của ngôn ngữ đích không đủ để diễn tả nghĩa của một từ trong ngôn ngữ nguồn và ngược lại;

- Dịch dựa trên luật (*Rule Based Machine Translation - RBMT*) là phương pháp dựa trên luật cú pháp, ngữ nghĩa và một từ điển khá đầy đủ thông tin. Câu được dịch thường không đạt độ chính xác như mong đợi do lỗi mâu thuẫn giữa các luật hoặc do tập luật không bao quát (Satoshi & Makoto, 1990);

- Dịch tự động dựa trên ví dụ (*Example-Based Machine Translation - EBMT*) được Antal và Peter (2009) tổng kết là cách tiếp cận không đòi hỏi phải có sự phân tích ngôn ngữ học về cú pháp, ngữ nghĩa vì mọi câu dịch đều dựa vào việc "so khớp" mẫu. Việc "so khớp" mẫu dựa hoàn toàn vào kho ngữ liệu song ngữ để xác định mẫu nào gần đúng nhất, sau đó hiệu chỉnh và xuất ra thành phần dịch tương ứng của mẫu đó;

- Dịch tự động dựa trên thống kê (*Statistics Machine Translation - STMT*) là một phương pháp mà các bản dịch được tạo trên cơ sở các mô hình thống kê có các tham số được bắt nguồn từ việc phân tích các cặp câu song ngữ. Ý tưởng dịch tự động bằng thống kê mang tính thuần túy về toán học, cách tiếp cận này không đòi hỏi sự phân tích sâu về ngôn ngữ, quá trình dịch được thực hiện dựa trên kết quả thống kê có được từ kho ngữ liệu (Adam, 2008).

Về mặt ngôn ngữ, tiếng K'Ho thuộc ngữ hệ Nam Á, nhóm ngôn ngữ Môn - Khmer (Trần, 1999). Vào đầu thế kỷ XX, ngôn ngữ K'Ho được xây dựng bằng hệ thống chữ Latin với mục đích truyền đạo, về sau tiếng K'Ho đã được cải tiến nhiều lần và được sử dụng phổ biến bởi các nhóm dân tộc thiểu số tại Lâm Đồng, Đăk Nông và các tỉnh Đông Nam bộ (Trần, 1999). Đến nay, tiếng K'Ho được giảng dạy trong một số trường tiểu học tại vùng dân tộc thiểu số và để phục vụ cho công tác quản lý, phát triển kinh tế - xã hội, giữ gìn an ninh quốc phòng. Điều này đòi hỏi đội ngũ cán bộ công chức công tác ở các vùng có đồng bào dân tộc thiểu số phải biết sử dụng tiếng dân tộc bản địa trong giao tiếp và trong công tác theo qui định.

Nhằm góp phần ứng dụng khoa học công nghệ vào việc nghiên cứu ngôn ngữ của đồng bào thiểu số và cung cấp thông tin dự báo thời tiết cho đồng bào dân tộc K'Ho trên địa bàn tỉnh Lâm Đồng, đồng thời các bản tin dự báo thời tiết mang một lượng lớn thông tin mang tính cập nhật, do vậy một ứng dụng dịch tự động từ tiếng Việt sang tiếng K'Ho trong phạm vi bản tin dự báo thời tiết của Đài Phát thanh và Truyền hình tỉnh Lâm Đồng đã được xây dựng. Do tiếng Việt và tiếng K'Ho cùng ngữ hệ Nam Á nhưng lại thuộc nhóm ngôn ngữ khác nhau (Trần, 1999) nên phần chuyển ngữ thường được xử lý bằng cách sử dụng từ vựng, cụm từ và câu, thay vì bằng quy tắc cú pháp tổng quát. Qua nghiên cứu tổng quan các phương pháp thì phương pháp dịch tự động dựa vào thống kê (STMT) là phương pháp phù hợp với yêu cầu và mục tiêu của đề tài.

Trong báo cáo này, phương pháp STMT sẽ được trình bày trong việc áp dụng để xây dựng hệ dịch tự động Việt - K'Ho. Nội dung bài viết sẽ đề cập chi tiết về phương pháp STMT, các nguyên tắc thiết kế của ứng dụng cùng một số kết quả dịch tự động sẽ được trình bày để minh họa cho khả năng ứng dụng phương pháp STMT. Cấu trúc của bài viết được tổ chức như sau: Mục 2 trình bày phương pháp STMT; Mục 3 đề cập đến kết quả thực nghiệm. Cuối cùng là phần kết luận và hướng phát triển.

2. PHƯƠNG PHÁP STMT

2.1. Phương pháp

Dịch máy thống kê là quá trình dịch văn bản từ một ngôn ngữ này sang một ngôn ngữ khác dựa trên mô hình được sinh ra một cách tự động từ ngữ liệu song ngữ (*parallel corpus*). Phương pháp dịch máy thống kê lần đầu tiên được Antal và Peter (2009) đề cập trong bài báo với phương pháp sử dụng là mô hình kênh nhiễu. Bài toán được phát biểu như sau:

Cho một câu ngôn ngữ nguồn $v = v_1^J = v_1, v_2, \ldots, v_J$ (tiếng Việt), ta cần dịch sang câu ngôn ngữ đích $k = k_1^I = k_1, k_2, \ldots, k_I$ (tiếng K'Ho). Dịch máy thống kê sẽ chọn một câu k_{max} (có xác suất cao nhất) trong rất nhiều khả năng dịch được đưa ra.

$$k_{max} = \underbrace{arg\,max}_{k_1^I} p(k_1^I | v_1^J) \quad (1)$$

Sử dụng luật quyết định Bayes, $p(k|v)$ được tính như sau:

$$p(k|v) = \frac{p(v|k)*p(k)}{p(v)} \quad (2)$$

Do $p(v_1^J)$ và $p(k_1^J)$ không thay đổi với mỗi câu cần dịch khi dựa vào mô hình ngôn ngữ (ngữ pháp) nên công thức (1) có thể được viết lại như sau:

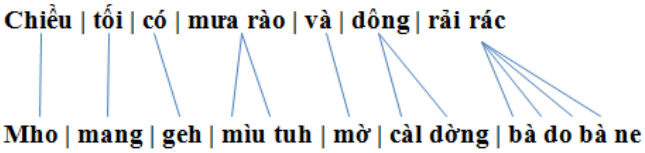
$$k_{max} = \underbrace{arg\,max}_{k_1^I} p(v_1^I | k_1^J) \quad (3)$$

Như vậy, để tính được k_{max} thì phải tính được các xác suất $p(v_1^I | k_1^J)$ phụ thuộc vào mô hình dịch với câu dịch thích hợp hơn sẽ có xác suất cao hơn. Xác suất này được ước lượng bằng cách sử dụng ngữ liệu song ngữ và sử dụng ý tưởng cách tiếp cận Maximum và mô hình gióng hàng.

Xác suất $p(v_1^I | k_1^J)$ được phân tích qua biến ẩn được thêm vào như công thức (4):

$$p(v_1^I | k_1^J) = \sum p(v_1^I, a_1^I | k_1^J) \quad (4)$$

Trong đó $p(v_1^I, a_1^I | k_1^J)$ được gọi là mô hình gióng hàng thống kê và gióng hàng a_1^I được gọi là biến ẩn. Gióng hàng xác định ánh xạ $i \rightarrow j = a_i$: Từ vị trí i của câu nguồn tương ứng với vị trí $j = a_i$ của câu đích. Chẳng hạn với trường hợp như Hình 2.



Hình 2. Mô hình gióng hàng

Gọi $a = (v_i, k_j)$ là một liên kết, ta có: $p(v_1^I | k_1^J) = \sum p(v_1^I, a_1^I | k_1^J)$. Trong đó, $\sum p(v_1^I, a_1^I | k_1^J)$ được xác định thông qua biểu thức (5).

$$\mathrm{p(v, a \mid k)} = \frac{S + \sum_{j=0}^{m} p(v_j \mid k_{a_j})}{l + m} \quad (5)$$

Trong đó: l, m lần lượt là độ dài (số từ) của v và k; S là số lần khi $p(v_j|k_{aj}) > 0$; $p(v_j|k_{aj})$ chính là xác suất của v_j khi có k_{aj} (hay nói cách khác là xác suất hai từ này có liên kết với nhau). Xác suất này hoàn toàn có thể thống kê được nhờ tập mẫu.

Chẳng hạn, với ví dụ trên, ta có công thức (6).

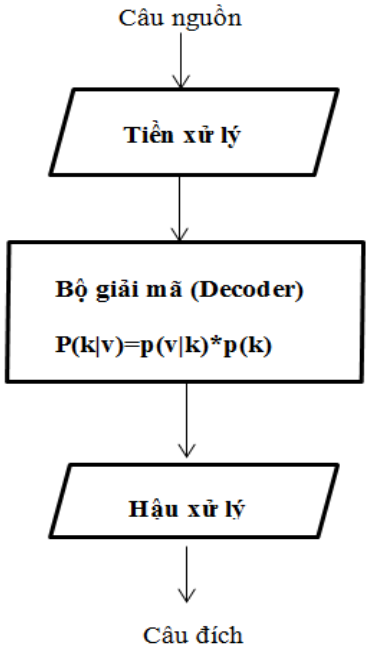
$$p(v,a|k) = \frac{7 + [p(chiều|mho) + \ldots + p(rải\ rác|bà\ do\ bà\ ne)]}{7+7} \quad (6)$$

Như vậy, xác suất $p(v,a|k)$ hoàn toàn tính được, do vậy $p(k|v)$ là tính được. Trong số các câu (k) trong tập mẫu, câu nào cực đại hoá (Maximum) được $p(k|v)$ chính là câu dịch cần chọn. Do vậy, thay vì xây dựng các từ điển, các quy luật chuyển đổi bằng phương pháp thủ công, hệ dịch này tự động xây dựng các từ điển, các quy luật dựa trên phương pháp thống kê. Rõ ràng, cách tiếp cận này không đòi hỏi một sự phân tích sâu về ngôn ngữ, chúng thực hiện hoàn toàn tự động các quá trình phân tích, chuyển đổi, tạo câu dựa trên kết quả thống kê có được từ kho ngữ liệu.

Trong phần dưới đây sẽ trình bày cụ thể công thức đã nêu trên và thay vì sử dụng mô hình gióng hàng a với từng cặp tiếng Việt-K'Ho mà thay vào đó là câu tiếng Việt với bản dịch của tiếng K'Ho thông qua kho dữ liệu ngữ liệu song ngữ.

2.2. Sơ đồ dịch máy thống kê (Statistical Machine Translation - SMT)

Sơ đồ dịch máy thống kê được mô tả như Hình 3 sau đây.



Hình 3. Sơ đồ dịch máy thống kê

2.2.1. Tiền xử lý

Khi nhận được văn bản đầu vào (tiếng Việt) hệ thống cần thực hiện việc xử lý phân đoạn từ tiếng Việt để tiện cho việc "xử lý" trong STMT. Tiếng Việt là một ngôn ngữ đơn lập, không biến hình, các ký tự được dựa trên hệ chữ cái Latin. Từ trong tiếng Việt ở đây lại không được xác định bởi khoảng trắng. Một từ tiếng Việt có thể được tạo bởi một hoặc nhiều hình vị và mỗi hình vị phân tách nhau bởi các khoảng trắng. Do vậy để tiến tới những ứng dụng xa hơn về xử lý ngôn ngữ tiếng Việt như gán nhãn chức năng cú pháp, phân tích cú pháp hay cụ thể để phục vụ cho việc dịch máy thì việc đầu tiên là phải giải quyết bài toán tách từ. Các nhà nghiên cứu đã đề xuất một số hướng tiếp cận để giải quyết bài toán tách từ. Nhìn chung, các hướng tiếp cận đó được chia thành hai loại: i) Tiếp cận dựa trên từ điển và ii) Tiếp cận dựa trên thống kê.

Nghiên cứu này thực hiện việc phân đoạn bằng phương pháp tiếp cận dựa trên từ điển. Ý tưởng của phương pháp này là duyệt một câu từ trái sang phải và chọn từ có nhiều tiếng nhất xuất hiện trong từ điển. Đây là một trong những phương pháp đơn giản nhưng có thể gặp phải rất nhiều các trường hợp nhập nhằng trong tiếng Việt. Tuy nhiên, nghiên cứu này chủ yếu triển khai thử

nghiệm hệ thống dịch tự động trong một chủ đề nhất định, cụ thể là bản tin dự báo thời tiết, vì vậy vấn đề về nhập nhằng ngữ nghĩa sẽ ít xuất hiện.

Sau khi kết thúc giai đoạn tiền xử lý thì mỗi câu trong đoạn văn bản đã được tách từ. Mỗi câu này sẽ là đầu vào của khối xử lý tiếp theo, khối lọc ra những câu có xác suất cao nhất. Có thể nói đây là khối xử lý phức tạp nhất, tốn nhiều thời gian nhất và chất lượng của nó sẽ ảnh hưởng đến hiệu suất dịch của toàn bộ hệ thống.

2.2.2. Bộ giải mã (Decoder)

Phần tiếp theo của một hệ dịch máy thống kê là chức năng tìm kiếm câu đích (giải mã). Chức năng của một bộ giải mã là từ câu nguồn V sẽ tìm câu cần dịch K sao cho tích của hai xác suất mô hình dịch và mô hình ngôn ngữ là lớn nhất. Như đã trình bày ở trên, theo như công thức (5) và theo mô hình gióng hàng a thì $p(v_1^J|k_1^J)$ sẽ được xác định bởi:

$$\mathrm{p(v,\ a\ |\ k)} = \frac{S + \sum_{j=0}^{m} p(v_j \mid k_{a_j})}{l+m} \tag{7}$$

Ví dụ: Với một câu đâu vào như sau: *v = Hôm nay nắng nhiều quá (hôm nay | nắng | nhiều | quá).* Trong kho dữ liệu ngữ liệu song ngữ có câu: *k = ngai tơngai gel du ết (có bản dịch là Ngày nắng ít quá (ngày | nắng | ít | quá)).* Mô hình gióng hàng a sẽ là cặp câu liên kết giữa tiếng Việt và bản dịch của tiếng K'Ho thông qua kho dữ liệu ngữ liệu song ngữ như sau:

v: Hôm nay | nắng | nhiều | quá

k:Ngày | nắng | ít | quá

Dựa vào từ điển đồng nghĩa ta có các xác suất như sau:

p(hôm nay|ngày)=0.7, p(nắng|nắng)=1, p(nhiều|ít)=0.3, p(quá|quá)=1

Áp dụng công thức (7), ta có:

$$p(v,a|k) = \frac{4+[p(\text{hôm nay}|\text{ngày})+\ldots+p(\text{quá}|\text{quá})]}{4+4} = \frac{4+3.0}{4+4} = 0.875$$

Như vậy, sau khi chọn được câu có k có xác suất cao nhất, thì sẽ đến bước tiếp theo là hậu xử lý.

2.2.3. Hậu xử lý

Đầu vào của khối xử lý này là câu có k_{max} xác suất cao nhất được chọn đối với câu cần dịch thông qua bộ giải mã. Chỉ còn một pha cuối cùng chính là pha thay thế, thêm và xóa các từ cho câu đầu vào để có được câu dịch cần tìm. Thực chất đây là sự điều chỉnh phần câu dịch (ngôn ngữ đích) của mẫu để nó trở thành câu dịch cuối cùng.

Ví dụ: *v: hôm nay nắng nhiều quá; k: ngày nắng ít quá (K'Ho: ngai tơngai gel du ết).*

- *Bước 1*: Đánh dấu các từ có thể thay thế, ví dụ ở đây từ "nhiều" là thành phần thay thế của "ít".

- *Bước 2*: Thực hiện xóa các từ ở câu k. Kết quả hiện tại của câu k:

Việt: ~~ngày~~ nắng ít quá.

K'Ho: ~~ngai~~ tơngai gel du ềt.

- *Bước 3*: Thực hiện thêm các từ còn thiếu, xóa các từ dư thừa của câu k để giống hoàn toàn với câu input. Kết quả hiện tại của câu k:

  Việt: hôm nay ~~ngày~~ nắng ít quá.

  K'Ho: ngai do ~~ngai~~ tơngai gel du ềt.

- *Bước 4*: Thực hiện pha thay thế, ở đây, từ "ít" chính là từ được thay thế bởi "nhiều". Tra trong từ điển song ngữ từ nhiều có nghĩa là "rà". Việc chính ở đây là thay thế trong câu ví dụ từ "ít" trong câu K'Ho thành từ "nhiều" tương ứng. Trong câu song ngữ như đã nói ở phần kho ngữ liệu, có một trường được gọi là đánh dấu liên kết sẽ biết được từ "ít" trong câu ví dụ tiếng Việt sẽ tương ứng với từ nào trong câu ví dụ K'Ho. Ở đây "ít" chính là từ "du ềt" và cuối cùng chỉ cần thay thế từ "du ềt" thành từ "rà". Kết quả: *ngai do ~~ngai~~ tơngai gel rà*.

Vậy câu dịch cuối cùng là "*ngai do tơngai gel rà*". Ở đây, có một đánh giá đối với pha thay thế. Thay vì phải xóa từ "ít" và thêm vào câu ví dụ từ "nhiều" thì kết quả sau khi thực hiện pha tạo mẫu: "*ngai do ~~ngai~~ tơngai rà gel*". Bởi thuật toán thêm từ sẽ dựa trên từ đứng sau nó, ở đây từ "quá" đứng sau nó vậy từ "nhiều" sẽ được thêm trước từ "quá". Vì vậy, kết quả sẽ có một chút sai lệch so với câu k, từ đó mà thể hiện được vai trò của pha thay thế trong trường hợp này.

3. KẾT QUẢ THỰC NGHIỆM

Ứng dụng dịch văn bản Việt - K'Ho dựa trên phương pháp STMT đã được xây dựng với phạm vi là dịch các bản tin dự báo thời tiết của Đài Phát thanh và Truyền hình tỉnh Lâm Đồng. Chức năng cơ bản là dịch văn bản tiếng Việt thành tiếng K'Ho với phạm vi như trên. Theo như thiết kế, hệ thống cần sử dụng đến ba loại dữ liệu chính: Từ điển song ngữ Việt – K'Ho; Từ điển đồng nghĩa; và Kho dữ liệu song ngữ. Để việc xử lý trong chương trình sau này được thuận tiện thì dữ liệu sẽ được cấu trúc và quản lý bằng hệ quản trị SQL Server. SQL Server là viết tắt của Structure Query Language, nó là một công cụ quản lý dữ liệu được sử dụng phổ biến ở nhiều lĩnh vực. Hầu hết các ngôn ngữ bậc cao đều có trình hỗ trợ SQL như VisualBasic, Oracle,Visual C… Các chương trình ứng dụng và các công cụ quản trị cơ sở dữ liệu (CSDL) cho phép người sử dụng truy cập tới CSDL mà không cần sử dụng trực tiếp SQL. Nhưng khi chạy những ứng dụng đó thì phải sử dụng SQL. Chương trình thực nghiệm với cấu trúc từ điển song ngữ Việt - K'Ho như Hình 4, cấu trúc từ điển đồng nghĩa được mô tả như Hình 5 và cấu trúc kho ví dụ song ngữ được mô tả như Hình 6.

ID	VietWord	KHoEqua	VietExp	KHoExp
1	tất cả	ală	NULL	NULL
2	các	ală	NULL	NULL
3	rộng	ananøg	NULL	NULL
4	phá hoại	aniai	NULL	NULL
5	chỗ này	anít do	NULL	NULL
6	buổi sáng	àng drim	NULL	NULL
7	phát hiện	bàn gõ	NULL	NULL
8	chiếu	bël	NULL	NULL
9	đầy	bềng	NULL	NULL

Hình 4. Cấu trúc từ điển Việt - K'Ho

ID	VietEntry	VietEqual1	VietEqual2	VietEqual5	VietCategory	VietLevel
1	tất cả	hết thảy	tất thảy	NULL	NULL	0
2	rộng	rộng rãi	mênh mông	NULL	NULL	0
3	phá hoại	phá	phá hủy	NULL	NULL	0
4	nhiệt độ	độ ẩm	NULL	NULL	NULL	0
5	thời tiết	khí hậu	trời	NULL	NULL	0
6	buổi sáng	sáng mai	NULL	NULL	NULL	0
7	đầy	đầy đủ	đầy ắp	NULL	NULL	0
8	thúc đẩy	phát triển	NULL	NULL	NULL	0
9	ở ngoài	bên ngoài	NULL	NULL	NULL	0
10	dịu dàng	mềm mại	nhẹ nhàng	NULL	NULL	0
11	đẹp	đẹp đẽ	xấu	NULL	NULL	0
12	mây	nước	biển	NULL	NULL	0
13	hôm kia	ngày kia	NULL	NULL	NULL	0
14	thêm	tăng	cộng	NULL	NULL	0
15	tỉnh	thành phố	huyện	NULL	NULL	0

Hình 5. Cấu trúc từ điển đồng nghĩa

ID	VietPhr	SplitPhr	KHoPhr	WordAligh
1	nước lũ tràn về cuốn trôi cả chòi	nước lũ,tràn,về,cuốn trôi,cả,chòi	dà sòr kŏ gơ pơrdô jơh kù	1_2,3,4,5,6,7
2	thời tiết xấu	thời tiết,xấu	mìu càl ờ diơng	1_2,3_4
3	mưa nhiều	mưa,nhiều	rà mìu	2,1
4	mưa kéo dài	mưa,kéo dài	mìu jit mìu jir	1_3,2_4
5	ngày có nắng ít quá	ngày,có,nắng,ít,quá	ngai tơngai gel du ết	1,0,2,4_5,3
6	có hôm gió thổi mạnh như bão	có,hôm,gió,thổi,mạnh,như,bão	geh ngai càl khŏm pràn...	1,2,3,4,5,6,7
7	chiều tối có mưa rào và dông r...	chiều,tối,có,mưa rào,và,dông,rải rác	mho mang geh mìu tuh...	1,2,3,4_5,6,7_8,9...
8	bản tin dự báo thời tiết đêm 01...	bản tin,dự báo,thời tiết,đêm,01,ng...	bản tin yal lai mìu càl m...	1_2,3_4,5_6,7,8,...
9	bản tin dự báo thời tiết	bản tin,dự báo,thời tiết	bản tin yal lai mìu càl	1_2,3_4,5_6

Hình 6. Cấu trúc tập mẫu

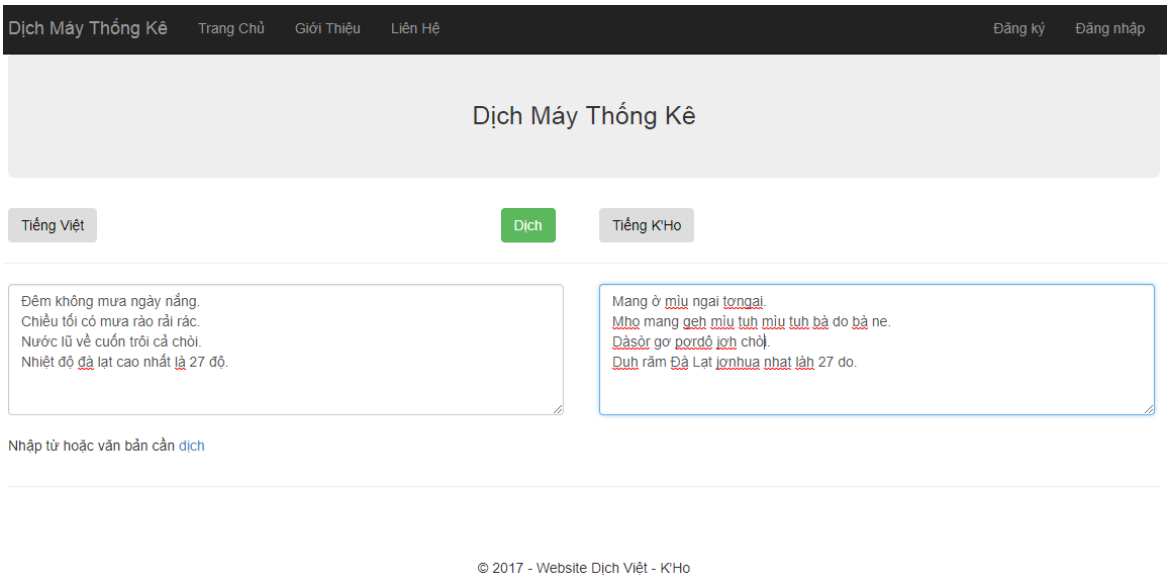
Với phạm vi bản tin thời tiết của Đài Phát thanh và Truyền hình tỉnh Lâm Đồng, các kho dữ liệu đã được xây dựng bao gồm:

- *Kho dữ liệu mẫu*: Gồm 212 cặp câu song ngữ Việt – K'Ho được trích từ các bản tin dự báo thời tiết của các năm 2015, 2016 và 2017 của Đài Phát thanh và Truyền hình Lâm Đồng; Báo Lâm Đồng; và Đài Tiếng nói Việt Nam;

- *Từ điển Việt - K'Ho*: Gồm 622 từ (Trần, 2014);

- *Từ điển đồng nghĩa*: Bao gồm 64 bản từ đồng nghĩa, mỗi từ sẽ có một hay nhiều từ đồng nghĩa (Nguyễn, 2001) và ứng với chúng là xác suất đồng nghĩa; Cùng nghĩa có thể thay thế nhau thì xác suất p(k|v) bằng 1 và ngược lại thì khoảng cách bằng 0, càng sát nghĩa thì xác suất càng gần 1.

Ngôn ngữ lập trình C#.NET đã được sử dụng với môi trường phát triển là Visual Studio 2013 để xây dựng ứng dụng chạy trên hệ điều hành Windows. Giao diện của ứng dụng như Hình 7.



Hình 7. Giao diện của ứng dụng

4. KẾT LUẬN

Ứng dụng dịch văn bản Việt - K'Ho dựa trên phương pháp STMT đã được xây dựng thành công. Ứng dụng dịch khá hiệu quả và câu dịch có chất lượng tốt trong phạm vi bản tin thời tiết của Đài Phát thanh và Truyền hình tỉnh Lâm Đồng. Nhược điểm của hệ thống là đòi hỏi phải có kho ví dụ song ngữ phong phú, từ điển song ngữ và từ điển đồng nghĩa đầy đủ thông tin thì độ chính xác của câu dịch sẽ càng cao. Ứng dụng cần hoàn thiện các nhược điểm trên để tiến tới xây dựng các công cụ phức tạp hơn trong xử lý ngôn ngữ tiếng K'Ho như: Dịch văn bản cho nhiều lĩnh vực; Nhận dạng và tổng hợp tiếng K'Ho…

TÀI LIỆU THAM KHẢO

Adam, L. (2008). Statistical machine translation. *ACM Computing Surveys, 40*(3), 1-49.

Antal, V. D. B., & Peter, B. (2009). Memory-based machine translation and language modelling. *The Prague Bulletin of Mathematical Linguistics*, (91), 17-26.

Đào, N. T. (2007). *Nghiên cứu về dịch thống kê dựa vào cụm từ và thử nghiệm với cặp ngôn ngữ Anh - Việt*. (Luận văn Thạc sĩ), Học viện Công nghệ Bưu chính Viễn thông, Việt Nam.

Nguyễn, V. T. (2001). *Từ điển từ đồng nghĩa tiếng Việt.* Hà Nội, Việt Nam: NXB. Giáo dục.

Pushpak, B. (2006). *Machine translation.* Florida, USA: CRC Press.

Satoshi, S., & Makoto, N. (1990). *Toward memory-based translation*. Paper presented at The 13th Conference on Computational Linguistics, Finland.

Trần, L. Q. (2006). *Kỹ thuật dịch máy và ứng dụng vào tài liệu hàng không*. (Luận văn Thạc sĩ), Trường Đại học Bách khoa Hà Nội, Việt Nam.

Trần, S. T. (1999). *Dân tộc - dân cư Lâm Đồng*. Hà Nội, Việt Nam: NXB. Thống kê.

Trần, V. L. (2014). *Từ điển K'Ho - Việt*. Hà Nội, Việt Nam: NXB. Giáo dục.