CHIA TÁCH TỪ VÀ DÁN NHÃN TỪ: ỨNG DỤNG CHO BỘ DỮ LIỆU CHUYÊN NGÀNH KHOA HỌC KĨ THUẬT TIẾNG VIỆT

TS. LÊ CHÍ HIẾU*

Tiếng Việt là ngôn ngữ của hơn 90 triệu dân, tuy nhiên, công việc xử lí ngôn ngữ tự động (TAL) hiện nay rất sơ sài trong cộng đồng ngôn ngữ nói chung và trong ngôn ngữ chuyên nghành (LSP) nói riêng. Những vấn đề cơ bản về phân tích tự động ngôn ngữ tiếng Việt (như là phân tích cú pháp (annotation), chia tách từ, dán nhãn, chia tách câu, nhóm từ) luôn là vấn đề gai góc, phức tạp do thiếu các dữ liệu ngôn ngữ và đặc tính của tiếng Việt. Trong bài nghiên cứu này, chúng tôi tập trung vào công việc phân tách và dán nhãn cho bộ dữ liệu chuyên ngành khoa học kĩ thuật. Đầu tiên, chúng tôi giới thiệu sơ qua về ngôn ngữ tiếng Việt, các đặc điểm từ vựng ngữ pháp, mô tả từ vựng, cú pháp; sau đó, chúng tôi giới thiệu các phương pháp chia tách và các công cụ dán nhãn. Trong trường hợp này, chúng tôi đã sử dụng các công cụ phát triển dựa trên nền phần mềm Qtag (Mason và Tufis-1998) và minh hoạ phương pháp bằng cách phân tích một câu trong bộ dữ liệu bằng cách sử dụng các công cụ phân tách và dán nhãn VnQTAG và Vntoolkit.

1. Vài nét về tiếng Việt

Tiếng Việt thuộc ngôn ngữ Việt - Mường, một nhánh của Môn-Khơme. Đây là cách sắp xếp do Andre-Goerges Haudricourt đề xuất. Điều này có nghĩa là, tiếng Việt bị ảnh hưởng ít nhiều bởi các ngôn ngữ thuộc nhóm Việt - Mường. Ví dụ: các số đếm xuất phát từ ngôn ngữ Môn-Khơme: một, hai, ba, v.v... Sau đó, dưới thời kì đô hộ của phong kiến phương Bắc, rất nhiều từ Hán Việt xuất hiện do có yếu tố ngôn ngữ Hán có trong hệ thống từ vựng tiếng Việt. Sang thế kỉ XVII, nhà truyền giáo người Pháp Alexandre de Rodes đã sáng tạo ra hệ thống chữ viết được gọi là "*quốc ngữ*" mà ngày nay trở thành ngôn ngữ chính thức. Cuối cùng, dưới thời Pháp thuộc, tiếng Việt đã mượn rất nhiều cấu trúc ngữ pháp và từ tiếng Pháp để sử dụng trong hệ thống tiếng Việt.

2. Bộ dữ liệu (corpus)

Cách đây một vài năm, vì nhiều lí do khác nhau, việc xây dựng một bộ dữ liệu bằng tiếng Việt là một nhiệm vụ vô cùng khó khăn. Nhưng sự phát triển của khoa học kĩ thuật, đặc biệt là sự tiến bộ của Internet đã góp phần rất lớn vào tiến trình này. Đối với chúng tôi, việc xây dựng một bộ dữ liệu chuyên ngành không phải là một ngoại lệ. Chúng tôi đã tìm những bài báo chuyên ngành khác nhau thuộc lĩnh vực khoa học kĩ thuật, các hiệp định, hiệp ước. Sau đó, chúng tôi đã sử dụng mã code UTF-8 để lưu trữ các văn bản dưới định dạng *.txt* nhằm tạo thuận lợi cho việc xử lí văn bản về sau này. Tất cả những thông tin đã lưu trữ sẽ giúp chúng tôi xác định được nguồn gốc của các bài báo, thời gian xuất bản và tên tác giả. Trong bộ dữ liệu này, chúng tôi tập trung các văn bản, hiệp định, hiệp ước về khoa học kĩ thuật và Bộ dữ liệu được giới hạn trong khoảng 500 ngàn từ.

3. Phân tách bộ dữ liệu (tokenize)

Nhằm xử lí bộ dữ liệu của mình, chúng tôi đã sử dụng phần mềm dán nhãn VnQTAG tại phòng thí nghiệm Loria (Nancy) phát triển. Bộ dán nhãn này thực hiện mã hóa bộ dữ liệu qua 2 bước: phân tách văn bản thành các đơn vị (token) và dán nhãn ngữ pháp. Vì những từ ghép xuất hiện rất nhiều trong tiếng Việt nên việc phân tách từ không thể thực hiện dựa trên căn cứ dấu hiệu khoảng trống (space) và chấm câu (punctuation) giống như các ngôn ngữ Anh, Pháp.

Bước đầu tiên, phần mềm VnQTAG sử dụng cơ sở âm tiết và từ vựng để công nhận tất cả khả năng phân tách trong bộ dữ liệu (sự phân tách này được định hướng bởi dấu hiệu chấm câu). Ý tưởng này đã đem lại một số kết quả khả quan. Một số trường hợp nhầm lẫn được giải quyết một cách thủ công (tức là kiểm tra thủ công sau khi sử dụng phần mềm). Cuối

* Trường Đại học sư phạm Hà Nội

cùng hệ thống cho phép người sử dụng thực hiện các lựa chọn sau khi phân tách từ.

Trong bước thứ hai, VnQTAG sẽ gắn nhãn cho mỗi token đã được xác định sau bước đầu tiên. Bộ gắn nhãn cho tiếng Việt bao gồm khoảng 11 nhóm từ loại (động từ, tính từ, danh từ,...).

Để minh hoạ, chúng ta hãy quan sát một câu được phân tách và dán nhãn như sau :

Người chuyên chở thực tế là bất kì người nào được chuyên chở ủy thác thực hiện việc chuyên chở hàng hóa hoặc một phần việc chuyên chở đó và bao gồm bất kì người nào khác được giao thực hiện việc chuyên chở đó (trích trong *Công ước về vận tải hàng hoá bằng đường biển*).

Tạm dịch (tiếng Pháp):

Le vrai transporteur: est n'importe quelle personne qui est mandaté de transporter les marchandises ou une partie de ce transport et comprend n'importe quelle autre personne est chargé d'effectuer ce transport.

Đầu tiên, câu này sẽ được phân tách trên cơ sở danh sách các âm tiết. Mỗi *token* phù hợp với một âm tiết hay một dấu hiệu chấm câu..

lNgười l chuyên l chở l thực l tế l là l bất l kì l người l nào l được l chuyên l chở l ủy l thác l thực l hiện l việc l chuyên l chở l hàng l hóa l hoặc l một l phần l việc l chuyên l chở l đó l và l bao l gồm l bất l kì l người l nào l khác l được l giao l thực l hiện l việc l chuyên l chở l đól .l

Sau đó, câu sẽ được dán nhãn trên cơ sở từ vựng. Mỗi *token* phù hợp với một loại từ (như danh từ, động từ, tính từ, etc...). Dựa theo kết quả phân tách từ, chúng ta nhận được nhiều giải pháp khác nhau.

Giải pháp 1:

<w pos="Np">Người</w> <w pos="Vt">chuyên chở</w> <w pos="Na">thực tế</w> <w pos="""">"</w> <w pos="Vla">là</w> <w pos="Jd">bất kỳ</w> <w os="Nc">người</w> <w pos="Pi">nào</w> <w pos="Jd">được</w> <w pos="Nc">người</w> <w pos="Vt">chuyên chở</w> <w pos="Nc">ủy</w> <w pos="Nc">thác</w> <w pos="Vt">thực hiện</w> <w pos="Ng">việc</w> <w pos="Vt">chuyên chở</w> <w pos="Ng">hàng hoá</w> <w pos="Cc">hoặc</w>

<w pos="Nn">một</w> <w pos="Na">phần</w> <w pos="Ng">việc</w> <w pos="Vt">chuyên chở</w> <w pos="Pd">đó</w> <w pos="Cc">và</w> <w pos="Vt">bao gồm</w> <w pos="Jd">bất kỳ</w> <w pos="Nc">người</w> <w pos="Pi">nào</w> <w pos="Aa">khác</w> <w pos="Jd">được</w> <w pos="Vt">giao phó</w> <w pos="Vt">thực hiện</w> <w pos="Ng">việc</w> <w pos="Vt">chuyên chở</w> <w pos="Pd">đó</w> <w pos=".">.</w>

Giải pháp 2:

<w pos="Np">Người</w> <w pos="Vt">chuyên</w> <w pos="Vt">chở</w>

<w pos="Na">thực tế</w> <w pos="""">"</w> <w pos="Vla">là</w> <w pos="Jd">bất kỳ</w> <w pos="Nc">người</w> <w pos="Pi">nào</w> <w pos="Jd">được</w> <w pos="Nc">người</w> <w pos="Vt">chuyên chở</w> <w pos="Nc">ủy</w> <w pos="Nc">thác</w> <w pos="Vt">thực hiện</w> <w pos="Ng">việc</w> <w pos="Vt">chuyên</w><w pos="Vt">chở</w> <w pos="Ng">hàng hoá</w> <w pos="Cc">hoặc</w> <w pos="Nn">một</w> <w pos="Na">phần</w> <w pos="Ng">việc</w> <w pos="Vt">chuyên</w><w pos="Vt">chở</w> <w pos="Pd">đó</w> <w pos="Cc">và</w> <w pos="Vt">bao gồm</w> <w pos="Jd">bất kỳ</w> <w pos="Nc">người</w> <w pos="Pi">nào</w> <w pos="Aa">khác</w> <w pos="Jd">được</w> <w pos="Vt">giao phó</w> <w pos="Vt">thực hiện</w> <w pos="Ng">việc</w> <w pos="Vt">chuyên</w><w pos="Vt">chở</w> <w pos="Pd">đó</w> <w pos=".">.</w>

Từ *"chuyên chở"* (transporter) là một từ ghép khá là "gai góc", bởi vì chúng ta có thể phân tách từ này như là một từ ghép (giải pháp 1) *"chuyên chở"* và gắn cho từ này nhãn ngoại động từ. Nhưng trong trường hợp 2, từ *"chuyên chở"* được tách làm 2 token l*chuyên*l và l*chở*l và được gián nhãn như 2 từ đơn <w pos="Vt">chuyên</w> (réservé) và <w pos="Vt">chở</w> (porter). Vậy thì, cách thức khả quan nhất là lựa chọn các token một cách thủ công và dán nhãn cho chúng.

4. Kết luận

Bài viết đã giới thiệu một số đặc điểm của tiếng Việt nhằm làm nổi bật những khó khăn nội hàm trong nghiên cứu tiếng Việt, ngôn ngữ duy nhất tại châu Á có sử dụng bảng chữ cái latin. Đây là nền tảng cho việc nghiên cứu các bước phân tích ngôn ngữ tự

động: phân tách từ và dán nhãn từ. Đối với việc xử lí tự động ngôn ngữ, tiếng Việt là một ngôn ngữ khá phức tạp ở cấp độ hình thái và kết hợp ngôn ngữ. Bởi vậy, việc kiểm tra tất cả các biến thể của ngôn ngữ cũng như việc dán nhãn cho từ mà chúng tôi đã thực hiện trong bài báo này là vô cùng cần thiết, giúp chúng tôi tạo ra một bảng từ vựng và tìm các kết hợp từ (collocations) nhằm tạo ra một công cụ giúp cho công việc soạn thảo và dịch thuật trong lĩnh vực khoa học kĩ thuật. ❑

Tài liệu tham khảo

1. A. Wirot. **Collocation and Thai Word Segmentation**. Proceeding of Joint International Conference of SNLP-Oriental COCOSDA. Thammasat University, Bangkok, 2002.
2. Agnès Tutin. **Le dictionnaire de collocations est-il indispensable**. LIDILEM, Grenoble, Revue française de linguistique appliquée, 2005.
3. Benoit Habert, Adeline Nazarenko, André Salem. **Les linguistiques de corpus**. Armand Colline, 1997.
4. Brigitte Bigi, Viet-Bac Le. **Normalisation et alignement de corpus français et vietnamiens: Format et Logiciels**. JADT 2008: 9es Journées internationales d'Analyse statistique des Données Textuelles, 2008.
5. Cam Tu N., Trung Kien N., Xuan Hieu P., Le Minh N., Quang Thuy H. **Vietnamese word segmentation with CRFs and SVMs: an investigation**. 6th international conference on Language Resources and Evaluation - LREC 2008.

SUMMARY

Automatic language processing (ALP) is still a new field to the language community in Vietnam. The problem approaching and word processing are posed thorny issue for the linguists of application. In the framework of a short article, the author generally introduces the processing measures and steps to automatic processing of this specialized corpus. Firstly, the Vietnamese language is analyzed in terms of semantic features to highlight the tokenization and tagging. Then, the Vietnamese tokenization and tagging are illustrated by applying to the scientific and technical corpus.

Giáo dục giá trị sống cho sinh viên...

(Tiếp theo trang 45)

sắc mãnh liệt, có ích và có ý nghĩa. Hộ trong ***Đời thừa*** hằng tâm niệm: "*Kẻ mạnh không phải là kẻ giẫm lên vai kẻ khác để thoả mãn lòng ích kỉ. Kẻ mạnh chính là kẻ giúp đỡ kẻ khác trên đôi vai của mình*" và mơ ước viết được "*một tác phẩm thực sự có giá trị (...) làm cho người gần người hơn*",... Trong các sáng tác ấy, Nam Cao đòi hỏi mỗi cá nhân được phát triển đến tận độ với một ý thức sống đầy trách nhiệm trong mối quan hệ mật thiết với sự phát triển chung của xã hội loài người.

* * *

Mỗi tác phẩm văn học giảng dạy trong trường học đều chuyển tải những GTS đích thực giúp HS, SV hoàn thiện nhân cách và làm giàu có hơn tâm hồn cũng như tri thức của mình. Đối với SV CĐSP, việc nhận thức GTS và vận dụng những bài học sống ấy vào thực tế là vô cùng quan trọng. Bởi những thầy cô giáo tương lai không chỉ phải sống đẹp, sống có ý nghĩa mà họ còn phải biết chọn lọc những GTS thiêng liêng trong từng bài học từng giờ giảng để giáo dục các thế hệ học sinh tại trường trung học cơ sở. Thiết nghĩ, đó cũng là cách tốt nhất để nhân lên những hành động sống đẹp trong nhà trường và xã hội. ❑

Tài liệu tham khảo

1. Nguyễn Văn Long. *Nguyễn Khải và sự đổi mới quan niệm về con người trong "Một người Hà Nội"* (*tapchinhavan.vn*).
2. Đỗ Lai Thuý. "*Xuân Diệu - nỗi ám ảnh thời gian*" (trong **Con mắt thơ**). NXB *Văn học*, H. 1998.
3. Trần Đăng Suyền. "Nam Cao nhà nhân đạo chủ nghĩa" (trong **Nam Cao, tác giả tác phẩm**). NXB *Giáo dục*, H. 2002.
4. Lưu Phát - Ánh Tuyết. "Sự cần thiết của giáo dục kĩ năng sống cho sinh viên sư phạm" (spnttw.edu.vn).

SUMMARY

The education of the life values is to get students, adolescents to realize the true value of life - the noblest ones which have been recognized, treasured and strived for by the community. As for literature students, literature works would be the most effective way to educate them on life values. It is due to the fact that one of the leading functions of literature is education which includes: the enhancement of ideal, tastes and help the readers to realize the eternal values of life. Human life's meaning is always involved in literature such as: the awareness of time's value and the appreciation of life which is inspired by XuanDieu's work; the inspiration of stuff, self-respect and elegant lifestyle in "Mot nguoi Hanoi" (A Hanoian -Nguyen Khai); or the sense of sympathy, the appreciation of human being's true value by Nam Cao's literary work. The first valuable lessons will assist the youngers to become part of life, to improve themselves step by step as well as foster the desire to live positively and beautifully.