XÂY DỰNG WORDNET CHO DANH TỪ TIẾNG VIỆT

Văn Chí Nam, Đinh Điền, Phạm Minh Tuấn
Khoa CNTT, Trường Đại học Khoa học Tự Nhiên - ĐHQG tp.HCM

Tóm tắt:

WordNet là một hệ cơ sở tri thức khổng lồ về ngôn ngữ học của từ vựng tiếng Anh, được xây dựng bởi các nhà ngôn ngữ học - máy tính, ngôn ngữ học tâm lý, và ngôn ngữ học tri nhận ở Đại học Priceton (Mỹ) từ đầu thập niên 1980. Trong WordNet, các từ được xếp vào các nhóm đồng nghĩa (tập các nghĩa có thể thay thế nhau trong một ngữ cảnh nào đó). Các nhóm đồng nghĩa liên kết với nhau thông qua các quan hệ ngữ nghĩa được xây dựng thông qua các nghiên cứu về cách sử dụng, cách lưu trữ các tri thức về ngôn ngữ trong bộ não con người. Nhờ cách tổ chức như vậy, WordNet đã cung cấp nhiều tri thức hữu dụng cho việc xử lý ngôn ngữ tự nhiên. WordNet hiện được dùng như một chuẩn phổ biến trong các ứng dụng xử lý ngôn ngữ tự nhiên.

Danh từ là loại từ phổ biến và phổ dụng trong mọi ngôn ngữ. Đến nay, đã có nhiều cách phân lớp danh từ tiếng Việt theo các tiêu chí khác nhau, nhưng ít nhiều các cách này đều mang tính chủ quan và chỉ được thực hiện trên một số ít các ví dụ cụ thể. Tuy nhiên, trong thực tế, khi phân giải ngữ nghĩa của một danh từ tiếng Việt trong một ngữ cảnh bất kỳ, chúng ta lại cần đến một hệ thống phân lớp hoàn chỉnh cho tất cả các danh từ tiếng Việt theo những ý niệm chung nhất trong tư duy của con người. Việc xây dựng một hệ thống phân lớp như thế đã được thực hiện thành công lần đầu tiên đối với tiếng Anh qua mạng WordNet, và cũng chính từ đây, các mạng tương tự cho tiếng Pháp, Tây Ban Nha, Đức, Hoa, Nhật, . đã được hình thành trên cơ sở mạng này.

Việc xây dựng một mạng từ vựng tương tự WordNet có nhiều ý nghĩa. Nó hỗ trợ cho việc phát triển các ứng dụng xử lý ngôn ngữ tiếng Việt, cho các nghiên cứu về ngôn ngữ học tiếng Việt. Trong bài báo này, chúng tôi sẽ trình bày cách thức rút trích (bán) tự động mối liên hệ ngữ nghĩa trong WordNet tiếng Anh và thông qua một số từ điển song ngữ để xây dựng một mạng từ vựng tiếng Việt ở phần danh từ. Bài báo sẽ gồm các phần : phần 1 - Giới thiệu về cấu trúc của mạng WordNet tiếng Anh ; phần 2 - tóm tắt một số công trình tương tự ; phần 3 - các phương pháp được áp dụng trong rút trích ; phần 4 - thử nghiệm và đánh giá ; phần 5 - kết luận và hướng phát triển.

BUILDING WORDNET FOR VIETNAMESE NOUNS

Van Chi Nam, Dinh Dien, Pham Minh Tuan
Department of Information Technology, University of Natural Sciences - VNU.HCM

Abstract:

WordNet, a very large knowledge-based lexicon in English linguistics, has been being constructed by the computational linguists, psycholinguists at Priceton University (United States) since the first years of 1980s. In WordNet, words are organized into synonym sets - synsets (that is a set of words are interchangeable in some context). The synsets connect each other through the sematic relationships which are built based on the studies of storing and using languages in human brains. With this organizing, WordNet provides lots of useful knowledge in processing natural language. WordNet now is used as a de facto standard in the natural language processing applications.

Noun is a popular part of speech of all the natural languages. Although there are lots of ways applied to categorize Vietnamese nouns, these ways are still subjective and only used in a few examples. However, in reality, to explain a meaning of Vietnamese noun in a specific context, we need a fully-made categorized system for all Vietnamese nouns basing on the most common concepts of human ideas. The construction of such categorized system has first succeeded in English with WordNet. The similar networks for French, Spanish, German, Chinese, Japanese.., are also built based on WordNet.

Constructing a sematic network similar to WordNet has full of meaning. Its usefulness can be found in the Vietnamese natural language processing applications, in Vietnamese linguistics researches. In this paper, we present the way to (semi)automatically extract the sematic relationships in English WordNet. With these relationships and some bilingual dictionaries, we build a Vietnamese noun sematic network. This paper is organized as follow : we introduce the structure of English WordNet in part 1, then we summarize the similar works in part 2, the methods used to extract are in part 3, and part 4 are the evaluations and testing, finally, in part 5, we conclude this paper and give some ideas to develop in the future.