MỞ RỘNG KHO NGỮ LIỆU DỊCH TỰ ĐỘNG THEO HƯỚNG NGỮ NGHĨA

SEMANTIC ORIENTED EXTENSION FOR MACHINE TRANSLATION CORPORA

Đặng Đại Thọ, Huỳnh Công Pháp

Trường Cao đẳng Công nghệ Thông tin, Đại học Đà Nẵng

Email: ddtho.dt@gmail.com, hcphap@gmail.com

TÓM TẮT

Kho ngữ liệu là tài nguyên ngôn ngữ căn bản và rất cần thiết để phát triển và cải tiến các hệ thống dịch tự động. Hiện nay đã tồn tại rất nhiều kho ngữ liệu dùng cho dịch tự động. Tuy nhiên, việc khai thác chúng còn rất nhiều hạn chế. Nguyên nhân là các kho ngữ liệu dịch tự động hiện nay chủ yếu tồn tại dưới dạng văn bản hoặc có liên kết các dạng dữ liệu khác như âm thanh, hình ảnh, đồ thị,… mà chưa được tổ chức ở dạng ngữ nghĩa. Vì thế, trong bài báo này, chúng tôi đề xuất mở rộng kho ngữ liệu dịch tự động theo hướng ngữ nghĩa bằng cách thêm tầng ngữ nghĩa vào các kho ngữ liệu hiện tại nhằm nâng cao hiệu quả của các hệ thống khai thác dịch tự động hiện nay.

Từ khóa: kho ngữ liệu; dịch tự động; ngữ nghĩa; hệ thống khai thác; mở rộng kho ngữ liệu

ABSTRACT

Corpora play a crucial role in the development and improvement of automatic translation systems. There are currently many corpora used in the machine translation (MT) domain. However, exploiting and using these corpora are still challenging and limited because of some reasons, of which the main reason is that most corpora are in terms of raw texts or linked to other different kinds of data such as audio, images, graphs.... But they are not organized into semantic layers. Therefore, in this paper, we want to propose an idea of extending and enlarging corpora by adding to them a semantic layer so that the performance of corpus exploitation systems will be much improved.

Key words: corpus; machine translation; semantic; exploitation system; corpus extension

1. Đặt vấn đề

Dữ liệu dịch tự động, còn gọi là kho ngữ liệu (corpus), là tài nguyên ngôn ngữ căn bản và rất cần thiết để phát triển và cải tiến các hệ thống dịch tự động.

Hiện nay có nhiều phương pháp phát triển các hệ thống dịch tự động: dịch theo kinh nghiệm, dịch thống kê, dịch dựa vào tập mẫu, dịch chuyên gia,… Trong đó, mỗi loại hệ thống dịch tự động sử dụng một loại kho ngữ kho liệu khác nhau. Chẳng hạn, loại hệ thống dịch tự động thống kê sử dụng các kho ngữ liệu rất lớn, liên kết ở mức từ (word alignment); loại hệ thống dịch dựa vào tập mẫu sử dụng kho ngữ liệu có liên kết ở mức câu (sentence alignment) hoặc mức đoạn (paragraph alignement); loại hệ thống dịch chuyên gia sử dụng kho ngữ liệu được làm giàu bởi nhiều loại thông tin khác nhau như âm thanh, ngôn ngữ trung gian (như IF, UNL,…) hoặc hình ảnh [2].

Bất cứ hệ thống dịch tự động hay hệ thống khai thác kho ngữ liệu thuộc loại nào cũng đều có một quá trình tìm kiếm và so khớp "phần tử" trong kho ngữ liệu với đầu vào của nó để suy luận và sinh ra đầu ra tương ứng. Chẳng hạn, các hệ thống dịch tự động dựa vào tập mẫu sử dụng kho ngữ liệu song song.Với mỗi câu đầu vào hệ thống đều tìm kiếm và so khớp với tập dữ liệu nguồn trong kho ngữ liệu để tìm ra câu ngôn ngữ đích liên kết với câu nguồn mà khớp với đầu vào của hệ thống. Tương tự như vậy, các hệ tìm kiếm, hệ hỏi đáp, từ điển… đều phải bao hàm quá trình này. Điều đó cho thấy quá trình so khớp đầu vào với kho ngữ liệu (cơ sở dữ liệu) của hệ thống khai thác rất quan trọng, quyết định hiệu quả và tính thông minh của một hệ thống.

Chính vì vậy mà ngày nay, trong các hệ thống khai thác kho ngữ liệu người ta đã nghiên cứu, xây dựng nhiều thuật toán tìm kiếm, so khớp thông minh giữa đầu vào, dữ liệu trong kho ngữ liệu của hệ thống. Tuy nhiên, gần như tất cả

các hệ thống hiện nay đều dừng lại ở mức so khớp dạng chuỗi ký tự hoặc dạng văn bản bằng các thuật toán như tính khoảng cách, tính xác suất, tính tần suất ký tự… Điều này đã hạn chế rất nhiều việc khai thác hiệu quả các kho ngữ liệu. Nguyên nhân là các kho ngữ liệu hiện nay chủ yếu tồn tại dưới dạng văn bản hoặc có liên kết các dạng dữ liệu khác như âm thanh, hình ảnh, đồ thị,… mà chưa được tổ chức ở dạng ngữ nghĩa.

Trong bài báo này, chúng tôi đề xuất mở rộng kho ngữ liệu dịch tự động theo hướng ngữ nghĩa bằng cách thêm tầng ngữ nghĩa vào các kho ngữ liệu hiện tại. Tầng ngữ nghĩa có thể đơn giản chỉ là tầng dữ liệu mô tả thêm dữ liệu nguồn của kho ngữ liệu 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 của ontology nào đó.

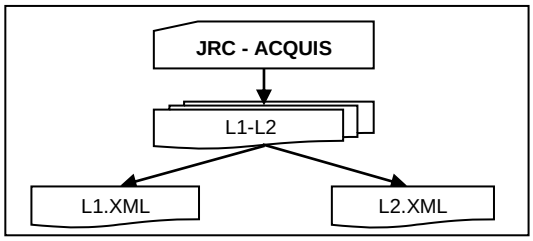
Để có cái nhìn rõ hơn về thực trạng các kho ngữ liệu hiện nay, phần đầu của bài báo sẽ giới thiệu một số kho ngữ liệu phổ biến dùng trong dịch tự động hiện nay, tiếp theo sẽ giới thiệu một số dạng đơn giản của ngữ nghĩa đã được định nghĩa trong các kho ngữ liệu và phần cuối cùng là đề xuất thêm tầng ngữ nghĩa vào các kho ngữ liệu hiện tại.

2. Các kho ngữ liệu phổ biến

Dưới đây là một số kho ngữ liệu dịch tự động phổ biến. Mặc dù các kho ngữ liệu này đã được làm giàu thông tin ở dạng khác văn bảnnhưng đều chưa được tổ chức theo dạng ngữ nghĩa 7.[1].

2.1. Kho ngữ liệu JRC-ACQUIS

JRC-ACQUIS là kho ngữ liệu song song đa ngôn ngữ, được xây dựng từ các văn bản pháp lý của Liên minh châu Âu. Phiên bản hiện tại là 3.0 gồm 22 ngôn ngữ với khoảng 23.000 tài liệu cho mỗi ngôn ngữ. Kho ngữ liệu này được cấu trúc gồm nhiều thư mục chứa các cặp ngôn ngữ được liên kết với nhau. Mỗi thư mục gồm các tệp ở dạng XML, mỗi tệp được nhóm theo ngôn ngữ, theo định dạng TEI. Trong đó, mỗi tệp XML theo định dạng TEI chứa tiêu đề cho biết thông tin về ngôn ngữ và các tài liệu thực, thông tin URL chỉ nguồn gốc dữ liệu. Cấu trúc vật lý của kho ngữ liệu này được mô tả như sau:

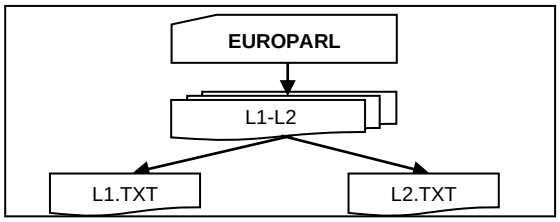


Hình 1. *Cấu trúc ngữ liệu JRC-ACQUIS*

Kho ngữ liệu JRC-ACQUIS được liên kết ở mức đoạn giữa các cặp ngôn ngữ, các đoạn rất ngắn, thường chứa một câu, thậm chí một phần của câu.

2.2. Kho ngữ liệu EUROPARL

EUROPARL là một trong những kho ngữ liệu song song phổ biến hiện nay, được xây dựng nhằm phục vụ cho việc nghiên cứu và phát triển các hệ thống dịch tự động. Kho ngữ liệu song song này được xây dựng từ các bài phát biểu của các cuộc họp Quốc hội châu Âu, gồm 11 ngôn ngữ chính thức của các nước thành viên của Liên minh châu Âu. Phiên bản hiện tại là 5.0, gồm hơn 50 triệu từ cho mỗi ngôn ngữ [3].



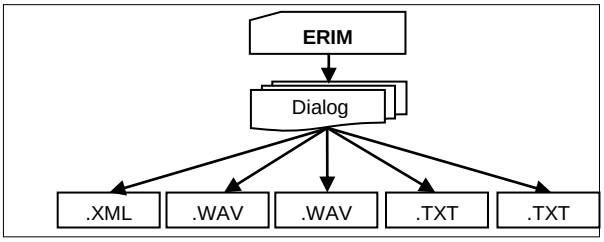
Hình 2. *Cấu trúc ngữ liệu EUROPARL*

Cấu trúc vật lý (Hình 2) của kho ngữ liệu này tương tự như kho ngữ liệu JRC-ACQUIS, gồm nhiều thư mục chứa đựng các cặp ngôn ngữ được liên kết với nhau. Tuy nhiên, mỗi thư mục gồm các tệp ở dạng TXT có cấu trúc gồm nhiều đoạn có liên kết với nhau. Kho ngữ liệu này được liên kết ở mức đoạn, trong đó tiếng Anh được xem như ngôn ngữ gốc liên kết với 10 ngôn ngữ còn lại. Việc liên kết được thực hiện bởi thuật toán Church and Gale.

2.3. Kho ngữ liệu ERIM

Kho ngữ liệu ERIM được xây dựng từ dự án ERIM nhằm phát triển môi trường cho phép

phiên dịch thông qua intranet hay extranet [1]. Môi trường này hỗ trợ nhiều phương tiện giao tiếp khác nhau như âm thanh, văn bản và hình ảnh. Đến nay, kho ngữ liệu ERIM đã có khoảng 600 phút hội thoại Pháp - Trung Quốc, 630 phút Pháp - Việt, 150 phút Pháp - Hindu và 540 phút Pháp - Tamil.

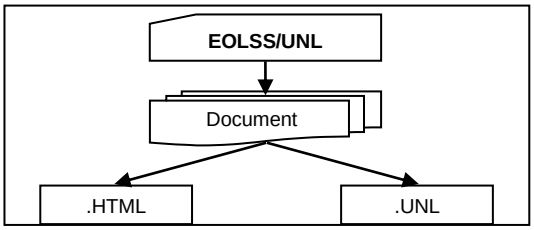


Hình 3. *Cấu trúc ngữ liệu ERIM*

Tương tự như hai kho ngữ liệu trên, cấu trúc vật lý của kho ngữ liệu ERIM (Hình 3) gồm nhiều thư mục. Mỗi thư mục chứa nhiều tệp tin ở định dạng khác nhau như TXT, XML, WAV (âm thanh). Mỗi thư mục biểu diễn một đoạn hội thoại, mỗi tệp. WAV chứa âm thanh của một câu thoại, mỗi tệp. TXT chứa đựng câu thoại ở dạng văn bản và mỗi. XML mô tả thông tin về câu thoại như độ dài, người nói,…

2.4. Kho ngữ liệu EOLSS/UNL

Kho ngữ liệu EOLSS/UNL gồm có 6600 bài viết (khoảng 250.000 trang) trong 6 ngôn ngữ thuộc UNESCO là tiếng Anh, tiếng Pháp, tiếng Arập, tiếng Nhật, tiếng Tây Ban Nha và tiếng Nga [1].



Hình 4. *Cấu trúc ngữ liệu EOLSS/UNL*

Cấu trúc vật lý của kho ngữ liệu này (Hình 4) cũng tương tự như các kho ngữ liệu phân tích ở trên gồm nhiều thư mục, mỗi thư mục biểu diễn một tài liệu ở định dạng HTML và UNL. Mỗi đoạn trong tệp. HTML được liên kết với một đoạn trong tệp. UNL tương ứng.

3. Các loại định dạng dữ liệu được làm giàu trong các kho ngữ liệu

Như trình bày ở phần trên, mặc dù các kho ngữ liệu có cấu trúc và định dạng khác nhau nhưng chúng ta có thể phân loại các kho ngữ liệu theo 2 loại, dựa vào mức độ thông tin được làm giàu đối với kho ngữ liệu, đó là: kho ngữ liệu thô (kho ngữ liệu văn bản) và kho ngữ liệu đã được làm giàu.

Đối với các kho ngữ liệu được làm giàu, chúng ta có thể tìm thấy các loại dữ liệu được làm giàu phổ biến như sau:

3.1. Gán nhãn từ loại

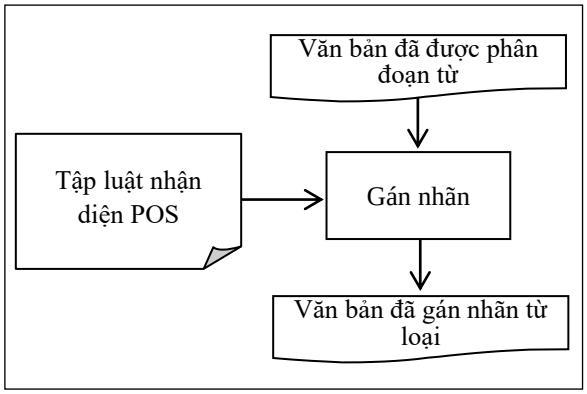
Một trong những phương pháp khai thác hiệu quả kho ngữ liệu là phân tích ngôn ngữ của kho ngữ liệu bằng cách phân loại các từ thành các lớp từ loại dựa vào ngữ cảnh của từ trong kho ngữ liệu. Mỗi từ loại tương ứng với một hình thái và một vai trò ngữ pháp nhất định. Để thể hiện chức năng ngữ pháp của mỗi từ, người ta sử dụng nhãn từ loại: danh từ, tính từ, động từ,… Ví dụ trong câu "I want to book a book", từ "book" có hai nhãn từ loại là động từ và danh từ. Công việc gán nhãn từ loại cho một văn bản là xác định từ loại của mỗi từ trong phạm vi văn bản đó. Danh sách các từ loại có thể có của một ngôn ngữ được gọi là bộ nhãn từ loại (POS-tagset) của ngôn ngữ đó [9].

Câu "Explosives found on Hampstead Heath" được lưu trữ trong kho ngữ liệu BNC corpus như sau:

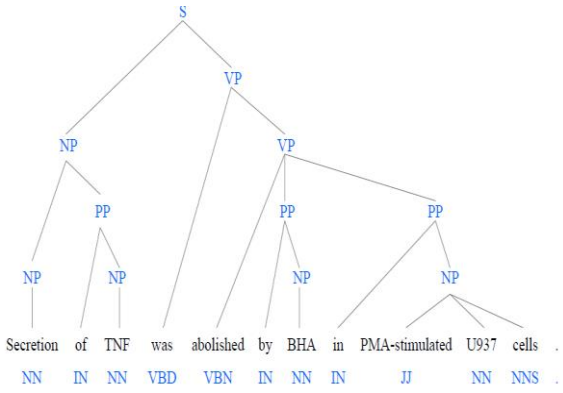
```
<s>
<w NN2>Explosives
<w VVD>found
<w PRP>on
<w NP0>Hampstead
<w NP0>Health
</PUN>
</s>
```

Trong đó s là câu, w là từ, NN2 là danh từ số nhiều, VVD là động từ ở thì quá khứ, PRP là giới từ, NP0 là danh từ riêng, PUN là dấu chấm câu [6].

Mô hình gán nhãn từ loại như Hình 5.



Hình 5. *Mô hình gán nhãn từ loại*



Hình 6. *Gán nhãn cây cú pháp*

3.2. Gán nhãn ranh giới ngữ

Một phương pháp khác liên quan đến làm giàu thông tin cho kho ngữ liệu đó là gán nhãn ranh giới ngữ, được thực hiện sau khi gán nhãn chú thích từ loại. Nó mô tả các mối quan hệ cú pháp giữa các đơn vị từ vựng và cấu trúc cú pháp khác nhau: cụm danh từ, cụm động từ, cụm tính từ,…[9].

Chẳng hạn, câu "Corpus annotation is the practice of adding interpretative linguistic information to a corpus" được gán nhãn ranh giới ngữ như sau:

```
[NP (NN Corpus) (NN annotation) ]
(VBZ is)
[NP (DT the) (NN practice) ]
(IN of) (VBG adding)
[NP (JJ interpretative) (JJ linguistic) (NN information) ]
[PP (TO to) [NP (DT a) (NN corpus) ]
```

Trong đó S là câu, NP là cụm danh từ, VP là cụm động từ, ADJP là cụm tính từ [7].

3.3. Gán nhãn cây cú pháp

Gán nhãn cây cú pháp nhằm mục đích phân tích một câu thành những thành phần văn phạm có liên quan với nhau và được thể hiện thành cây cú pháp [5].

4. Hạn chế của kho ngữ liệu

Như trình bày ở trên, mặc dù các kho ngữ liệu hiện tại cũng đã được làm giàu bằng những định dạng dữ liệu khác nhau như hình ảnh, âm thanh, đồ thị… và thậm chí các đơn vị từ của kho ngữ liệu cũng đã được gán nhãn từ loại hoặc ranh giới ngữ… Tuy nhiên, thông tin được làm giàu cho các kho ngữ liệu vẫn chưa thật sự đầy đủ để có thể cho phép khai thác hiệu quả các kho ngữ liệu này. Các hạn chế của các kho ngữ liệu hiện tại có thể thấy như sau:

4.1. Hạn chế về mặt ngữ nghĩa

Vấn đề ngữ nghĩa của các kho ngữ liệu còn ở mức độ đơn giản, đó là các khối chú giải thông tin. Chú giải là phần giải thích các thông tin đặc thù làm rõ nghĩa cho các văn bản trong kho ngữ liệu như là chú giải bên ngoài ngôn ngữ (ví dụ, chú giải về tác giả: tên, tuổi, giới tính, năm sinh… và về văn bản: tác giả, tên văn bản, năm và nơi xuất bản, thể loại, phong cách ngôn ngữ…); hoặc là chú giải cấu trúc (ví dụ, chương, đoạn, câu, hình thái từ…); hoặc là chú giải cho chính ngôn ngữ văn bản về từ vựng, cú pháp.

Thực tế hiện nay các kho ngữ liệu chưa giúp cho việc giải quyết nhập nhằng ngữ nghĩa hiệu quả.

Nhận diện ranh giới từ đối với các ngôn ngữ biến hình (tiếng Pháp, tiếng Nga, tiếng Anh) trong các kho ngữ liệu hiện nay đã giải quyết tốt. Tuy nhiên, đối với các ngôn ngữ đơn lập (tiếng Việt, tiếng Hoa, tiếng Lào,…) đến nay vẫn còn rất nhiều hạn chế. Nguyên nhân là đối với các ngôn ngữ biến hình, ranh giới từ được xác định

chủ yếu dựa vào khoảng trắng hay dấu câu. Còn trong các ngôn ngữ đơn lập, từ vựng chủ yếu là các từ ghép vì thế khoảng trắng không phải luôn luôn là ranh giới chính xác [9].

Chẳng hạn, trong tiếng Anh, câu "He is a doctor" được phân định ranh giới dễ dàng là "He / is / a /doctor". Còn câu tương ứng trong tiếng Việt là "Anh ấy là bác sĩ" nếu phân định ranh giới từ theo khoảng trắng trở thành "Anh / ấy / là / bác / sĩ". Ở đây, "bác sĩ" là từ ghép nay bị chia thành hai từ đơn là "bác" và "sĩ", cách phân định này là sai. Cách phân định ranh giới đúng của câu trên phải là "Anh ấy / là / bác sĩ".

Từ loại là một yếu tố quan trọng trong việc xác định nghĩa chính xác của từ và sắp xếp các từ thành câu hoàn chỉnh trong dịch tự động. Cho đến nay, đối với các ngôn ngữ đơn lập, việc xác định từ loại còn gặp rất nhiều khó khăn. Trong đó, việc nhập nhằng ranh giới từ cũng góp phần gây ra sự nhập nhằng từ loại.

Ví dụ, câu tiếng Việt "Ông già đi nhanh quá!" nếu được phân định ranh giới là "/Ông/ già đi /nhanh / quá !" sẽ có nghĩa tiếng Anh tương ứng là "The man becomes old so fast !". Còn nếu được phân định ranh giới là "Ông già / đi / nhanh /quá !" thì tương ứng là "The old man goes so fast !".

Từ trên, chúng ta thấy với các chú giải ngữ nghĩa của các kho ngữ liệu hiện nay, các hệ thống khai thác chưa thể giải quyết hiệu quả vấn đề nhập nhằng về ranh giới từ và từ loại.

Bất cứ ngôn ngữ nào cũng có từ đa nghĩa. Chẳng hạn trong tiếng Việt, từ "ăn" trong các câu "Tôi đi ăn sáng", "Nó đi ăn cướp", "Phanh không ăn", "Một đô-la Mỹ ăn 20 ngàn đồng Việt Nam", "Tàu thủy ăn hàng" vừa có những nét nghĩa giống và khác nhau. Với các kho ngữ liệu hiện nay, các hệ thống khai thác rất khó dịch từ đa nghĩa theo nghĩa nào trong nhóm nghĩa của nó. Bởi vì việc chọn lựa nghĩa phù hợp trong câu là một vấn đề khó khăn, cần phải hiểu được mối quan hệ của từ với ngữ cảnh xung quanh để nhận biết nghĩa chính xác của từ.

Ngoài ra, việc nhập nhằng ngữ nghĩa còn ở mức cấu trúc, mức liên câu và mức văn bản.

4.2. Hạn chế của các hệ thống chú giải ngữ nghĩa [9]

Chúng ta có thể thấy, mỗi từ có thể mang nhiều nghĩa khác nhau, nhưng trong một ngữ cảnh cụ thể thì nó mang một nghĩa nhất định nào đó. Chẳng hạn trong tiếng Anh, danh từ "bank" có thể là "ngân hàng", hoặc "bờ sông", hoặc "dãy". Để dễ phân biệt nghĩa các từ vựng khác nhau, người ta tiến hành gán nhãn ngữ nghĩa của tất cả các từ trong kho ngữ liệu. Có nghĩa là phân chia toàn bộ ý nghĩa từ vựng thành hệ thống các ý niệm. Chẳng hạn, với danh từ "bank" nói trên, các nghĩa tương ứng của chúng sẽ là "ngân hàng" thuộc về ý niệm "công trình xây dựng nhân tạo"; "bờ sông" thuộc về ý niệm "công trình thiên tạo"; "dãy" thuộc về ý niệm "sự sắp xếp tổ chức".

Tuy nhiên, cho đến nay chưa có một hệ thống nhãn ngữ nghĩa nào giúp cho việc giải quyết nhập nhằng ngữ nghĩa của từ một cách ổn thỏa. Cụ thể như sau:

Hệ thống LLOCE sắp xếp các mục từ thành các chủ đề, mỗi chủ đề được chia thành nhiều nhóm, mỗi nhóm được chia thành nhiều lớp, mỗi lớp gồm các mục từ có quan hệ ngữ nghĩa với nhau (đồng nghĩa, gần nghĩa,…). Hệ thống này chỉ gồm 3 cấp nên giữa các lớp khó tìm mối quan hệ với nhau.

Hệ thống LDOCE chỉ chú trọng đến danh từ. Bên cạnh đó nó phân chia lớp ngữ nghĩa quá thô (chỉ 32 lớp) nên không thể khử nhập nhằng cho các từ cùng lớp nhưng khác nghĩa.

Hệ thống WordNet là một hệ thống các ý niệm có quan hệ nhiều mặt với nhau, tạo thành một mạng lưới phức tạp. Nó phân cấp chi tiết và giữa các lớp còn có nhiều kiểu quan hệ khác nhau. Tuy vậy, nó không có sự phân biệt về ngguyên tắc giữa từ đồng nghĩa và đa nghĩa.

Hệ thống CoreLex được xây dựng từ các lớp cơ bản của WordNet. Tuy nhiên, nó chỉ dành cho danh từ mà thôi.

5. Đề xuất theo hướng ngữ nghĩa

Như trình bày ở phần trên, hạn chế hiện tại của các kho ngữ liệu dùng trong dịch tự động có thể thấy không chỉ ở kích cỡ của kho ngữ liệu mà chính là 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 thanh, các loại đồ thị,… có vẻ như 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 cần 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 đó là mỗi kho ngữ liệu được mô tả thông tin bởi phần header của kho (như tên kho, ngôn ngữ, tác giả, kích thước, lĩnh vực,…) mà 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.

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 nghĩacác 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: Vơi 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.

6. Bàn luận

Các kho ngữ liệu dùng trong dịch tự động hiện tại có kích thước tương đối lớn và đã ít nhiều được làm giàu bởi một số định dạng thông tin khác nhau. Tuy nhiên, gần như chưa có một kho ngữ liệu nào được làm giàu hay mở rộng theo hướng ngữ nghĩa. Điều này gây nên hạn chế rất lớn đối với việc khai thác hiệu quả các kho ngữ liệu hiện nay. Các kho ngữ liệu hiện tại chỉ cho phép các hệ thống tìm kiếm và so khớp dựa vào các thuật toán so sánh chuỗi. Vấn đề mà bài báo đề cập là cần mở rộng các kho ngữ liệu theo hướng ngữ nghĩa nhằm cho phép các hệ thống tìm kiếm và so khớp hiệu quả và chính xác hơn. Tuy nhiên, một vấn đề phát sinh là kích thước của kho ngữ liệu sẽ tăng lên đáng kể nếu kho ngữ liệu được thêm một tầng ngữ nghĩa. Vấn đề này cũng sẽ kéo theo tốc độ xử lý và tìm kiếm của các hệ thống bị ảnh hưởng rất lớn, do đó cần phải xây dựng những thuật toán tối ưu nhằm tăng tốc độ so khớp và tìm kiếm cho các hệ thống.

7. Kết luận

Các kho ngữ liệu có vai trò quan trọng trong lĩnh vực xử lý ngôn ngữ tự nhiên và dịch tự động. Do đó, hiện nay tồn tại rất nhiều kho ngữ liệu được xây dựng bởi các nhà phát triển và tổ chức khác nhau. Tuy nhiên, các kho ngữ liệu này lại có cấu trúc và định dạng khác nhau, đa số chỉ tồn tại dưới dạng văn bản hoặc chỉ có liên kết với một số định dạng dữ liệu cơ bản. Chính vì thế, việc khai thác và sử dụng các kho ngữ liệu này chưa thật sự hiệu quả và gặp không ít khó khăn. Để khai thác và sử dụng các kho ngữ liệu này một cách hiệu quả, chúng tôi đã đề xuất ý tưởng mở rộng các kho ngữ liệu theo hướng ngữ nghĩa ở nhiều cấu độ khác nhau: ở cấp độ đơn

giản, ngữ nghĩa được xây dựng có thể bao gồm các chú thích, các từ/ cụm từ đồng nghĩa, các từ/ cụm từ trái nghĩa; cấp độ phức tạp, tầng ngữ nghĩa được xây dựng dưới dạng 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. Bài báo chỉ dừng lại ở mức đề xuất ý tưởng, việc triển khai và thực nghiệm ý tưởng này chắc chắn được chúng tôi thực hiện trong thời gian tới. Ý tưởng này còn mở ra một hướng nghiên cứu tiềm năng về việc khai phá dữ liệu từ các kho ngữ liệu.

TÀI LIỆU THAM KHẢO

[1] 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 multimédia*, PhD thesis-National Polytechnic Institute of Grenoble, 228 p.

[2] 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.

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

[4] Mosleh H. A., Tang E. K. (1999), *Example-Based Machine Translation Based on the Synchronous SSTC Annotation Schema*, Procceding of the Machine Translation Summit VII. Singapore, pp. 244-249.

[5] KimJ-D. (2003), *The GENIA corpus – Linguistic and Semantic Annotation of Biomedical Literature, Tsujii Laboratory*, University of Tokyo.

[6] McEnery T. and Wilson A. (2001), *Corpus Linguistics*, Edinburgh University Press.

[7] Matthew B-O.*Corpus Mark-up, http://www.lexically.net/courses/sessions/markup/Corpus%20Mark-up.ppt*

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

[9] *Tổng quan về xử lý ngôn ngữ tự nhiên trong dịch máy*, http://www.mediafire.com/?thwbuuub32yq4zu

(BBT nhận bài: 07/10/2013, phản biện xong: 22/10/2013)