

KHỐI LIỆU – ĐẶC TRƯNG VÀ PHÂN LOẠI (phần II)

Đào Hồng Thu
(TS, Đại học Bách khoa HN)

Trong phần 1, chúng tôi đã đưa ra các khái niệm: khối liệu, ngôn ngữ học khối liệu (ngôn ngữ học corpus), corpus, corpus văn bản. Trong đó, các thuật ngữ *corpus* và *corpus văn bản* là tương đồng về ngữ nghĩa. Về tiếng Việt, các thuật ngữ trên có thể đặt ở vị trí tương đồng với thuật ngữ "*khối liệu*".

Khối liệu là tập hợp các dữ liệu tương đồng về mặt ngôn ngữ, được trình bày dưới dạng model văn bản điện tử, theo các cấu trúc nhất định. Khối liệu được sử dụng để giải quyết các vấn đề về ngôn ngữ trong các lĩnh vực khoa học cụ thể. Việc tìm kiếm dữ liệu trong khối liệu theo bất kì từ nào đều cho phép lập danh mục tất cả các trường hợp sử dụng từ đã cho trong ngữ cảnh với đầy đủ dẫn nguồn. Khối liệu có thể được sử dụng làm cẩm nang hướng dẫn và tìm kiếm thông tin, cũng như dữ liệu thống kê về các đơn vị ngôn ngữ và lời nói. Trên cơ sở khối liệu chúng ta có thể nhận được các dữ liệu về tần số hình thái từ, đơn vị từ vựng, phạm trù ngữ pháp; có thể theo dõi được các thay đổi về tần số và ngữ cảnh ở các thời điểm khác nhau v.v. Cuối cùng, khối liệu được sử dụng làm cơ sở và công cụ biên soạn các thể loại từ điển lịch sử và hiện đại khác nhau; được sử dụng để xây dựng và giải thích ngữ pháp; để phục vụ cho việc dạy học bản ngữ và ngoại ngữ cũng như dịch thuật.

1. Các đặc trưng cơ bản của khối liệu

Có thể nói rằng khối liệu là mô hình nhỏ nhất của ngôn ngữ. Căn cứ vào hoạt động của khối liệu trong các hệ thống của nó có thể thấy các đặc trưng cơ bản sau:

a. Tính điển hình

Đây là khái niệm quan trọng nhất của Ngôn ngữ học khối liệu. Đặc trưng này xác định tính thực tiễn của khối liệu, có nghĩa là khối liệu phải là tập hợp của nhiều loại văn bản phù hợp cho

việc nghiên cứu đối với nhiều ngôn ngữ. Điều đặc biệt quan trọng là khối liệu chung phải hàm chứa được một tỉ lệ xác định các văn bản thuộc nhiều thể loại, tại nhiều thời điểm lịch sử, của nhiều tác giả khác nhau v.v.

Tính điển hình của khối liệu có thể được hiểu là sự thể hiện tỉ lệ cần và đủ trong khối liệu về dung lượng, thể loại, phong cách, số lượng tác giả, vùng địa lí, thời kì lịch sử v.v.

b. Kích cỡ xác định

Khái niệm "khối liệu" có nghĩa thông thường là tập hợp các văn bản có kích cỡ xác định. Theo thời gian, kích cỡ (dung lượng và thành phần) của khối liệu có thể thay đổi. Tuy nhiên, những thay đổi này hoặc là không được kéo theo sự thay đổi về tính điển hình của toàn khối hoặc là phải thay đổi nền của khối liệu cần thay đổi.

Các khối liệu ban đầu như khối liệu Brown hoặc khối liệu tiếng Nga Upsanski bao gồm một triệu đơn vị từ và cụm từ sử dụng. Ngày nay, các khối liệu cần chứa được ít nhất là 100 triệu đơn vị từ và cụm từ sử dụng.

Khi xây dựng khối liệu cần lưu ý đến mục đích sử dụng ngữ liệu khối và người sử dụng khối liệu được tạo lập. Quá trình xây dựng khối liệu cho thấy mỗi nhóm người sử dụng khối liệu đều có nhu cầu riêng của mình. Ví dụ, phần lớn người sử dụng tìm trong khối liệu các từ hoặc cụm từ khi gặp phải các vấn đề chính tả hoặc phong cách của chúng, thường là họ tìm đồng nghĩa của từ hoặc cụm từ. Đối với các trường hợp như vậy thì chỉ cần xây dựng khối liệu ngôn ngữ vừa đủ, có thể chấp nhận một số từ hoặc cụm từ trong khối chưa mang đầy đủ tính điển hình của khối liệu được tạo dựng với nguồn từ điển thông dụng của ngôn ngữ toàn dân. Một ví dụ khác, đối với người sử dụng là chuyên gia một chuyên ngành công nghệ chẳng hạn thì người xây dựng khối liệu cần

tuân thủ đầy đủ các nguyên tắc về tính điển hình, kích cỡ xác định và phong cách ngôn ngữ của chuyên ngành đó bên cạnh một lượng dự trữ từ và cụm từ của từ điển thông dụng. Hoặc đối với người sử dụng là nhà từ vựng học thì ngoài nguồn từ và cụm từ của từ điển thông dụng ra còn cần một lượng các đơn vị từ theo chuyên ngành cụ thể và các phương tiện ngôn ngữ kèm theo. Đối với người sử dụng là các nhà lí luận ngôn ngữ và chuyên gia trong lĩnh vực soạn thảo ngôn ngữ máy tính thì cần khối liệu đặc trưng về mặt ngôn ngữ, bao gồm các dữ liệu về từ vựng - ngữ nghĩa, cú pháp và hình thái học.

c. Tính chú giải

Để giải quyết được các vấn đề ngôn ngữ khác nhau trong khối liệu, khi xây dựng khối liệu cần có lượng đáng kể chú giải thông tin ngôn ngữ và ngoại ngôn ngữ thuộc các thể loại khác nhau. Vì vậy, trong Ngôn ngữ học khối liệu hình thành các khối chú giải thông tin. Chú giải (tagging) là phân giải thích các thông tin đặc thù làm rõ nghĩa cho các văn bản trong khối liệu như là chú giải bên ngoài, ngoại ngôn ngữ (ví dụ, chú giải về tác giả: tên, tuổi, giới tính, năm sinh v.v. và về văn bản: tác giả, tên văn bản, năm và nơi xuất bản, thể loại, phong cách ngôn ngữ v.v.); hoặc là chú giải cấu trúc (ví dụ, chương, đoạn, câu, hình thái từ v.v.); hoặc là chú giải cho chính ngôn ngữ văn bản về từ vựng, cú pháp.

Các nghiên cứu về Ngôn ngữ học khối liệu đã cho thấy việc xác định và lựa chọn các loại hình chú giải phải do các nhà ngôn ngữ tiến hành trên cơ sở các văn bản được lựa chọn cho việc xây dựng khối liệu.

Chú giải trong khối liệu có thể được chia thành:

- **Chú giải hình thái học (POS-tagging)**

Đây là dạng cơ bản trong phân tích và xây dựng khối liệu bởi vì phần lớn các khối liệu lớn chính là các khối liệu nhỏ được chú giải hợp lại về mặt hình thái học. Trong quá trình xây dựng khối liệu, phân tích hình thái học được xem là cơ sở của phân tích cú pháp và phân tích ngữ nghĩa.

Chú giải hình thái học bao gồm chú giải các thành phần lời nói (viết và nói) và phạm trù ngữ pháp của các thành phần đó trong khối liệu. Đơn vị chú giải hình thái học là từ (<w>) hoặc cụm từ.

Hiện nay, các thành tựu về hình thái học máy tính đã phát triển ở mức độ cho phép tự động chú

giải các khối liệu có kích cỡ lớn. Chú giải hình thái học ngày nay thường được thực hiện một cách tự động nhưng bắt buộc phải có sự tham gia của con người vào việc lựa chọn nghĩa từ xác định trong ngữ cảnh cụ thể cho khối liệu do trong phân tích hình thái học có thể xảy ra các hiện tượng đa nghĩa của từ.

Các nghiên cứu về khối liệu cho thấy nếu chú giải hình thái học được lựa chọn càng nhiều bao nhiêu thì phân tích văn bản càng được chi tiết bấy nhiêu. Ngày nay đang có xu hướng giảm số lượng chú giải hình thái học do dung lượng các khối liệu ngày càng tăng. Hệ thống làm đơn giản quá trình mã hóa tạo điều kiện cho việc tránh mắc lỗi không cần thiết, sử dụng một cách lôgic các dữ liệu, tránh các trường hợp đa nghĩa của từ trong văn bản và tăng tốc độ chú giải của các khối liệu lớn với hàng triệu từ.

- **Chú giải cú pháp (parsing)**

Chú giải cú pháp là kết quả của phân tích cú pháp được thực hiện trên cơ sở dữ liệu về phân tích hình thái học. Đây là dạng chú giải mô tả các mối quan hệ cú pháp giữa các đơn vị từ vựng và cấu trúc cú pháp khác nhau (ví dụ, mệnh đề phụ thuộc, mệnh đề độc lập, thành ngữ v.v.).

Hiện nay, ở mức độ phân tích cú pháp đang tồn tại xu hướng chi tiết hóa nhỏ nhất việc mã hóa chú giải để tăng tốc độ và trình tự phân tích văn bản. Điều này cũng diễn ra đối với chú giải hình thái học và chú giải ngữ nghĩa.

- **Chú giải ngữ nghĩa (semantic tagging)**

Trong quá trình xây dựng khối liệu thì phần chú giải ngữ nghĩa là phần phức tạp và khó khăn nhất. Cho đến nay, đối với ngữ nghĩa học dành cho khối liệu vẫn chưa có sự đồng nhất chung về mặt lí luận. Tuy vậy, các phạm trù ngữ nghĩa của các đơn vị từ và cụm từ sử dụng đã được xác định khá rõ ràng cho việc xây dựng các khối liệu cụ thể.

Do tính chất phức tạp trong việc lựa chọn các chú giải ngữ nghĩa cho khối liệu nên hiện nay các nhà nghiên cứu ngôn ngữ học khối liệu đang tập trung rất nhiều công sức vào lĩnh vực đang phát triển. Một điều hiển nhiên là chú giải ngữ nghĩa hiện đang được phát triển rất mạnh bởi tính xác thực của nó đối với hoạt động của khối liệu.

Kiểu chú giải ngữ nghĩa là các mã (code) gồm các con chữ và chữ số. Các nghiên cứu cho thấy có khoảng 250 - 300 đơn vị mã để phân loại

toàn bộ từ vựng.

- Ngoài các loại hình chú giải chủ yếu trên còn tồn tại các chú giải khác như chú giải từ loại, chú giải ngôn điệu, chú giải bình luận hoặc tranh luận v.v.

2. Phân loại các khối liệu chủ yếu

Trong Ngôn ngữ học khối liệu, khối liệu có thể được phân loại theo phương thức phân nhóm hoặc theo kiểu chú giải ngôn ngữ. Mặc dù tồn tại nhiều kiểu chú giải, trên thực tế, phần lớn các khối liệu có kiểu chú giải cú pháp / hình thái học (treebanks - ngân hàng cấu trúc cú pháp - ĐHT dịch). Một điều hiển nhiên và cần nhấn mạnh là khối liệu có chú giải cú pháp luôn hàm chứa các đặc trưng hình thái học của các đơn vị từ vựng.

Căn cứ vào mục đích nghiên cứu - sử dụng khối liệu và đặc điểm hoạt động của nó, có thể phân loại như sau:

- Theo kiểu dữ liệu, các khối liệu được chia thành khối liệu viết, khối liệu nói, khối liệu kết hợp;
- Theo ngôn ngữ văn bản, các khối liệu được chia thành khối liệu tiếng Việt, khối liệu tiếng Anh, khối liệu tiếng Nga.
- Theo đặc tính song song của văn bản, các khối liệu được chia thành khối liệu đơn ngữ, khối liệu song ngữ và khối liệu đa ngữ.
- Theo thể loại chuyên ngành, các khối liệu được chia thành khối liệu khoa học phổ biến, khối liệu không phổ biến, khối liệu ngôn ngữ học ứng dụng, khối liệu khoa học kỹ thuật và công nghệ.
- Theo tính chất chú giải, các khối liệu được chia thành khối liệu có chú giải và khối liệu không có chú giải.
- Theo chức năng và mục đích sử dụng, các khối liệu được chia thành khối liệu nghiên cứu, khối liệu minh họa, khối liệu tham khảo.
- Theo tính chất xã hội của văn bản, các khối liệu được chia thành khối liệu cộng đồng, khối liệu tác phẩm.

Tài liệu tham khảo

1. Андрющенко В.М. Концепция и архитектура машинного фонда русского языка / Отв.ред.А.П. Ершов. М., 1989.
2. Баранов А.Н. Корпусная лингвистика // Баранов А.Н. Введение в прикладную лингвистику. М., 2001. С.112-137.
3. Вербицкая Л.А., Казанский Н.Н., Касевич В.Б. Некоторые проблемы создания национального корпуса русского языка // Научно-техническая информация. Сер. 2. 2003. № 6. С. 2-8.

4. Доклады научной конференции «Корпусная лингвистика и лингвистические базы данных» / Под ред. А.С. Герда. СПб., 2002.

5. Марчук Ю.Н. Корпус текстов и сверхбольшие базы лингвистических данных // Сборник: Труды международной конференции «Корпусная лингвистика – 2002». - Издательство Санкт-Петербургского университета, 2002.

6. Holmes-Higgin P., Ahmad K. Assembling and Viewing a Corpus of Texts: Self-organisation, Logical Deduction and Spreading Activation as Metaphors // Euralex'96 Proceedings. – Stockholm, 1996.

7. Рыков В.В. Прагматически ориентированный корпус текстов // Тверской лингвистический меридиан Вып. 3. Тверь, 1999. С. 89-96..

8. Труды Международного семинара по компьютерной лингвистике и ее приложениям «Диалог-2003», «Диалог-2004», «Диалог-2005»..

9. Труды Международной научной конференции «Корпусная лингвистика 2004» / Под ред. А.С. Герда. СПб., 2004.

10. Чардин И.С. Лингвистические корпуса с синтаксической разметкой и их применение // Научно-техническая информация. Сер. 2. 2003. № 6. - стр. 18-24.

11. Шимкова М. Репрезентативность корпуса как лингвистическая проблема // Сборник: Труды международной конференции «Труды международной конференции «MegaLing'2005. Прикладная лингвистика в поиске новых путей» - 2005. - стр.124.

12. English Corpus Linguistics: Studies in Honour of Jan Svartvik / Aijmer K., Altenberg B. (eds.). London, 1991.

13. Fillmore C.J., Atkins B.T.S. Starting Where the Dictionaries Stop: the Challenge of Corpus Lexicography // Atkins B.T.S., Zampolli A. (eds.). Computational Approaches to the Lexicon. 1994.

14. Kennedy G. An Introduction to Corpus Linguistics. London, 1998.

15. Leech G. The State of Art in Corpus Linguistics // English Corpus Linguistics / Aijmer K., Altenberg B. (eds.). London, 1991. P. 8-29.

16. McEnery A., Wilson A. Corpus Linguistics. Edinburgh, 1996.

17. Francis N.W. Language Corpora B.C. // Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82. Stockholm, 4.-6. August 1991. / Svartvik J. (ed.), p. 17-32.

18. Proceedings of the LREC (Language Resource Evaluating Conference). 2004, 2005.

19. Quirk R. On Corpus Principles and Design // Directions in Corpus Linguistics. Berlin; New York, 1992. P. 461-462.

20. Sinclair J. M. The Automatic Analysis of Corpora // Directions in Corpus Linguistics. Berlin, 1992.

21. Svartvik, J. (ed.). Directions in Corpus Linguistics, Berlin. 1992.

22. Zakharov V. Russian Corpus of the 19th Century // Text, Speech and Dialogue: Proceedings of the 6th International Conference TSD 2003, p. 146-151. (Lecture Notes in Artificial Intelligence, 2807.

23. Dao Hong Tху. Корпус параллельных текстов в аспекте корпусной лингвистики. // Проблемы современной филологии и лингводидактики, сб. научных трудов, СПб, изд. РГПУ им. А.И.Герцена, 2006, с.23-28;

(Bài này gửi đến Ban biên tập ngày 06-02-2007)