CHIẾU CHÚ THÍCH ĐỒNG THAM CHIẾU SỬ DỤNG SONG NGỮ ANH-VIỆT

Lê Tuấn Thu[1], Lương An Vinh[2]
[1]Khoa Công nghệ Thông tin,
Trường Đại học Khoa Học Tự nhiên, ĐHQG-HCM
[2]Khoa Công nghệ Thông tin,
Trường Đại học Công nghệ Sài Gòn
thuletuan@gmail.com, vinh.luongan@stu.edu.vn

Tóm tắt

Phân giải đồng tham chiếu là xác định các cụm danh từ mà tham chiếu tới các thực thể trong thế giới thực. Nghiên cứu này sử dụng phép chiếu dữ liệu đồng tham chiếu được chú thích thủ công từ tiếng Anh sang tiếng Việt trong ngữ liệu song ngữ. Phương pháp này thích hợp với các ngôn ngữ nghèo tài nguyên vì không yêu cầu phải có dữ liệu lớn đã chú thích đồng tham chiếu bằng tay để huấn luyện trình phân giải cho các ngôn ngữ này. Để chiếu các chú thích đồng tham chiếu từ tiếng Anh sang tiếng Việt, đầu tiên, chúng tôi gióng hàng các từ tiếng Anh và tiếng Việt trong mỗi cặp câu song song và sau đó chiếu các chú thích đồng tham chiếu tiếng Anh sang văn bản tiếng Việt sử dụng gióng hàng từ. Kết quả đánh giá các chuỗi đồng tham chiếu chuyển giao trên biểu thức tham chiếu được phát hiện chính xác với B-cubed độ đo F khoảng 85%. Phương pháp tiếp cận chiếu chú thích có tiềm năng cho phép xây dựng ngữ liệu tiếng Việt có thông tin chú thích đồng tham chiếu.

Từ khoá: đồng tham chiếu, phân giải đồng tham chiếu, chiếu chú thích, chú thích đồng tham chiếu

PROJECTING COREFERENCE ANNOTATIONS USING ENGLISH-VIETNAMESE PARALLEL CORPUS

Le Tuan Thu[1], Luong An Vinh[2]
[1]Faculty of Information Technology, University of Science, VNU-HCM
[2]Faculty of Information Technology, Saigon Technology University
thuletuan@gmail.com, vinh.luongan@stu.edu.vn

Abstract

Co-reference resolution is a task of identifying noun phrases that refer to entities in the real world. This research involves projecting hand–annotated co-reference data from English to Vietnamese via a parallel corpus. This approach is appropriate for low-resource languages because it does not require a large collection of manually annotated data to train co-reference resolvers in these languages. To project co-reference annotations from English to Vietnamese, we first aligned the English and Vietnamese words in each pair of parallel

sentences, and then projected the English co-reference annotations onto the Vietnamese text using the word alignment. The result were evaluated the transferred co-reference chains on correctly detected referential expression the B-cubed F-measure with about 85%. The projection approach has the potential to create co-reference-annotated data for the Vietnamese language.
Key words: co-reference, co-reference resolution, projection, projecting annotations