PHÂN TÍCH CÚ PHÁP SONG SONG CHO SONG NGỮ ANH - VIỆT

Nguyễn Lưu Thuỳ Ngân, Đỗ Xuân Quang, Nguyễn Văn Toàn
Khoa CNTT, Trường Đại học Khoa học Tự Nhiên - ĐHQG tp.HCM

Tóm tắt:

Hiện nay với sự phát triển trong lĩnh vực dịch máy, nhiều hệ dịch đã lần lượt ra đời trên thế giới. Các hệ dịch dù khác nhau về hướng tiếp cận thì mục tiêu cuối cùng vẫn là thu được văn bản ở ngôn ngữ đích có chất lượng cao. Trong các cách tiếp cận khác nhau đó có thể nói hiện nay các hệ dịch dựa trên nền tảng chuyển đổi cú pháp giữa hai ngôn ngữ nguồn và đích là hệ dịch có kết quả cao nhất. Trong các hệ dịch này công đoạn phân tích ngữ pháp đóng một vai trò quyết định, đảm bảo kết quả dịch theo đúng văn phong của ngôn ngữ đích. Đặc biệt đối với hệ dịch Anh-Việt, việc phân tích ngữ pháp tiếng Anh và tiếng Việt tốt mới có thể cho kết quả cao trong công đoạn chuyển đổi cây cú pháp Anh-Việt và là cơ sở cho toàn bộ hệ dịch.

Trong bài báo này chúng tôi trình bày một phương pháp phân tích ngữ pháp song song trên hai ngôn ngữ , được xây dựng dựa trên mô hình ngữ pháp tương quan đảo trật tự sử dụng thống kê (Stochastic Inversion Transduction Grammar). Khác với những loại ngữ pháp thông thường chỉ mô hình hoá cho một ngôn ngữ nhất định, SITG mô hình hoá cho một cặp ngôn ngữ có mối quan hệ với nhau về mặt cấu trúc. Hơn nữa SITG có tính chất phi ngữ cảnh, không dựa trên trạng thái xác định, do đó hiệu quả khi ứng dụng cho phân tích cú pháp, kết hợp với ưu điểm của phương pháp thống kê để tìm ra cách phân tích cú pháp có tính chính xác cao nhất.

Trái ngược với tiếng Anh, các bộ phân tích ngữ pháp tiếng Việt phần lớn đều cho kết quả không cao do bởi chưa có một công trình nghiên cứu về việc hình thức hoá ngữ pháp tiếng Việt một cách hoàn chỉnh. Trong khi đó, phương pháp phân tích cú pháp song song này là nhằm khai thác các ràng buộc ngữ pháp từ ngữ liệu song ngữ, do đó tận dụng được thế mạnh riêng của mỗi ngôn ngữ và mối tương quan giữa chúng. Lợi ích lớn nhất được mang lại từ công trình nghiên cứu này là: "ánh xạ" các kết quả nghiên cứu khá hoàn chỉnh của tiếng Anh sang tiếng Việt. Vì vậy, thành công của phân tích ngữ pháp song song sẽ mang lại rất nhiều thuận lợi cho việc phân tích cú pháp tiếng Việt nói riêng và các xử lý tiếng việt khác nói chung.

BILINGUAL PARSING FOR ENGLISH - VIETNAMESE CORPUS

Nguyen Luu Thuy Ngan, Do Xuan Quang, Nguyen Van Toan
Department of Information Technology, University of Natural Sciences - VNU.HCM

Abstract:

With the development of machine translation nowadays, many translation systems have been built. Although their approaches are various, their only goal is producing a high-quality translation in the target language. Among these approaches, transfer-based systems have drawn much attention because of their best results in reality. Problems of structural transfer in transfer-based systems can be resolved by parsing each half of the bilingual corpus separately and drawing transfer rules . In these systems parsing plays an important role. It means that we have to find out a believable method to parse Vietnamese in English-Vietnamese transfer-based system.

In this paper we present a bilingual parsing of parallel corpora, which is based on Stochastic Inversion Transduction Grammar, introduced by Dekai Wu (1997). Other than usual grammars which describe the structure of only a language, transduction grammar describes a structurally correlated pair of languages. SITG is a context-free model, rather than finite-state model. So it is very useful for parsing. Moreover a probability is associated with each rewrite rule in a SITG so that we can compute an optimal parse a specific pair of sentences.

On contrary to English, up to now there is almost no Vietnamese parsing which shows a high performance because Vietnamese grammars have not been well-studied yet. Meanwhile this bilingual parsing method collects constraints from Vietnamese-English corpus and makes use of correlations to bear on the task of extracting linguistic information for Vietnamese. And that is the most advantage of this method. Success in bilingual parsing will contribute much to Vietnamese parsing and other tasks of Vietnamese processing.