ĐẠI HỌC QUỐC GIA TP HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA KHOA HỌC VÀ KĨ THUẬT MÁY TÍNH

BÁO CÁO THỰC TẬP TỐT NGHIỆP

Phân giải đồng tham chiếu trên bệnh án điện tử

Giáo viên hướng dẫn:

Cao Hoàng Trụ

Sinh viên thực hiện:

Nguyễn Duy Hưng – 51101475
Vương Anh Tuấn – 51104040

15/05/2015

Mục lục

1 Giới thiệu vấn đề

- Giới thiệu về bệnh án điện tử và xu thế của nó trên thế giới
- Nêu lên vấn đề về trích xuất các kiến thức từ nguồn dữ liệu lớn như bệnh án điện tử
- Một trong những vấn đề liên quan là coref, tuy rất được quan tâm nghiên cứu cho các lĩnh vực khác nhưng cho lĩnh vực bệnh án điện tử thì vẫn chưa được xem xét tới
- Giới thiệu cụ thể bài toán: phân giải đồng tham chiếu trên bệnh án điện tử

2 Các công trình liên quan

2.1 Bệnh án điện tử

- Giới thiệu bệnh án điện tử là gì
- Đưa ví dụ

2.2 Phân giải đồng tham chiếu

- Giới thiệu bài toán coreference là gì
- Coreference resolution nói chung có 3 kiểu hệ thống
    1. Mention-pair model
    2. Entity-mention model
    3. Ranking model

2.3 Phân giải đồng tham chiếu cho bệnh án điện tử

- Coreference cho văn bản y khoa, cụ thể là bệnh án điện tử
- Có 3 hướng tiếp cận:
    1. Rule-based learning system
    2. Supervised learning system
    3. Hybrid system

3 Kiến thức và công nghệ

3.1 Named-Entity-Regconition

- Giải thích, giới thiệu, đưa ví dụ về NER

3.2 Những vấn đề trong phân giải đồng tham chiếu trong bệnh án điện tử

- Đưa ra các key observation trong bài báo
- Đưa ra các feature design
- Nói rõ về coreference là gì và phân giải nó là như thế nào

4 Bài toán đề xuất

4.1 Phạm vi đề tài

4.1.1 Nội dung bài toán

- Xây dựng hệ thống phân giải đồng tham chiếu trên các bệnh án điện tử với các thực thể đã được cho biết trước

4.1.2 Dữ liệu đầu vào

- Là các bệnh án điện tử cùng với danh sách các thực thể đã được gán nhãn có trong bệnh án đó theo một định dạng nhất định
- Giải thích rõ đầu vào
- Đưa ra các ví dụ thực thể trong bệnh án

4.1.3 Kết quả đầu ra

- Là chuỗi đồng tham chiếu các thực thể và nhãn cho chuỗi đó
- Giải thích rõ kết quả
- Đưa ra ví dụ kết quả mong muốn

4.2 Thiết kế hệ thống

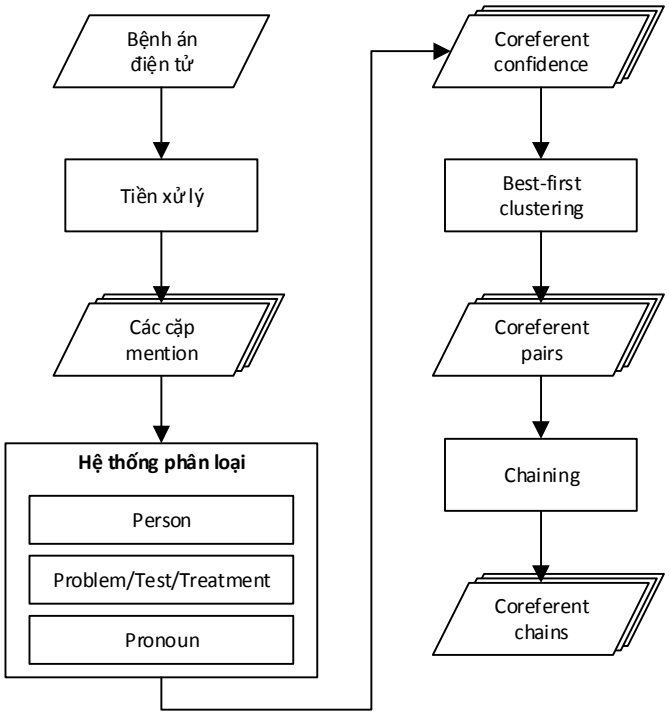
4.2.1 Định nghĩa nhãn

- Định nghĩa 5 nhãn Person, Problem, Test, Treatment, Pronoun
- Đưa ra ví dụ về các thực thể và nhãn trong một bệnh án cụ thể

4.2.2 Chi tiết hệ thống

Hệ thống sẽ gồm các bước (Hình 1)

- Tiền xử lý (kết quả là các cặp thực thể có khả năng đồng tham chiếu)
- Học máy có giám sát sử dụng 3 module riêng biệt để phân loại 3 nhóm nhãn thực thể (kết quả là độ tin cậy việc đồng tham chiếu của căp thực thể)
- Áp dụng giải thuật best-first clustering (kết quả là các cặp thực thể đã được xác định là đồng tham chiếu)
- Xây dựng chuỗi đồng tham chiếu



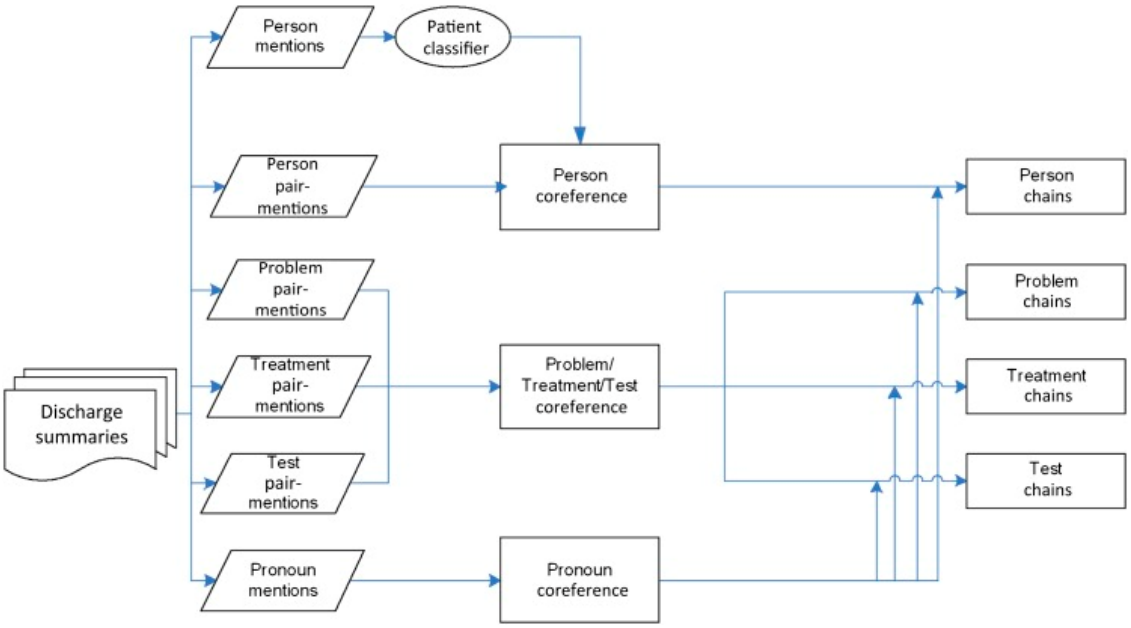
Hình 1. Sơ đồ khối

4.2.3 Tiền xử lý

- Giải thích bước tiền xử lý
- Đưa ví dụ: "her CT scan" và "a CT scan" sau khi được tiền xử lý đều trở thành "CT scan"
- Cặp mention được xây dựng là lọc từ $C(n, 2)$ các cặp mention.

4.2.4 Học máy có giám sát

- Sử dụng 3 module riêng biệt ứng với 5 nhãn
    1. Module Person (thêm vào thuộc tính patient hoặc family hoặc hospital person)
    2. Module Non-person (Problem – Test – Treatment)
    3. Module Pronoun



Hình 2. Sơ đồ khối

4.2.5 Best-first clustering

- Giải thích thuật toán best-first clustering

4.2.6 Xây dựng chuỗi đồng tham chiếu

- Ghép các cặp thực thể đồng tham chiếu để xây dựng chuỗi đồng tham chiếu

5 Tập dữ liệu và phương pháp đánh giá

5.1 Tập dữ liệu

- Nói về bộ dữ liệu i2b2/VA
- Quy trình lấy và cam kết bảo mật dữ liệu
- Số lượng mẫu trong từng tập (training và test)

5.2 Phương pháp đánh giá

- Sử dụng 3 độ đo: F-measure, Precision và Recall
- Tính 3 độ đo trên theo ba cách khác nhau: MUC, B-CUBED và CEAF (giải thích kĩ), sau đó lấy trung bình không trọng số
- Kết quả đánh giá cuối cùng là trung bình không trọng số của 3 độ đo trên

6 Kết luận

7 Tài liệu tham khảo

[1] Y. Xu, J. Liu, J. Wu, Y. Wang, Z. Tu, J.-T. Sun, J. Tsujii and E. I-Chao, "A classification approach to coreference in discharge summaries: 2011 i2b2 challenge," *Journal of the American Medical Informatics Association : JAMIA,* vol. 19, no. 5, pp. 897-905, 2012.