GÁN NHÃN TỪ LOẠI TỰ ĐỘNG CHO TIẾNG VIỆT

Đinh Điền, Nguyễn Văn Toàn, Diệp Chí Cường*
Khoa CNTT, Trường Đại học Khoa học Tự Nhiên - ĐHQG tp.HCM
* Trung tâm Phát triển CNTT - ĐHQG tp.HCM

Tóm tắt:

Gán nhãn từ loại (POS Tagger) của một từ là một công đoạn rất quan trọng và bắt buộc phải có đối với mọi hệ xử lý ngôn ngữ tự nhiên. Đối với tiếng Anh, bài toán này đã được giải quyết khá tốt (95%), nhưng đối với tiếng Việt thì đây là điều còn rất mới mẻ. Trong bài báo này, chúng tôi sẽ trình bày mô hình gán nhãn từ loại tự động bằng phương pháp học TBL (Transformation-Based Learning) trên tập mẫu tiếng Việt đã được gán nhãn chính xác được rút ra từ kho ngữ liệu song ngữ Anh-Việt của chúng tôi.

Trong bài toán này, đối với các ngôn ngữ thông dụng như Anh, Pháp, Nga,. thì phương pháp học TBL đã được áp dụng thành công vì họ đã có những kho ngữ liệu đã được đánh dấu chính xác (annotated corpora) và họ không cần quan tâm vào việc nhận diện ranh giới từ trước khi đánh dấu (do đó là những ngôn ngữ biến hình). Nhưng đối với tiếng Việt (một ngôn ngữ đơn lập), chúng tôi lại phải giải quyết hai trở ngại nói trên trước khi tiến hành gán nhãn từ loại tự động. Với trở ngại thứ nhất, chúng tôi đã khắc phục thành công với chương trình tách từ tiếng Việt tự động với độ chính xác 95% (đã công bố ở Hội nghị Quốc tế về Xử lý ngôn ngữ Châu Á - Thái Bình Dương NLPRS01 tại Tokyo vào 11/2001). Với trở ngại thứ hai, chúng tôi cũng đã khắc phục bằng việc xây dựng kho ngữ liệu song ngữ Anh-Việt với hơn 500.000 câu mà trong đó hơn 25.000 câu tiếng Việt đã được gán nhãn từ loại chính xác nhờ kết quả liên kết từ Anh-Việt và phép chiếu từ loại từ Anh sang Việt (đã công bố ở Hội nghị Quốc tế về Xử lý ngôn ngữ APIS02 tại Bangkok, Thái lan vào 2/2002).

Chúng tôi đã áp dụng thử nghiệm mô hình này và bước đầu nhận được kết quả rất khả quan (trên 80%). Trong tương lai, chúng tôi sẽ tăng khối lượng tập mẫu huấn luyện lên 50.000 câu và chắc chắn độ chính xác sẽ tăng theo. Thành công của việc gán từ loại tiếng Việt tự động sẽ là cơ sở cho những bước đi tiếp theo trong việc xử lý tiếng Việt, như: xác định ranh giới ngữ (danh ngữ, động ngữ, .), phân tích cú pháp, phân tích ngữ nghĩa, .

AUTOMATIC POS TAGGER FOR VIETNAMESE

Dinh Dien, Nguyen Van Toan, Diep Chi Cuong*
Department of Information Technology, University of Natural Sciences - VNU.HCM
* Center of Information Technology Development - VNU.HCM

Abstract:

Assigning the Part-of-Speech tag (POS Tagger) to a word is a very important and obligational task for every Natural Language Processing Systems. For English, this problem has been solved with high accuracy (approx. 96%), but in Vietnamese, this problem hasnt been solved yet. In this paper, we will present a model for automatically POS tagging by TBL (Transformation-Based Learning) method running on our annotated Vietnamese corpus which is drawn out from our aligned English - Vietnamese bilingual corpus.

In this problem, for popular languages such as English, French, Russia, etc., the TBL method have been applied successfully because they have already had large annotated corpora and they needed not to solve the word segmentation problem before tagging (because these languages are inflectional languages). But in Vietnamese (an isolating language), we had to solve two abovementioned drawbacks before POS tagging. In the first drwaback, we have solved by our automatic Vietnamese Word Segmenter with the accuracy of 95% (published in the international conference on Natural Language Processing of Pacific-Rim languages titled NLPRS01 in Tokyo, Nov. 2001). In the second drwaback, we have overcome it by building a 500,000 - sentence English Vietnamese bilingual corpus in which 25,000 Vietnamese sentences are POS tagged by English - Vietnamese Word Alignment Program (published in APIS02 - Bangkok, Feb. 2002) .

We have experimented and evaluated this model and its performance appeared to be good result (approx. 80%). Later, we will enhance our bilingual corpus to 50,000 - sentence one and the accuracy of model will certainly be improved. In the future, we will apply this model for chunker (NP, VP,.), parser and semantic analyzer for Vietnamese.