CẢI TIẾN CHẤT LƯỢNG DỊCH TỰ ĐỘNG BẰNG GIẢI PHÁP MỞ RỘNG KHO NGỮ LIỆU

Huỳnh Công Pháp, Đặng Đại Thọ, Nguyễn Văn Bình

Đại học Đà Nẵng

hcphap@gmail.com, ddtho.dt@gmail.com, binhsht@gmail.com

TÓM TẮT *- Chất lượng dịch tự động, nhất là đối với các ngôn ngữ ít được đầu tư và ít phổ biến (under resourced languages) như là tiếng Việt và tiếng các dân tộc thiểu số là vấn đề rất được quan tâm hiện nay. Bên cạnh nhiều giải pháp như cải tiến các thuật toán và phương pháp dịch tự động, thì giải pháp nâng cao chất lượng dịch bằng cách mở rộng kho ngữ liệu luôn là một hướng giải quyết rất hiệu quả đã được chứng minh [7]. Do đó, trong bài báo này chúng tôi tập trung đề xuất nhiều giải pháp hiệu quả nhằm cho phép xây dựng được các kho ngữ liệu không chỉ lớn về kích thước, đa dạng về ngôn ngữ mà còn tốt về chất lượng. Tất cả các giải pháp đã được nhóm tác giả tiến hành triển khai thực hiện rất công phu và đã thu được những kết quả đáng kể.*

Từ khóa *- Dịch tự động, kho ngữ liệu, chất lượng dịch, mở rộng kho ngữ liệu, cải tiến chất lượng dữ liệu.*

I. ĐẶT VẤN ĐỀ

Nhằm đáp ứng nhu cầu giao tiếp của con người, hiện nay các hệ thống dịch tự động đang được phát triển đáng kể cả về số lượng và chất lượng. Mặc dù vậy, chất lượng dịch tự động vẫn còn khá thấp so với mong muốn, đặc biệt là các ngôn ngữ ít được đầu tư và ít phổ biến (under resourced languages) như tiếng Việt, các tiếng dân tộc thiểu số. Ngay cả các ngôn ngữ phổ biến và có sự đầu tư rất lớn như tiếng Anh, tiếng Pháp, tiếng Trung Quốc, tiếng Nhật,… chất lượng dịch tự động qua lại giữa các ngôn ngữ này cũng còn nhiều vấn đề cần phải cải tiến. Một trong những nguyên nhân quan trọng ảnh hưởng đến chất lượng dịch tự động chính là chất lượng các kho ngữ liệu được sử dụng cho các hệ thống dịch. Thật vậy, các chiến dịch đánh giá chất lượng dịch tự động (evaluation campanges) như CSTAR, NESPOLE, IWSLT [1] đã được tổ chức rất quy mô với nhiều phương pháp đánh giá chủ quan (subjective evaluations) và khách quan (objective evaluations) khác nhau đã cho thấy rằng chất lượng dịch tự động đối với các cặp ngôn ngữ phổ biến cũng chỉ chấp nhận được với một số lĩnh vực nhất định và có chất lượng rất kém nếu như kho ngữ không đảm bảo cả về chất lượng và khối lượng.

Mặc dù hiện nay tồn tại nhiều kho dữ liệu dùng trong lĩnh vực tự động. Song, tất cả các kho ngữ liệu này đều có những nhược điểm đáng kể. Các kho ngữ liệu như EuroParl (11 ngôn ngữ, 34-55 triệu từ), JRC-Acquis (22 ngôn ngữ, 11-22 triệu từ), XinHua News (2 ngôn ngữ, 12-14 triệu từ),… có số lượng ngôn ngữ và khối lượng dữ liệu tương đối lớn. Tuy vậy, so với số lượng ngôn ngữ tự nhiên trên thế giới hiện nay (khoảng 6500 ngôn ngữ nói hiện nay) và khối lượng dữ liệu đủ để bao phủ hầu hết các lĩnh vực dịch thì các kho ngữ liệu trên so ra còn quá khiêm tốn để có thể cho phép xây dựng được một hệ thống dịch tự động chất lượng cao. Mặt khác, các kho ngữ liệu hiện nay có chất lượng còn khá thấp, dữ liệu đa số ở dạng thô và nhập nhằng, bởi lẽ đa số các kho ngữ liệu tồn tại dưới dạng văn bản hoặc chỉ có một số ít thông tin mô tả đi kèm [6].

Do đó, để có thể xây dựng được một hệ thống dịch tự động chất lượng và hoàn chỉnh, ngoài việc nghiên cứu cải tiến phương pháp dịch tự động, vấn đề quan trọng cần giải quyết đó là nghiên cứu giải pháp xây dựng được kho ngữ liệu lớn không chỉ về khối lượng dữ liệu, số cặp ngôn ngữ mà còn tốt về chất lượng.

Trong bài báo này, chúng tôi đề xuất các giải pháp mở rộng kho ngữ liệu cả về khía cạnh khối lượng và khía cạnh chất lượng, nhằm hướng đến giải pháp cải tiến chất lượng dịch của các hệ thống dịch tự động nói chung và nhất là các hệ thống dịch tự động tiếng Việt và tiếng dân tộc thiểu số ở Việt Nam.

Đối với việc mở rộng kho ngữ liệu theo khía cạnh khối lượng, chúng tôi đề xuất không chỉ giải pháp mở rộng về số lượng câu mà còn đề xuất giải pháp mở rộng số lượng ngôn ngữ. Đối với khía cạnh cải tiến chất lượng, chúng tôi đề xuất giải pháp khử nhập nhằng dữ liệu của các kho ngữ liệu bằng cách làm giàu thông tin cho dữ liệu của kho ngữ liệu. Thông tin được làm giàu có thể đơn giản chỉ là các mô tả thêm cho các thành phần dữ liệu ở dạng đơn giản như các chú thích, các từ đồng nghĩa, trái nghĩa,… hoặc có thể phức tạp đến mức mỗi thực thể từ hoặc cụm từ trong kho ngữ liệu sẽ được mô tả bởi một lớp hoặc tập các lớp dữ liệu của các ontology.

II. CÁC NGHIÊN CỨU LIÊN QUAN

Như đề cập ở trên, giải pháp mở rộng kho ngữ liệu của bài báo tập trung vào hai hướng nghiên cứu chính gồm mở rộng kho ngữ liệu theo hướng khối lượng và theo hướng chất lượng.

A. Tổng quan tình hình nghiên cứu mở rộng kho ngữ liệu theo khía cạnh khối lượng

Liên quan đến khía cạnh mở rộng kho ngữ liệu theo hướng khối lượng, chúng tôi tập trung nghiên cứu tổng quan các phương pháp và công trình liên quan đến hai vấn đề sau:

1. Mở rộng kho ngữ liệu theo hướng ngôn ngữ

Đối với vấn đề mở rộng kho ngữ liệu theo hướng ngôn ngữ, chúng ta trước hết có thể đề cập đến các công trình [2][3]. Các công trình này đã đưa ra giải pháp mở rộng kho ngữ liệu theo hướng ngôn ngữ bằng cách gọi các hệ thống dịch tự động để dịch dữ liệu hiện có của kho ngữ liệu sang các ngôn ngữ mới, sau đó đề xuất giải pháp cho các chuyên gia xem, chỉnh sửa để thu được dữ liệu có chất lượng trong ngôn ngữ mới tương đồng với dữ liệu gốc.

2. Mở rộng kho ngữ liệu theo hướng xây dựng/bổ sung dữ liệu

Liên quan đến vấn đề xây dựng/bổ sung dữ liệu mới vào kho ngữ liệu, chúng ta có thể kể đến các phương pháp xây dựng dữ liệu mới [2][4][5]. Các công trình này đề xuất việc xây dựng nội dung mới có chất lượng cho kho dữ liệu dịch tự động bằng cách trích lọc dữ liệu từ các kho dữ liệu hỗn tạp trong thực tế.

Ngoài ra, chúng ta phải kể đến đề xuất tại [6], công trình này đưa ra đề xuất xây dựng nội dung mới cho kho ngữ liệu bằng cách thu thập, trích lọc các câu song song từ các trang website đa ngữ.

B. Tổng quan tình hình nghiên cứu mở rộng kho ngữ liệu về khía cạnh chất lượng

Liên quan đến khía cạnh mở rộng kho ngữ liệu theo hướng chất lượng, chúng tôi tập trung nghiên cứu tổng quan các phương pháp và công trình liên quan đến hai vấn đề:

1. Mở rộng kho ngữ liệu theo hướng cải tiến chất lượng dữ liệu

Liên quan đến giải pháp cải tiến chất lượng kho ngữ liệu, chúng ta có thể kể đến các công cụ và hệ thống sinh mới và chỉnh sửa dữ liệu dịch. Trong đó, nổi bật nhất là hệ thống SECTra_w [6], hệ thống này cho phép người dùng nạp kho ngữ liệu vào và hiển thị kho ngữ liệu một cách trực quan, dạng song song rất thuận tiện cho việc kiểm tra và cải tiến chất lượng văn bản. Hệ thống này cũng là một môi trường cộng tác cho phép nhiều người tham gia cải tiến và chỉnh sửa kho ngữ liệu. Tiếp đến, chúng ta có thể kể đến các công cụ cục bộ như Mtpost-editor, phát triển bởi NIST, hoặc công cụ SYSTRAN Review Manager được dùng ở Công ty Systran. Ngoài ra, chúng ta cũng tìm thấy một số hệ thống được triển khai dưới mô hình mạng như Google Translator Toolkit, BEYtrans, Yakushite.net, Translationwiki.net, Traduwiki, Caitra [1].

2. Mở rộng kho ngữ liệu theo khía cạnh ngữ nghĩa

Liên quan đến giải pháp mở rộng kho ngữ liệu theo khía cạnh ngữ nghĩa, đây là một giải pháp khá mới và là một hướng đi hứa hẹn nhằm giúp cho việc khai thác kho ngữ liệu hiệu quả. Đây cũng là hướng mà chúng tôi đang tập trung nghiên cứu, trong bài báo này chúng tôi cũng sẽ đề cập đến khía cạnh này như là một giải pháp mở rộng chất lượng cho các kho ngữ liệu [11]. Trong công trình đã công bố, chúng tôi đã đề cập đến giải pháp mở rộng ngữ nghĩa song cũng mới dừng lại ở mức đề xuất giải pháp chung, trong bài báo này chúng tôi sẽ đưa ra các giải pháp cụ thể hơn.

III. GIẢI PHÁP MỞ RỘNG KHO NGỮ LIỆU

Như đã đề cập ở phần trên, nếu có được kho ngữ liệu đa ngữ đủ lớn về khối lượng, tốt về chất lượng thì chắc chắn chúng ta sẽ cải tiến được chất lượng dịch của các hệ thống dịch tự động hiện nay. Thật vậy, trong sự nghiệp nghiên cứu hơn 50 năm về lĩnh vực xử lý ngôn ngữ tự nhiên và dịch tự động, GS. Christian Boitet, phòng thí nghiệm LIG, Trường Đại học Joseph Fourier, Grenoble, Pháp đã chứng minh sự ảnh hưởng và mối quan hệ mật thiết giữa chất lượng và khối lượng của kho ngữ liệu với chất lượng dịch của các hệ thống dịch tự động trong công trình công bố của mình [4]. Do đó, vấn đề nghiên cứu và đề xuất các giải pháp mở rộng kho ngữ liệu là hết sức cần thiết để cải tiến chất lượng dịch tự động.

Giải pháp mở rộng kho ngữ liệu mà chúng tôi đề xuất trong bài báo này sẽ gồm các giải pháp:

- Mở rộng hay làm tăng thêm khối lượng kho ngữ liệu;

- Mở rộng hay cải tiến chất lượng dữ liệu của các kho ngữ liệu.

A. Mở rộng khối lượng kho ngữ liệu:

Đối với việc mở rộng khối lượng kho ngữ liệu, chúng tôi đề xuất hai giải pháp cụ thể gồm:

- Giải pháp thứ nhất là tìm cách hợp nhất các kho ngữ liệu hiện có để tạo ra một kho ngữ liệu lớn hơn.

- Giải pháp thứ hai là xây dựng hệ thống cho phép mở rộng ngôn ngữ cũng như thêm dữ liệu vào kho ngữ liệu.

1. Hợp nhất các kho ngữ liệu

Trước khi trình bày cụ thể giải pháp này, chúng ta cần làm rõ khái niệm hợp nhất kho ngữ liệu. Trong các công trình nghiên cứu [6][4][11] đã chỉ rõ rằng, hiện nay tồn tại rất nhiều kho ngữ liệu có kích thước, số lượng ngôn ngữ, định dạng và cấu trúc khác nhau. Hợp nhất các kho ngữ liệu chính là tìm cách trộn, liên kết các kho ngữ liệu này lại với nhau để tạo nên một kho ngữ liệu lớn hơn có cùng cấu trúc, định dạng và với nhiều cặp ngôn ngữ hơn. Ví dụ, có 2 kho ngữ liệu song song: kho thứ nhất (C_1) gồm 2 cặp ngôn ngữ Anh-Pháp và Anh-Việt gồm 5000 cặp câu; kho thứ hai (C_2) gồm 2 cặp ngôn ngữ Anh-Việt và Việt - Khmer gồm 5000 cặp câu. Sau khi hợp nhất 2 kho ngữ liệu trên, chúng ta sẽ có được một kho ngữ liệu lớn hơn có số lượng từ 5000 đến 10000 cặp câu với 4 cặp ngôn ngữ Anh-Pháp, Anh-Việt, Anh-Khmer và Việt-Khmer. Một cách tổng quát, nếu xem mỗi kho ngữ liệu là một tập hợp (Ci), gồm các cặp câu và các cặp

ngôn ngữ thì kho ngữ liệu hợp nhất (C_u) sẽ là kết quả của phép hợp của các kho ngữ liệu thành viên và được biểu diễn bởi công thức sau:

$$C_u = C_1 \cup C_2 \cup C_3 \cup \ldots\ldots \cup C_n = \bigcup_{i=1}^{n} C_i \quad (1)$$

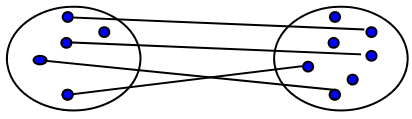
Giải pháp hợp nhất các kho ngữ liệu sẽ bao gồm một số vấn đề cụ thể cần giải quyết như sau:

a) Hợp nhất dữ liệu

Hợp nhất dữ liệu tức là liên kết dữ liệu (alignment) hay là xác định tính tương đồng giữa các đơn vị dữ liệu cùng hoặc khác ngôn ngữ của các kho ngữ liệu.

Liên kết các đơn vị dữ liệu cùng một ngôn ngữ giữa các kho ngữ liệu thực chất là quá trình so sánh văn bản để tìm ra được các cặp đơn vị dữ liệu giữa các kho ngữ liệu tương đồng với nhau. Nếu xét ở mức so sánh chuỗi ký tự, việc liên kết dữ liệu dễ dàng được thực hiện thông qua việc áp dụng một số thuật toán phổ biến hiện nay như Edit distance, BLEU, NIST, WER,… Nếu xét ở mức độ so sánh ngữ nghĩa, việc liên kết dữ liệu sẽ rất phức tạp. Tuy nhiên, với mục đích hợp nhất dữ liệu các kho ngữ liệu, chúng ta chỉ dừng lại ở mức so sánh chuỗi ký tự. Một cách tổng quát, liên kết dữ liệu cùng ngôn ngữ giữa hai kho ngữ liệu có thể biểu diễn bằng công thức sau:

$$C = \{(x,y) \mid x \in C_1 \wedge y \in C_2 \wedge x \approx y\} \quad (2)$$



Trong đó: x là đơn vị dữ liệu của kho ngữ liệu C1, y là đơn vị dữ liệu của kho ngữ liệu C_2 sao cho x tương đồng nội dung với y.

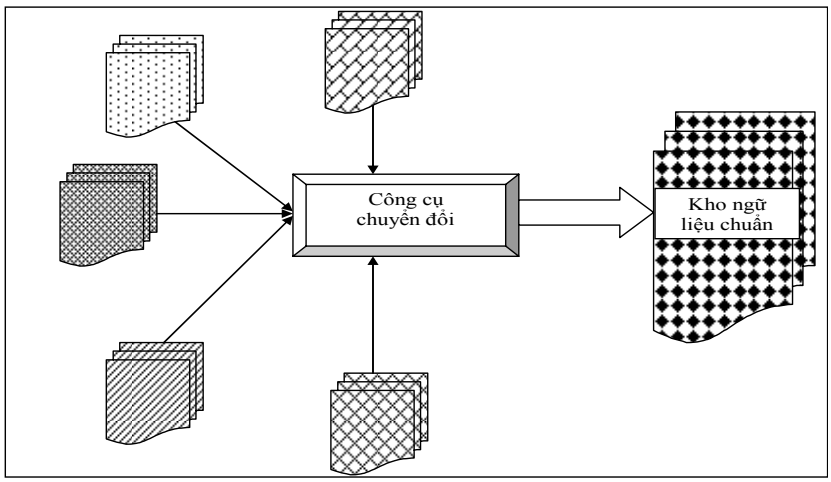
Đối với việc liên kết các đơn vị dữ liệu khác ngôn ngữ, việc liên kết dữ liệu sẽ được thực hiên thông qua các thuật toán và công cụ đối sánh văn bản (text aligner) đã tồn tại [5][8]. Một cách tổng quát, liên kết dữ liệu khác ngôn ngữ giữa hai kho ngữ liệu có thể biểu diễn bằng công thức sau:

$$C = \{(x,y) \mid x \in C_{1L1} \wedge y \in C_{2L2} \wedge f(x) \approx f(y)\} \quad (3)$$

Trong đó: x là đơn vị dữ liệu trong ngôn ngữ L_1 của kho ngữ liệu C_{1L1}, y là đơn vị dữ liệu trong ngôn ngữ L_2 của kho ngữ liệu C_{2L2} và f là hàm xác định tính tương đồng giữa x và y.

b) Hợp nhất định dạng và cấu trúc các kho ngữ liệu

Một vấn đề lớn cần giải quyết đối với bài toán hợp nhất các kho ngữ liệu đó là hợp nhất các định dạng và cấu trúc các kho ngữ liệu. Thật vậy, hiện nay các kho ngữ liệu được xây dựng bởi các tổ chức, cá nhân và nhóm nghiên cứu khác nhau. Do đó, các kho ngữ liệu sẽ khác nhau về kích thước, định dạng dữ liệu và cấu trúc. Để có thể hợp nhất được các kho ngữ liệu trước tiên chúng ta cần nghiên cứu đề xuất một cấu trúc và định dạng chuẩn có thể biểu diễn được tất cả các kho ngữ liệu. Sau đó nghiên cứu và xây dựng được công cụ chuyển đổi các kho ngữ liệu đang tồn tại để xây dựng được kho ngữ liệu với cấu trúc và định dạng chuẩn đã đề xuất [6][5].



Hình 1. Giải pháp chuyển đổi các kho ngữ liệu

2. Mở rộng ngôn ngữ của kho ngữ liệu

Mở rộng khối lượng kho ngữ liệu bằng cách mở rộng ngôn ngữ không phải là ý tưởng mới. Thực tế cho thấy giải pháp này đã mang lại kết quả nhất định. Tuy nhiên, để nâng cao hơn nữa hiệu quả của cách làm này cần phải có giải pháp toàn diện hơn.

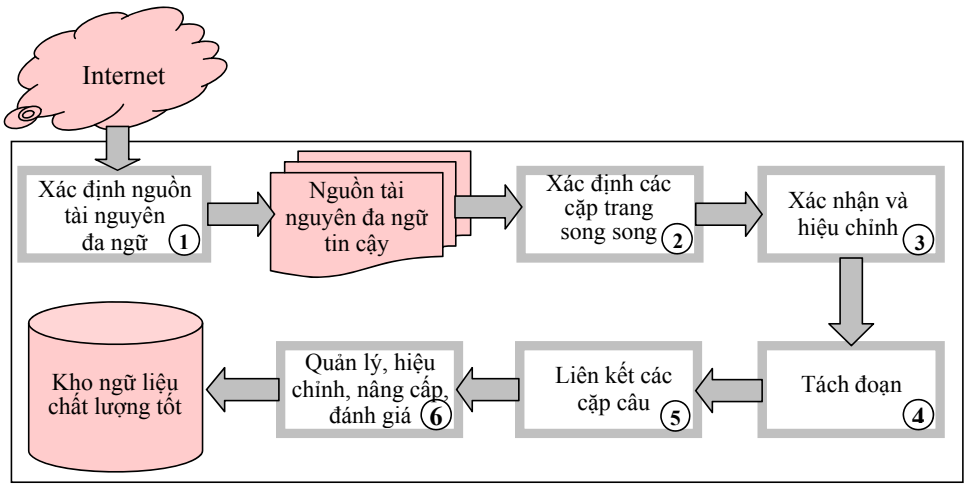
Việc mở rộng ngôn ngữ cho các kho ngữ liệu được thực hiện bằng cách gọi các hệ thống dịch tự động như Google Translator, Systrans, Reverso,… để dịch các dữ liệu nguồn sang ngôn ngữ cần mở rộng. Về mặt kỹ thuật đây không phải là vấn đề dễ thực hiện. Bởi vì để thực hiện ý tưởng này, chúng ta phải nghiên cứu cách thức gọi các hệ thống dịch một cách tự động. Hệ thống Google Translator cung cấp phương thức gọi dưới dạng dịch vụ (web services), chúng ta cần thực hiện là xây dựng công cụ đọc lần lượt từng bộ phận của kho ngữ liệu và gửi đến hệ thống Google Translator thông qua lời gọi dịch vụ của Google Translator. Còn đối với các hệ thống dịch như Systran, Reverso,… chúng ta cần phải nghiên cứu các hàm API hay thư viện lập trình mà các hệ thống này cung cấp để xây dựng các chương trình gọi các hệ thống này thực thi tự động.

Ngôn ngữ 1	Ngôn ngữ 2	….	Ngôn ngữ N
Dữ liệu	Dữ liệu		Dữ liệu
Dữ liệu	Dữ liệu		Dữ liệu
….	….		….

Hình 2. Giải pháp mở rộng ngôn ngữ cho kho ngữ liệu

3. Thêm dữ liệu vào kho ngữ liệu

Tương tự giải pháp mở rộng khối lượng kho ngữ liệu bằng cách mở rộng ngôn ngữ, thêm dữ liệu vào kho ngữ bằng cách thu thập dữ liệu từ các nguồn dữ liệu khác không phải là ý tưởng mới và đã được nhiều người thực hiện. Tuy nhiên, vấn đề khó ở đây là một giải pháp tổng thể và tự động cho bất kỳ ngôn ngữ nào. Bởi lẽ để thực hiện được điều này chúng ta phải giải quyết được tất cả các bước như mô tả ở hình dưới đây của quá trình này một cách tổng thể, tự động cho tất cả các ngôn ngữ [10].



Hình 3. Giải pháp thêm dữ liệu vào kho ngữ liệu

Theo hình trên, các vấn đề khó của quá trình thu thập dữ liệu để xây dựng và bổ sung dữ liệu cho kho ngữ liệu bao gồm nhận dạng ngôn ngữ tự động từ nguồn tài nguyên hỗn tạp; xác định văn bản song song ở nhiều mức khác nhau như trang, đoạn, câu; tách câu, tách đoạn; liên kết câu hoặc đoạn. Các vấn đề này được xem là rất khó để thực hiện một cách tổng thể cho bất kỳ ngôn ngữ nào vì đặc điểm của mỗi ngôn ngữ [6].

B. Cải tiến chất lượng kho ngữ liệu

Như chúng tôi đã phân tích ở phần trên, chất lượng dịch của các hệ thống dịch tự động phụ thuộc rất lớn vào các kho ngữ liệu ở hai khía cạnh khối lượng và chất lượng. Trong phần này chúng tôi đề xuất các giải pháp nhằm cải tiến chất lượng kho ngữ liệu thông qua quá trình hậu xử lý (post-edit) và mở rộng ngữ nghĩa cho kho ngữ liệu.

1. Cải tiến thông qua quá trình hậu xử lý

Một kho ngữ liệu song ngữ có thể được xây dựng tự động bằng cách thu thập dữ liệu song song từ các nguồn tài nguyên khác như các website hoặc có thể được xây dựng bằng cách mở rộng ngôn ngữ thông qua quá trình dịch tự động. Vì thế chất lượng của các kho ngữ liệu thường rất thấp, để cải tiến chất lượng dữ liệu của các kho ngữ liệu cần phải có sự tham gia kiểm tra, chỉnh sửa của con người trên dữ liệu của kho ngữ liệu.

Do đó, vấn đề cần giải quyết ở đây đó là nghiên cứu xây dựng được một hệ thống hỗ trợ cho quá trình hậu xử lý. Hệ thống này cần cho phép nạp các kho ngữ liệu lớn và hiển thị dữ liệu một cách trực quan và khoa học sao cho dễ dàng cho người dùng kiểm tra và cải tiến dữ liệu. Ngoài ra, hệ thống này cần phải hoạt động như một môi trường cộng tác, cho phép nhiều người dùng tham gia cải tiến dữ liệu.

2. Mở rộng ngữ nghĩa

Hạn chế hiện tại của các kho ngữ liệu dùng trong dịch tự động không chỉ ở kích cỡ của kho ngữ liệu mà còn ở thông tin được làm giàu cho kho ngữ liệu. Các loại định dạng thông tin phổ biến được làm giàu cho kho ngữ liệu như hình ảnh, âm, các loại đồ thị,… chưa thật sự đầy đủ để giúp cho các hệ thống khai thác có thể sử dụng hiệu quả các kho ngữ liệu hiện tại. Do đó, vấn đề đặt ra là cần phải mở rộng các kho ngữ liệu hiện tại theo hướng ngữ nghĩa. Khi đó, kho ngữ liệu sẽ được mô tả đầy đủ thông tin hơn.

Việc mô tả thông tin cho kho ngữ liệu không chỉ dừng lại ở mức chung như hiện nay như mô tả thông tin bởi phần header của kho (như tên kho, ngôn ngữ, tác giả, kích thước, lĩnh vực,…), mà cần phải mở rộng đến thực thể của kho ngữ liệu như mỗi đoạn, mỗi câu và thậm chí mỗi cụm từ, mỗi từ đều được mô tả thông tin rõ ràng hơn. Hay nói cách khác, việc mở rộng kho ngữ liệu theo hướng ngữ nghĩa chính là việc xây dựng thêm một tầng ngữ nghĩa cho kho ngữ liệu. Khi đó, mỗi thực thể trong kho ngữ liệu được gắn kết với tầng ngữ nghĩa. Ở mức độ đơn giản, tầng ngữ nghĩa có thể bao gồm các chú thích, các từ/cụm từ đồng nghĩa, các từ/cụm từ trái nghĩa… Ở mức độ phức tạp, tầng ngữ nghĩa được xây dựng thành mạng lưới ontology, trong đó mỗi ontology gồm tập hợp các lớp thuộc một lĩnh vực hẹp nào đó, định nghĩa cụ thể hơn cho các thực thể của kho ngữ liệu [11].

Vấn đề đặt ra là làm cách nào để xây dựng tầng ngữ nghĩa cho các kho ngữ liệu một cách bán tự động, tức là xây dựng những chương trình có thể tự xác định các thực thể trong kho ngữ liệu thuộc các lớp được xây dựng sẵn, tự trích rút giá trị để xây dựng thuộc tính cho các lớp. Các bước xây dựng tầng ngữ nghĩa cho kho ngữ liệu có thể như sau:

Bước 1: Với mỗi kho ngữ liệu, định nghĩa các loại lớp dựa vào ngữ cảnh của kho (lĩnh vực của kho) và mối quan hệ giữa chúng.

Chẳng hạn, với kho ngữ liệu thuộc lĩnh vực y tế chúng ta sẽ có các lớp như Bác_sĩ, Bệnh_nhân, Thuốc,…

Bước 2: Xây dựng thuộc tính cho các lớp đã định nghĩa ở bước 1.

Bước 3: Với mỗi thực thể trong kho ngữ liệu, nhận biết thực thể thuộc lớp đã định nghĩa theo ngữ cảnh.

Ở bước này, công việc chính là thực hiện việc phân lớp từ, cụm từ. Ví dụ, đối với cụm từ "Hồ Chí Minh", tùy theo từng trường hợp mà nó có thể thuộc lớp Danh_nhân, lớp Người, lớp Thành_phố, lớp Đường_phố,…

Bước 4: Với mỗi thực thể đã xác định và phân loại theo lớp, tiến hành xây dựng thông tin cho thực thể đó dưới dạng gán giá trị cho các thuộc tính của các đối tượng thực thể đã xác định.

IV. KẾT QUẢ THỰC NGHIỆM

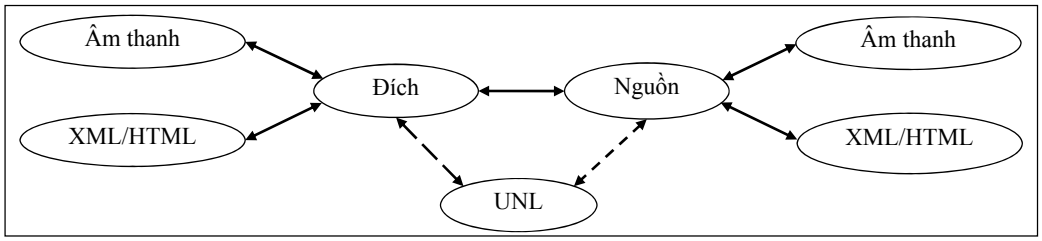
Từ các giải pháp được đề xuất như trình bày ở trên, chúng tôi đã tiến hành triển khai thực hiện và đã đạt được nhiều kết quả đáng kể.

Đối với giải pháp hợp nhất các kho ngữ liệu, chúng tôi đã tiến hành nghiên cứu và phân tích các kho ngữ liệu phổ biến đại diện cho các loại kho ngữ liệu phổ biến hiện nay gồm: JRC-ACQUIS, EUROPARL, ERIM và EOLSS/UNL. Từ việc phân tích bốn kho ngữ liệu trên, chúng tôi có được bảng tóm tắt đặc trưng của các kho ngữ liệu trên sau đây.

Bảng 1. Bảng tóm tắt đặc trưng các kho ngữ liệu phổ biến

Kho ngữ liệu	Tổ chức vật lý	Định dạng dữ liệu	Mức liên kết
JRC-ACQUIS	Gồm nhiều thư mục, mỗi thư mục chứa các tệp liên kết với nhau	XML theo chuẩn TEI	Tệp hoặc đoạn hoặc câu hoặc từ
EUROPARL		TXT	
ERIM		TXT, XML và WAV	
EOLSS/UNL		HTML và UNL	

Ở dạng trực quan, mối quan hệ giữa các dạng dữ liệu của các kho ngữ liệu có thể tóm tắt theo hình vẽ như sau:



Hình 4. Quan hệ giữa các dạng dữ liệu của các kho ngữ liệu

Từ việc phân tích các kho ngữ liệu trên, chúng tôi đã xác định được cấu trúc và định dạng chuẩn cho các kho ngữ liệu như sau: cấu trúc vật lý chuẩn của một kho ngữ liệu phải bao gồm nhiều thư mục và mỗi thư mục chứa đựng các tệp ở một định dạng nào đó. Trong đó, mỗi kho ngữ liệu phải chứa đựng các tệp mô tả ở mức độ khác nhau: mô tả

ở mức kho ngữ liệu, mô tả ở mức tài liệu và mô tả ở mức các phân đoạn bên trong. Ở mức độ chung nhất, chúng tôi đề xuất một kho ngữ liệu gồm hai phần:

Phần tiêu đề (header) chứa thông tin về ngữ liệu, ngôn ngữ, ngày tạo,…

Phần thân (body) chứa thông tin của các loại tài liệu: <doc>, <dialogue>,… Mỗi tài liệu chứa mô tả cấu trúc phân cấp của nó: chương, trang, mục,… và mô tả phân đoạn: (<seg>, <TP>, ,… Trong đó, mô tả đoạn chứa các thông tin: nguồn, bản dịch trước, bối cảnh, bài chỉnh sửa, âm thanh, điểm số, đồ thị UNL,…

Đối với định dạng chuẩn của kho ngữ liệu, chúng tôi sử dụng định dạng XML và đề xuất như sau:

```
<! ELEMENT corpus(header, body) >
<! ELEMENT header (name, date, domain, authors, project, Nlang, lang,
* othermeta *)>
<! ELEMENT name (# PCDATA)>
<! ELEMENT date (# PCDATA)>
<! ELEMENT domain (# PCDATA)>
<! ELEMENT authors (# PCDATA)>
<! ELEMENT project (# PCDATA)>
<! ELEMENT Nlang (# PCDATA)>
<! ATTLIST lang CDATA>
<! ELEMENT lang (# PCDATA)>
<! ATTLIST Othermeta CDATA>
<! MEMBER othermeta (# PCDATA)>
<! ELEMENT Othermeta  (# PCDATA)>
<! ELEMENT body (doc *) # REQUIRED>
<! ATTLIST doc CDATA>
<! ATTLIST doc id CDATA>
<! ATTLIST doc Nsegmments CDATA>
<! ATTLIST name CDATA doc>
<! ELEMENT doc (section *)>
<! ATTLIST article type CDATA>
<! ELEMENT section (segment *)>
<! ELEMENT section (segment *)>
<! ATTLIST segment id CDATA>
<! ATTLIST segment CDATA>
<! ELEMENT segment (case *)>
<! ATTLIST occurrence CDATA>
<! ATTLIST occurrence lang CDATA>
<! ATTLIST occurrence version CDATA>
<! ATTLIST occurrence producer CDATA>
<! ATTLIST occurrence level CDATA>
<! ATTLIST occurrence rating CDATA>
<! ATTLIST occurrence date CDATA>
<! ELEMENT occurrence (#PCDATA)>
```

Hình 5. Định dạng chuẩn biểu diễn kho ngữ liệu

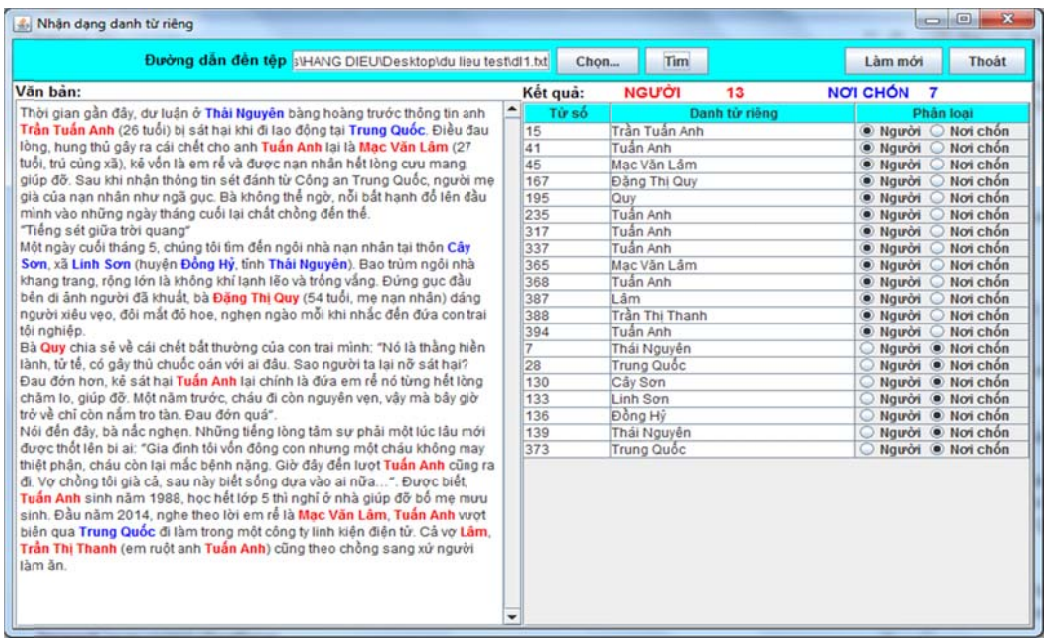
Sau khi có được định cấu trúc và dạng chuẩn cho các kho ngữ liệu, chúng tôi đã xây dựng công cụ và tiến hành chuyển đổi các kho ngữ liệu ở 4 dạng trên để xây dựng kho ngữ liệu đồng nhất về định dạng và cấu trúc.

Đối với giải pháp mở rộng ngôn ngữ, bổ sung và cải tiến dữ liệu cho các kho ngữ liệu, chúng tôi đã xây dựng được một môi trường làm việc cộng tác (collaborative environment) cho phép gọi các hệ thống dịch tự động để mở rộng ngôn ngữ cho kho ngữ liệu, thu thập dữ liệu song song từ các website đa ngữ cũng như cho phép cải tiến dữ liệu thông qua chức năng hậu xử lý.

Hình 6. Môi trường cho phép mở rộng kho ngữ liệu

Đối với giải pháp mở rộng kho ngữ liệu theo hướng ngữ nghĩa, như đã đề cập ở trên mặc dù đây là một hướng giải pháp rất hay, song cũng rất thách thức đối với quá trình thực nghiệm bởi lẽ hướng giải pháp này phải bao gồm nhiều công việc cần thực hiện. Để có thể thêm được tầng ngữ nghĩa cho kho ngữ liệu thì từng thực thể dữ liệu cần được nhận dạng và làm rõ nghĩa hơn. Thực thể dữ liệu của kho ngữ liệu bao gồm nhiều loại, trong đó ở giai đoạn này chúng tôi tập trung đối với danh từ riêng. Thực tế, danh từ riêng rất đa dạng và thường nhập nhằng, cùng một danh từ riêng có thể khác nhau với ngữ cảnh khác nhau. Vấn đề cốt lõi và quan trọng nhất đối với công việc này là nhận dạng và trích rút danh từ riêng để sau đó tiến hành thêm tầng ngữ nghĩa.

Chúng tôi đã đề xuất giải pháp kết hợp giữa thuật toán Maximum Matching và phân tích mối quan hệ giữa các thành tố văn bản, cụ thể là quan hệ của thực thể cần kiểm tra với các thực thể tiền tố và hậu tố của nó. Sau khi có danh sách các danh từ riêng, chúng ta tiến hành nhận biết các danh từ riêng đó thuộc lớp danh từ riêng nào. Việc nhận biết, phân loại danh từ riêng này dựa vào quan hệ giữa các thực thể trong văn bản và so khớp các thực thể tiền tố và hậu tố với tập hợp các từ ngữ cảnh nhằm chỉ địa danh hoặc tên người. Để nâng cao hơn nữa hiệu quả của giải pháp này, hệ thống cho phép người dùng hiệu chỉnh kết quả nhận diện bằng tay. Hệ thống sẽ hiển thị danh sách các từ, cụm từ đã được nhận diện để người dùng có thể xác nhận, chỉnh sửa,…



Hình 7. Hệ thống nhận dạng và phân loại thực thể danh từ riêng từ kho ngữ liệu

V. KẾT LUẬN

Cải tiến chất lượng dịch cho các hệ thống dịch tự động, đặc biệt dịch tự động tiếng Việt và tiếng các dân tộc thiểu số là vấn đề rất cần thiết hiện nay. Trong số nhiều giải pháp cải tiến chất lượng dịch, giải pháp mở rộng kho ngữ liệu theo cả hai khía cạnh khối lượng và cải tiến chất lượng là một giải pháp hiệu quả đã được chứng minh. Bài báo này đã đề xuất nhiều giải pháp hiệu quả nhằm cho phép tạo ra được kho ngữ liệu không chỉ lớn về kích thước, đa dạng về ngôn ngữ mà còn tốt về chất lượng. Tất cả các giải pháp đã được nhóm tác giả tiến hành triển khai thực hiện rất công phu và đã thu được những kết quả đáng kể. Mặc dù vậy, một số giải pháp như mở rộng ngữ nghĩa đối với kho ngữ liệu là một giải pháp rất thách thức cho vấn đề triển khai thực hiện và mang tính dài hơi. Do đó, trong thời gian đến nhóm tác giả sẽ tiếp tục nghiên cứu và tập trung vào hướng mở rộng ngữ nghĩa cho kho ngữ liệu đối với nhiều loại thực thể dữ liệu khác. Ngoài ra, giải pháp hợp nhất các kho ngữ liệu cũng cần được quan tâm giải quyết đối với nhiều loại kho ngữ liệu khác nhau.

VI. TÀI LIỆU THAM KHẢO

[1] Koehn Ph. (2005), Europarl: A Parallel Corpus for Statistical Machine Translation, In Proc. of the 10th Machine Translation Summit, Phuket, Thailand, pp. 79–86.

[2] Munteanu D.S., Marcu D. (2006), Extracting parallel sub-sentential fragments from non-parallel corpora. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, pp. 81-88.

[3] Blanchon H., and Boitet C. (2007), Pour l'évaluation externe des systèmes de TA par des méthodes externes fondées sur la tâche. TAL, vol. 48, 32 p.

[4] Boitet C. (2007), Corpus pour la TA: types, tailles, et problèmes associés, selon leur usage et le type de système. Revue française de linguistique appliquée. Vol. XII –2007, pp. 25-38.

[5] J. Brunning (2010), Alignment Models and Algorithms for Statistical Machine Translation, Ph.D. Thesis. Cambridge University, 191 p.

[6] Huynh C-P. (2010), Des suites de test pour la TA à un système d'exploitation de corpus alignés de documents et métadocuments multilingues, multiannotés et multimedia, PhD thesis-National Polytechnic Institute of Grenoble, 228 p.

[7] Huynh C-P (2011), New approach for collecting high quality parallel corpora from multilingual Websites. iiWAS11 Conference. Proceedings of the 13th International Conference on Information Integration and Web-based Applications & Services.

[8] R. Nazar (2011), Parallel corpus alignment at the document, sentence and vocabulary levels, Procesamiento del Lenguaje Natural 47, p 129-136.

[9] Huỳnh Công Pháp, Đặng Đại Thọ (2012), Giải pháp chuẩn hóa các kho ngữ liệu dùng trong lĩnh vực dịch tự động, Tạp chí Khoa học công nghệ - Đại học Đà Nẵng. Số 9(58).2012 Quyển 3, Trang: 111-117.

[10] Huynh C-P (2012), Nghiên cứu và xây dựng hệ một hệ thống hỗ trợ khai thác dữ liệu dịch tự động, Đề tài nghiên cứu khoa học cấp Đại học Đà Nẵng.

[11] Đặng Đại Thọ, Huỳnh Công Pháp (2013), Mở rộng kho ngữ liệu theo hướng ngữ nghĩa, Tạp chí Khoa học & Công nghệ Đại học Đà Nẵng, Số: 12(73).2013 Quyển 2, Trang: 110-116.

[12] Maheshwari S., Himanshu S. (2014): Corpus quality improvements for statistical machine translation, IJAICT Volume 1, Issue 3, pp. 351-353.

IMPROVING QUALITY OF MACHINE TRANSLATION BASED ON SOLUTIONS OF EXTENDING MT CORPORA

Huynh Cong Phap, Dang Dai Tho, Nguyen Van Binh

ABSTRACT *- Improving machine translation (MT) quality, especially under resourced languages such as Vietnamese, ethnic minority languages, is an important problem to be addressed. Beside many proposed solutions like enhencing automatic translation algorithms and methods, the solution of improving machine translation quality based on extending machine translation corpora at quantitive and quality aspects is the very efficient one which has been proved [7]. Therefore, in this paper we propose efficient and crucial solutions allowing to buid machine translation corpora with big quantitive, numberous languages, and better quality. All solutions have been implemented and we have achieved convincing results.*