KẾT HỢP CÁC NGUỒN TRI THỨC KHÁC NHAU ĐỂ XỬ LÝ NHẬP NHẰNG NGỮ NGHĨA CHO HỆ DỊCH ANH - VIỆT

Trần Lê Hồng Dũ, Đinh Điền, Nguyễn Anh Tuấn
Khoa CNTT, Trường Đại học Khoa học Tự Nhiên - ĐHQG tp.HCM

Tóm tắt:

Vấn đề giải quyết nhập nhằng ngữ nghĩa (word sense disambiguation) là một vấn đề cốt lõi của dịch máy nói riêng và xử lý ngôn ngữ tự nhiên nói chung. Giải quyết tốt vấn đề này sẽ giúp cho hệ dịch có chất lượng tốt hơn. Bên cạnh đó cũng góp phần giúp cho máy tính có khả năng hiểu được ngôn ngữ tự nhiên, trở nên gần gũi hơn với con người. Trong bài báo cáo này chúng tôi trình bày một phương pháp kết hợp các nguồn tri thức khác nhau để giải quyết nhập nhằng ngữ nghĩa cho từ tiếng Anh tập trung vào lớp từ mở: danh từ, động từ, tính từ và trạng từ. Kết quả này được dùng vào việc xử lý ngữ nghĩa cho hệ dịch tự động Anh-Việt.

Phương pháp chính được áp dụng trong bài báo là áp dụng thuật toán TBL (Transformation Based Learning) để xây dựng hệ khử nhập nhằng ngữ nghĩa. Các luật được học và rút luật từ một tập ngữ liệu huấn luyện có kích thước lớn (SemCor kích thước 778.587 từ). Sau đó áp dụng một mô hình tối ưu luật, hệ luật sau khi được tối ưu được dùng trong hệ thống xử lý nhập nhằng ngữ nghĩa. Kết quả chúng tôi đạt được độ chính xác 73.54% đối với hệ khử nhập nhằng theo phương pháp TBL và 80% đối với toàn bộ hệ thống xử lý nhập nhằng ngữ nghĩa kết hợp. Đây là một kết quả đáng khích lệ có ý nghĩa cho cho việc xây dựng một hệ dịch Anh - Việt.

COMBINE MULTIPLE KNOWLEDGE SOURCES FOR WORD SENSE DISAMBIGUATION IN THE ENGLISH - VIETNAMESE MACHINE TRANSLATION SYSTEM

Tran Le Hong Du, Dinh Dien, Nguyen Anh Tuan
Department of Information Technology, University of Natural Sciences - VNU.HCM

Abstract:

Word sense disambiguation (WSD) is not only a core problem in machine translation but also a destination of natural language processing. With a high accurate sense tagger, engaging an ability of translating better result. Moreover, it makes the computer can understand natural language, become more friendly to people. In this paper, we describe a method for combining multiple knowledge sources for word sense disambiguation and we focus on open word classes (noun, verb, adjective and adverb). The result of this report is used for semantic processing in English - Vietnamese translator.

The main method we describe in this paper is the Transformation-Based Learning algorithm (TBL) for contributing to a WSD system. An ordered rule set is extracted from a large training corpus (SemCor with about 800.000 word). This rules set is optimized by another module. The optimized rule set is used in WSD system. We achieved result with accurate 73.54% in sense tagger with TBL method and accurate 80% for whole WSD system which used multiple combined knowledge source. This is an encouraging result and bring a better English - Vietnamese translator.