DỊCH TỰ ĐỘNG VIỆT- K'HO SỬ DỤNG PHƯƠNG PHÁP DỰA VÀO VÍ DỤ MẪU

Nguyễn Minh Tuấn[a], Đinh Viết Tuấn[a*]

[a]Khoa Công nghệ Thông tin, Trường Đại học Đà Lạt, Lâm Đồng, Việt Nam

Nhận ngày 04 tháng 01 năm 2016
Chỉnh sửa ngày 30 tháng 03 năm 2016 | Chấp nhận đăng ngày 31 tháng 03 năm 2016

Tóm tắt

Một ứng dụng dịch tự động 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 trên ví dụ mẫu (EBMT). 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 EBMT.

Từ khóa: EBMT; Dịch máy; Dịch tự động; Dịch tự động dựa vào ví dụ mẫu; MT.

1. GIỚI THIỆU

Hiện nay, trên thế giới có khoảng 5650 ngôn ngữ khác nhau [1], với 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.

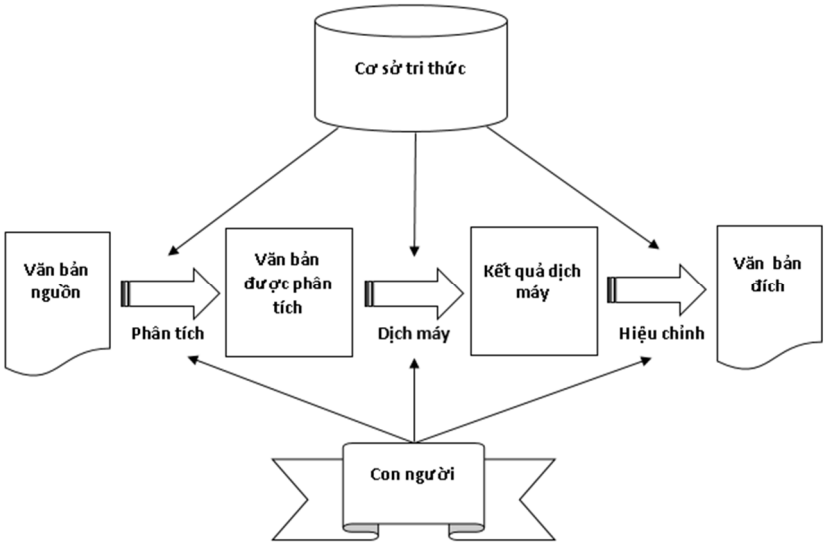
Hiện nay, khái niệm dịch tự động (dịch máy) đã được nhiều tác giả trong lĩnh vực xử lý ngôn ngữ tự nhiên định nghĩa, tuy có một vài điểm khác biệt nhưng hầu hết đều tương đương với định nghĩa của Pushpak Bhattacharyya, Indian Institute of Technology Bombay Mumbai [2] như sau:

* Tác giả liên hệ: Email: tuandv@dlu.edu.vn

"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 hay cho nhiều hơn hai ngôn ngữ".

Theo *"Kỹ thuật dịch tự động và ứng dụng vào tài liệu hàng không"* của Trần Lâm Quân thì quá trình xử lý tài liệu của dịch tự động được mô tả như Hình 1 [3]. 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: đầu tiên, văn bản được phân tích thành các thành phần, sau đó 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.

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:



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

- Dịch tự động dựa trên thống kê (Statistics Machine Translation - STMT) [4] 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 (corpus).

- Dịch dựa trên cụm từ (Phrase Based Machine Translation – PBMT) [1] 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) [5] 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.

- Dịch tự động dựa trên ví dụ (Example-Based Machine Translation - EBMT) [6] 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.

Về mặt ngôn ngữ, tiếng K’Ho thuộc ngữ hệ Nam Á, nhóm ngôn ngữ Môn – Khmer [7]. Vào đầu thế kỷ 20, 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 [7]. 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 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.

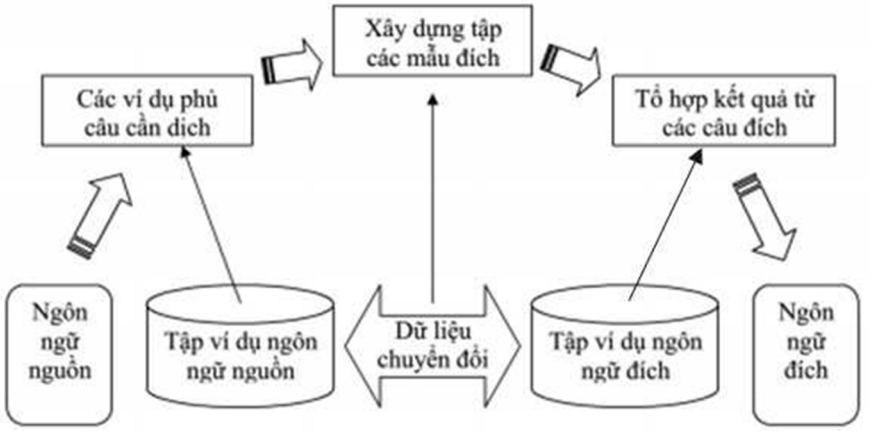
Nhằm góp phần ứng dụng khoa học công nghệ vào việc nghiên cứu ngôn ngữ của đồng bào thiểu số và cung cấp thông tin dự báo thời tiết cho đồng bào dân tộc K'Ho trên địa bàn tỉnh Lâm Đồng, đồng thời các bản tin dự báo thời tiết mang một lượng lớn thông tin mang tính cập nhật, do vậy một ứng dụng dịch tự động từ tiếng Việt sang tiếng K'Ho trong phạm vi bản tin dự báo thời tiết của đài Phát thanh 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 [7] 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 trên ví dụ mẫu (EBMT) là phương pháp phù hợp với yêu cầu và mục tiêu của đề tài.

Trong báo cáo này, phương pháp EBMT 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 EBMT, 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 EBMT. Cấu trúc của bài viết được tổ chức như sau: Phần 2 trình bày phương pháp EBMT. Phần 3 đề cập đến kết quả thực nghiệm. Cuối cùng là phần kết luận và hướng phát triển.

2. PHƯƠNG PHÁP EBMT

Ý tưởng của phương pháp EBMT được giới thiệu lần đầu tiên bởi Nagao trong dự án xây dựng hệ dịch tự động Nhật-Anh [8]. Sơ đồ một hệ EBMT, mà sau này được diễn giải bằng những thuật ngữ như: "Dịch bằng suy diễn từ ví dụ" hay "Dịch trên nguyên lý tương tự" [8], được mô tả như Hình 2.

Một hệ EBMT cụ thể đã được Sumita đề xuất với tên gọi là hệ dịch D3 (Dp-match Driven transDucer) dựa trên so khớp quy hoạch động [9]. Sở dĩ có tên như vậy, vì trong pha "tìm kiếm" (Retrieve), hệ chọn từ kho ví dụ những câu tương tự nhất với câu đầu vào căn cứ vào khoảng cách ngữ nghĩa giữa chúng thông qua giải thuật so khớp quy hoạch động (DP-Matching Algorithm) giữa hai dãy từ (word sequences).



Hình 2. Sơ đồ một hệ EBMT

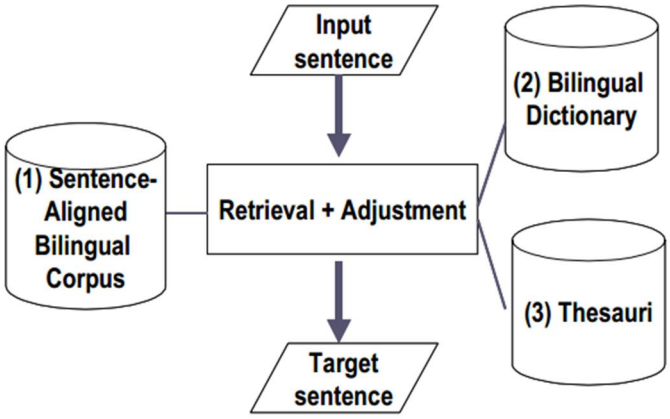
Hệ dịch D3 yêu cầu một tập mẫu, gồm các cặp câu song ngữ, nhưng trong quá trình dịch chúng phải được biểu diễn dưới dạng dãy từ (word sequence). Để dịch một câu đầu vào, hệ thống sẽ tìm kiếm trong tập ngữ liệu những cặp câu nào có phần ngôn ngữ nguồn tương tự nhất với nó. Khái niệm "tương tự" ở đây sẽ được lượng hoá bằng một độ đo ngữ nghĩa gọi là "edit-distance". Sau đó, với mỗi một cặp câu vừa được chọn ra, hệ thống sẽ so sánh phần ngôn ngữ nguồn của nó với câu đầu vào, lọc ra các thành phần khác nhau giữa chúng để tổng quát hoá câu ngữ liệu thành các mẫu (patterns). Công đoạn cuối cùng chỉ là chọn ra mẫu phù hợp nhất và thực hiện phép thay thế các thành phần khác nhau nói trên để có được câu dịch cần tìm từ phần ngôn ngữ đích của mẫu đó.

Hình 3 biểu diễn mô hình của một hệ dịch D3, giống như một hệ dịch Example-Based tổng quát hệ dịch D3 sử dụng 3 nguồn dữ liệu sau:

- Kho dữ liệu song ngữ (**Bilingual Corpus**): Tham gia vào giai đoạn "Tìm kiếm ngữ liệu tương tự", đây là kho dữ liệu quan trọng nhất.
- Từ điển đồng nghĩa (**Thesauri**): Sử dụng trong hai giai đoạn "Tìm kiếm ngữ liệu tương tự" và "Sản sinh mẫu".
- Từ điển song ngữ (**Bilingual Dictionary**): dùng trong 2 giai đoạn "Sản sinh mẫu" và "Thay thế".

Trong khối Retrieval và Adjustment bao gồm 4 bước:

- Tìm kiếm ngữ liệu tương tự **(Retrieve)**.
- Chọn ngữ liệu phù hợp nhất (**Select**).
- Sản sinh mẫu (**Generate**).
- Thay thế (**Substitute**).



Hình 3. Mô hình của hệ dịch D3

Giải thuật DP-Matching:

Duyệt từng câu trong tập mẫu, sau đó sử dụng giải thuật tính khoảng cách (Distance) giữa nó với câu đầu vào theo công thức (1):

$$dist = \frac{I+D+2\sum SEMDIST}{L_{input}+ L_{example}} \quad (1)$$

Trong đó:

- I, D lần lượt là số Insertion và Deletion (số từ cần thêm vào và xóa đi để thu được input từ example).
- SEMDIST: là khoảng cách về mặt ngữ nghĩa được dùng trong pha thay thế sau này. (SEMDIST giữa 2 từ giống nhau sẽ là 0)
- L_{input}, $L_{example}$ lần lượt là độ dài (số lượng từ của câu hoàn chỉnh đã được tách ra)

Ta xét một ví dụ đơn giản sau với câu input và example như sau:

- Hôm nay nắng nhiều quá (hôm nay | nắng | nhiều | quá).

- Ngày nắng ít quá (ngày | nắng | ít | quá).

Với 2 câu trên thì I=D=0, có hai cụm từ khác nhau giữa 2 câu là "nhiều" và "ít" và lúc này từ điển đồng nghĩa được sử dụng, nếu đo được khoảng cách (0<=SEMDIST<=1) thì sẽ lấy từ trong từ điển, ngược lại khoảng cách giữa 2 cụm từ là 1 vì chúng không có độ đồng nghĩa về ngữ nghĩa. Giả sử "nhiều" và "ít" tìm thấy trong từ điển với SEMDIST với khoảng cách là 0.7, "hôm nay" và "ngày" không tìm thấy trong từ điển thì SEMDIST là 1. Từ công thức (1), khoảng cách giữa hai câu trên được tính:

$$\text{dist}(\text{input}, \text{example}) = \frac{0+0+2*(1+0+0.7+0)}{4+4} = \frac{3.4}{8} = 0.425 \quad (2)$$

Từ công thức (1), giải thuật DP - Matching được phân tích như sau:

Một mảng 2 chiều m bao gồm cột là độ dài của câu Input và hàng là độ dài của câu Example đã được phân đoạn sẽ được duyệt. Mảng 2 chiều m sẽ được khởi tạo như sau: m[0,0]=0, các phần tử m[i,0] (i=1->L_{input})=m[i-1,0]+1 và tương tự phần tử m[0,j](j=1->$L_{example}$)= m[0,j-1]+1. Sau đó các phần tử của hàng i cột j sẽ được tính theo với công thức (với 1<=i<= L_{input} và 1<=j<= $L_{example}$):

$$m[i,j] = \min\begin{pmatrix} m[i-1,j-1] + 2 * \\ \text{SEMDIST}(m[i], m[j]), m[i-1,j] + 1, m[i,j-1] + 1 \end{pmatrix} \quad (3)$$

Theo đó m[L_{input}, $L_{example}$] sẽ được tính và chia cho tổng L_{input}, $L_{example}$ thì ta sẽ có được khoảng cách giữa 2 câu. Mảng m cho ví dụ trên được tính như Hình 4.

Sau khi duyệt hết kho ví dụ mẫu và lấy được câu ví dụ có độ so khớp thấp nhất, tiếp theo pha hiệu chỉnh, chỉnh sửa dữ liệu để phù hợp với câu đầu vào sẽ được thực hiện trước khi xuất ra kết quả câu dịch cuối cùng.

		ngày	nắng	ít	quá
	0	1	2	3	4
hôm nay	1	2	3	4	5
nắng	2	3	2	3	4
nhiều	3	4	3	3.4	4.4
quá	4	5	4	4.4	3.4

Với SEMDIST("hôm nay", "ngày")=1; SEMDIST("nắng", "nắng")=0;
SEMDIST("nhiều", "ít")=0.7; SEMDIST("quá", "quá")=0;

-> dist(input,example)=3.4/8=0.425

Hình 4. Ví dụ giải thuật DP-Matching

Hiệu chỉnh câu:

- Pha sinh mẫu (xóa, thêm)

Đầu vào của khối xử lý này là câu ví dụ được chọn sao cho nó tương tự nhất đối với câu cần dịch thông qua giải thuật DP-Matching.

Việc tiếp theo chính là thực hiện công việc đánh dấu từ thay thế, thêm và xóa các từ cho câu đầu vào để tổng quát câu ví dụ giống với câu cần dịch.

Ví dụ:

Input: hôm nay nắng nhiều quá. Example: ngày nắng ít quá.

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

Input: hôm nay nắng **nhiều** quá. Example: ngày nắng **ít** quá.

Bước 2: thực hiện xóa các từ không thể thay thế ở câu ví dụ:

Input: hôm nay nắng **nhiều** quá. Example: ~~**ngày**~~ nắng **ít** quá.

Bước 3: thực hiện thêm các từ còn thiếu cho câu ví dụ để giống hoàn toàn với câu input:

Input: hôm nay nắng **nhiều** quá.

Example: ***hôm nay*** ~~**ngày**~~ nắng **ít** quá.

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

- Pha thay thế

Sau khi qua quá trình tạo mẫu, hiện tại câu ví dụ hầu như đã giống hoàn toàn đối với câu cần dịch. Chỉ còn một pha cuối cùng chính là pha thay thế để có được câu dịch cần tìm. Thực chất đây là sự thay thế thành phần của câu dịch để nó trở thành câu dịch cuối cùng. Ở ví dụ trên từ *"ít"* chính là từ được thay thế bởi *"nhiều"*. Tra trong từ điển song ngữ từ nhiều ta được *"rà"*. Việc chính ở đây ta chỉ cần thay thế trong câu ví dụ từ *"ít"* trong câu ví dụ K'Ho thành từ *"nhiều"* tương ứng. Trong câu song ngữ ví dụ như đã nói ở phần kho ngữ liệu, có một trường được gọi là đánh dấu liên kết ta 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 ta chỉ cần thay thế từ *"du ềt"* thành từ "rà".

Input: hôm nay nắng **nhiều** quá.

Example:

- Việt: hôm nay nắng **ít** quá.

- K'Ho: ngai do tơngai gel **du ềt**.

Vậy câu dịch cuối cùng là *"ngai do tơngai gel rà"*. Ở đây, thay vì phải xóa từ "ít" và thêm vào câu ví dụ từ "nhiều" thì kết quả sau khi thực hiện pha tạo mẫu sẽ là: *"ngai do 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 ví dụ, 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 EBMT đã đề xuất đã được xây dựng với:

- Phạm vi: Dịch các bản tin dự báo thời tiết của đài Phát thanh truyền hình tỉnh Lâm Đồng.
- Chức năng: Dịch văn bản tiếng Việt thành tiếng K'Ho với phạm vi trên.
- Thiết kế và tổ chức lưu trữ dữ liệu:

Theo như thiết kế, hệ thống chúng ta cần sử dụng đến ba loại dữ liệu chính: từ điển song ngữ Việt – K'Ho, từ điển đồng nghĩa và kho ví dụ mẫu song ngữ. Để việc xử lý trong chương trình sau này được thuận tiện, truy xuất nhanh và sử dụng ít bộ nhớ, thì việc tổ chức và lưu trữ một cách hợp lý các loại dữ liệu này đóng vai trò hết sức quan trọng. Vì vậy dữ liệu sẽ được cấu trúc và tổ chức theo ngôn ngữ XML thay vì dùng các công cụ quản lý cơ sở dữ liệu quan hệ như SQL, Oracle...

Ngôn ngữ XML (Extensible Markup Language) là ngôn ngữ được định nghĩa bởi tổ chức mạng toàn cầu (World Wide Web Consortium - W3C), XML là một ngôn ngữ tổng quát dùng để biểu diễn thông tin dưới dạng các tài liệu có cấu trúc, định nghĩa dữ liệu thông qua các thẻ [10]. Để mô tả dữ liệu, các giản đồ XML (XML Schema) phải sẽ định nghĩa. Vì vậy, XML có thể mở rộng với các ngôn ngữ tự nhiên.

Cấu trúc từ điển song ngữ Việt - K'Ho như Hình 5:

```
<BilingualDictionary>
  <ID>1</ID>
  <VietWord>tất cả</VietWord>
  <KHoEqua>ală</KHoEqua>
  <VietExp></VietExp>
  <KHoExp></KHoExp>
</BilingualDictionary>
```

Hình 5. Cấu trúc từ điển Việt - K'Ho

Cấu trúc từ điển đồng nghĩa được mô tả như Hình 6.

```
<ThesaurusDictionary>
  <ID>1</ID>
  <VietEntry>tất cả</VietEntry>
  <VietEqual1>hết thảy</VietEqual1>
  <VietEqual2>tất thảy</VietEqual2>
  <VietEqual3></VietEqual3>
  <VietEqual4></VietEqual4>
  <VietEqual5></VietEqual5>
  <VietCategory></VietCategory>
  <VietLevel>0</VietLevel>
</ThesaurusDictionary>
```

Hình 6. Cấu trúc từ điển đồng nghĩa

Cấu trúc kho ví dụ song ngữ được mô tả như Hình 7:

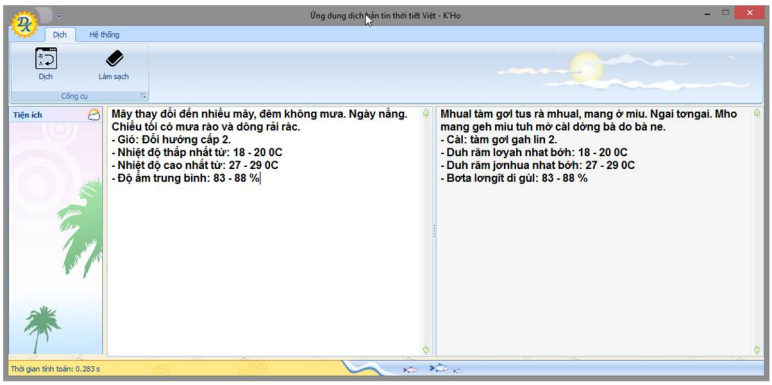
```
<ParallelCorpus>
  <ID>13</ID>
  <VietPhr>ngày nắng gián đoạn</VietPhr>
  <SplitPhr>ngày,nắng,gián đoạn</SplitPhr>
  <KHoPhr>ngai tongai tơm gơl</KHoPhr>
  <WordAlign>1,2,3_4</WordAlign>
</ParallelCorpus>
<ParallelCorpus>
  <ID>14</ID>
  <VietPhr>chiều tối và đêm có mưa rải rác</VietPhr>
  <SplitPhr>chiều,tối,và,đêm,có,mưa,rải rác</SplitPhr>
  <KHoPhr>mho mang mờ mang geh mìu vai nis</KHoPhr>
  <WordAlign>1,2,3,4,5,6,7_8</WordAlign>
</ParallelCorpus>
```

Hình 7. Cấu trúc kho ví dụ

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

- Kho ví dụ: bao gồm 212 cặp câu song ngữ Việt – K'Ho [11].
- Từ điển Việt - K'Ho: bao gồm 622 từ [12, 13, 14].
- Từ điển đồng nghĩa: bao gồm 64 bản từ đồng nghĩa, mỗi từ sẽ có một hay nhiều từ đồng nghĩa [15] và ứng với chúng là khoảng cách đồng nghĩa; cùng nghĩa có thể thay thế nhau thì khoảng cách bằng 0 và ngược lại thì khoảng cách bằng 1, càng sát nghĩa thì khoảng cách càng gần 0.

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 8.



Hình 8. Giao diện của ứng dụng

4. KẾT LUẬN

Ứng dụng dịch văn bản Việt - K'Ho dựa trên phương pháp EBMT 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 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, tuy nhiên điều này sẽ ảnh hưởng đáng kể đến tốc độ dịch của ứng dụng.

Dựa trên những kết quả đã nghiên cứu và xây dựng, cũng như những hạn chế gặp phải, chúng tôi đề xuất một số hướng phát triển của bài báo trong tương lai:

- Xây dựng website cho phép dịch trực tuyến, tạo diễn đàn cho người dùng thảo luận và đánh giá chương trình.

- Xây dựng ứng dụng dịch Việt-K'Ho và ngược lại với phạm vi mở rộng hơn và các kho dữ liệu được xây dựng phong phú hơn để có thể dịch văn bản trong các trong các ngữ cảnh khác nhau.

TÀI LIỆU THAM KHẢO

[1] Đào Ngọc Tú, *Nghiên cứu về dịch thống kê dựa vào cụm từ và thử nghiệm với cặp ngôn ngữ Anh – Việt*, Học viện công nghệ bưu chính viễn thông, (2007).

[2] Pushpak Bhattacharyya, *Machine Translation*, Indian Institute of Technology Bombay Mumbai, (2006).

[3] Trần Lâm Quân, *Kỹ thuật dịch máy và Ứng dụng vào tài liệu hàng không*, Hà Nội, (2006).

[4] Adam Lopez, *Statistical Machine Translation*, ACM Computing Surveys, Vol. 40, No. 3, Article 8, (2008).

[5] See Sato, S. Nagao, *Toward Memory-based Translation*, IPSJ-WG, (1990).

[6] Antal van den Bosch and Peter Berck, *Memory-Based Machine Translation and Language Modeling*, The Prague Bulletin of Mathematical Linguistics, (2009).

[7] Trần Sỹ Thứ, *Dân tộc - dân cư Lâm Đồng*, Việt Nam, (1999).

[8] Eiichiro Sumita and Hitoshi IIDA, *Example-Based Machine Translation*, ATR Interpreting Telephony Research Laboratories, Japan.

[9] Elichiro Sumita, *Exampled-based machine translation using DP-matching between word sequences*, Proceedings of the workshop on Data-driven methods in machine translation, Vol. 14, (2001).

[10] LINQ To XML Tutorials with Examples, http://www.dotnetcurry.com/showarticle.aspx?ID=564, (2014).

[11] http://www.lamdongtv.vn/ ; http://baolamdong.vn/ ; http://vov4.vov.vn/Kho.aspx

[12] Không rõ, *Từ vựng K'Ho - Việt*, Viêt Nam, (2014).

[13] Trần Văn Lệ, *Từ điển K'Ho - Việt*, Việt Nam, (2012).

[14] Sở Nội vụ - Sở Giáo dục và Đào tạo tỉnh Lâm Đồng, *Tài liệu dạy và học tiếng K'Ho*, Việt Nam, (2007).

[15] Nguyễn Văn Tu, Từ điển từ đồng nghĩa Tiếng Việt, Nhà xuất bản giáo dục, Việt Nam, (2001).

VIETNAMESE - K'HO MACHINE TRANSLATION USING EBMT APPROACH

Nguyen Minh Tuan[a], Dinh Viet Tuan[a*]

[a]The Faculty of Information Technology, Dalat University, Lamdong, Vietnam
[*]Corresponding author: tuandv@dlu.edu.vn

Article history
Received: January 04th, 2016
Received in revised form: March 30th, 2016
Accepted: March 31st, 2016

Abstract

This paper proposes the Example Based Machine Translation (EBMT) method for Vietnamese-K'Ho machine translation. Both Vietnamese and K'Ho are linguistically rooted in the South Asia but they belong to different language groups; therefore, Vietnamese-K'Ho or vice versa translation are widely conducted by exploiting vocabulary, phrases and sentences instead of the general syntax rules. The design principles of the application are described in details, along with the system interface. The machine translation results are also presented to illustrate the applicability of the EBMT method.

Keywords: Example based machine translation; MT; EBMT; Machine translation.