Báo cáo kỹ thuật

Đề tài nhánh SP.74

Xây dựng kho ngư liệu song ngữ Anh – Việt

Ghi chú :

Báo cáo này bao gồm các báo cáo về nghiên cứu – thiết kế liệt kê trong phụ lục hợp đồng :

1. *Nghiên cứu nội dung các kho ngữ liệu song ngữ. SP: 1 báo cáo*
2. *Nghiên cứu tham khảo cấu trúc các kho ngữ liệu song ngữ. SP: 1 báo cáo*
3. *Thiết kế nội dung kho ngữ liệu câu Anh- Việt. SP: 1 báo cáo*
4. *Thiết kế cấu trúc cho kho ngữ liệu câu Anh- Việt. SP: 1 báo cáo*
5. *Thiết kế xây dựng khuôn dạng dữ liệu cho hai kho ngữ liệu câu Anh- Việt. SP: 1 báo cáo*
6. *Nghiên cứu tiêu chí chọn mẫu ngữ liệu song ngữ Anh-Việt. SP: 1 báo cáo*

Nhóm thực hiện

1. Hồ Bảo Quốc
2. Đinh Điền
3. Đặng Bác Văn
4. Lương Vỹ Minh
5. Phạm Đào Duy Vũ

Mục lục

I. Giới thiệu

I.1 Mục tiêu của đề tài nhánh

Trong tính toán ngôn ngữ học (linguistic computing) một tài nguyên rất cần thiết đó là các kho ngữ liệu song ngữ song song (parallel corpus). Các kho ngữ liệu song ngữ song song này có thể được sữ dụng cho nhiều mục tiêu khác nhau như : nghiên cứu ngôn ngữ học so sánh, tìm kiếm thông tin xuyên ngữ, dịch máy .v.v. Các kho ngữ liệu song ngữ này là nguồn tài nguyên để các ứng dụng có thể học các tương ứng của các đơn vị ngôn ngữ (từ, ngữ, câu, đoạn, văn bản ...) của hai ngôn ngữ, từ đó giải quyết các vấn đề liên quan. Kết quả của các bài toán trên phụ thuộc rất nhiều vào ***độ lớn và chất lượng*** của kho ngữ liệu song song được sử dụng. Trên thế giới đã có rất nhiều kho ngữ liệu song ngữ song song được xây dựng để phục vụ cho các mục tiêu như trên (xin xem chi tiết ở phần II). Hiện nay chưa có một kho ngữ liệu song song Anh - Việt được công bố chính thức và cho phép cộng đồng nghiên cứu liên quan đến có thể chia sẽ sử dụng cho các mục tiêu nghiên cứu. Do đó đề tài nhánh này nhằm nghiên cứu các cách tiếp cận xây dựng kho ngữ liệu song ngữ song song, cấu trúc và định dạng lưu trữ của các kho ngữ liệu song ngữ song song và các tiêu chí và phương pháp đánh giá một kho ngữ liệu song ngữ song song Anh – Việt. Trong khuôn khổ cho phép của kinh phí đề tà***i, mục tiêu của đề tài nhánh là xây dựng được một kho ngữ liệu song ngữ Anh – Việt song song gióng hàng đến mức câu (Sentence Aligment) gồm 100.000 cặp câu***