# *Xây dựng Dữ liệu cho Phân tích Sự kiện Mở trên Web*

Nguyễn Kiêm Hiếu

Đại học Bách khoa Hà nội

KDE Lab – 4/2016

## Nội dung

- Sự kiện là gì?
- Tập dữ liệu MUC-4
- Xây dựng tập dữ liệu ASTRE
- Thí nghiệm đánh giá



SỰ KIỆN LÀ GÌ?



## Sự kiện đơn

- Một hành động xảy ra trong một thời gian cụ thể, không gian cụ thể
- Thành phần của sự kiện:
  - Bản thân hành động
  - Chủ thể của hành động
  - Đối tượng của hành động
  - Ngữ cảnh: Thời gian, không gian
  - Các thành phần khác tùy vào sự kiện cụ thể

# Sự kiện đơn (tiếp)

Ví dụ

- “Hàng ngày tôi đi tới trường bằng xe bus”
  - Hành động: đi
  - Chủ thể: tôi
  - Thời gian: hàng ngày
  - Đích: trường
  - Phương tiện: xe bus



# Sự kiện đơn (tiếp)

Ví dụ

- “AlphaGo đánh bại Lee Se-dol với tỉ số 4-1”
  - Hành động: đánh bại
  - Chủ thể: AlphaGo
  - Đối tượng: Lee Se-dol
  - Tỉ số: 4-1



# Sự kiện đơn (tiếp)

- Biểu diễn sự kiện:
  - Hành động:
    - Động từ: đi/đánh bại…
    - Danh từ chỉ sự kiện: cuộc thi, buổi cắm trại…
  - Chủ thế/đối tượng/thuộc tính: thực thể
    - Danh từ: trường, xe bus
    - Thực thể có tên: AlphaGo, Lee Se-dol



# Sự kiện phức hợp (tiếp)

- Sự kiện phức hợp: bao gồm một tập hợp các sự kiện có liên quan chặt chẽ với nhau theo quan hệ trình tự hay quan hệ nhân-quả
- Hệ quả: Chủ thể của sự kiện này có thể là đối tượng của sự kiện khác

# Sự kiện phức hợp (tiếp)

**Law & Justice Events:**

- LIFE {Injure, Die}
- CONFLICT {Attack}
- JUSTICE {Charge-Indict, Arrest-Jail, Release-Parole, Sentence, Convict, Appeal, Acquit, Execute, Extradite}



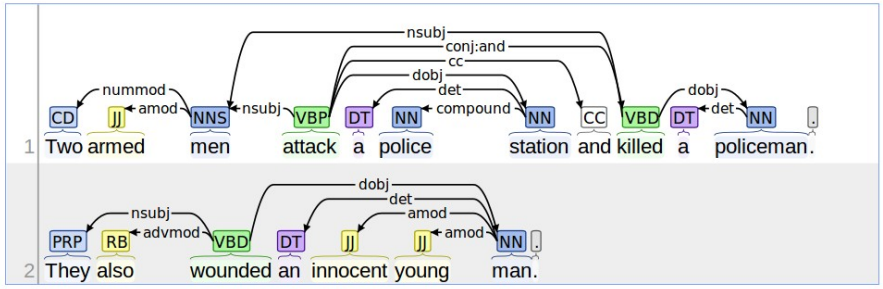
# Sự kiện phức hợp (tiếp)

Ví dụ

- Sự kiện đánh bom
  - Thủ phạm: chủ thể của hành động đánh bom/nhận trách nhiệm/phá hủy/bị tiêu diệt, đối tượng của hành động điều tra...
  - Mục tiêu: đối tượng của hành động đánh bom, chủ thể của hành động bị phá hủy...
  - Nạn nhân: chủ thể của hành động bị thương/chết...
  - Phương tiện: đối tượng của hành động sử dụng, chủ thể của hành động phát nổ/phá hủy...



# Biểu diễn sự kiện hướng thực thể

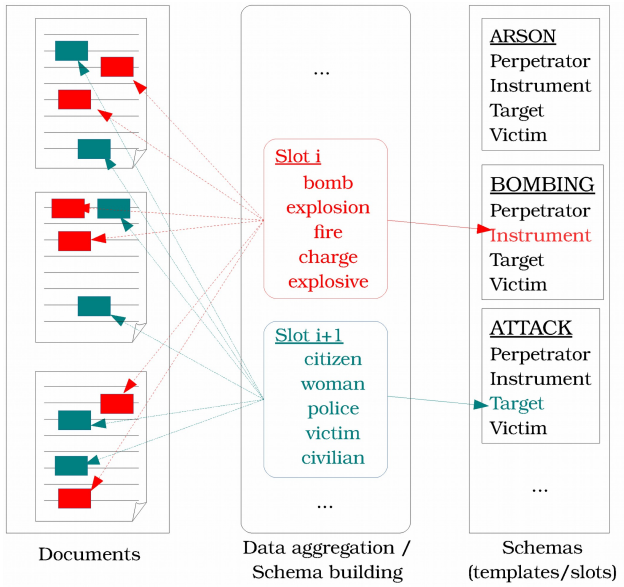



| | Attributes | Head | Triggers |
|---|---|---|---|
| #1 | [armed:amod] | man | [attack:nsubj, kill:nsubj, wound:nsubj] |
| #2 | [police:nn] | station | [attack:dobj] |
| #3 | [] | policeman | [kill:dobj] |
| #4 | [innocent:amod, young:amod] | man | [wound:dobj] |



# Hai bài toán

- Bài toán 1: Xác định cấu trúc sự kiện phức hợp
- Bài toán 2: Liệt kê các sự kiện phức hợp được miêu tả trong văn bản theo cấu trúc đã xác định






# Phương pháp tiếp cận

- Giải quyết đồng thời hai bài toán
- Sử dụng công cụ NLP để tạo ra cách biểu diễn mạnh và giàu thông tin cho sự kiện
- Sử dụng mô hình sinh để học cấu trúc sự kiện thông qua suy diễn và liệt kê sự kiện thông qua gán nhãn



TẬP DỮ LIỆU MUC-4



# Tập dữ liệu MUC-4

- 1300 tài liệu
- Các vụ tấn công khủng bố ở Nam Mỹ trong giai đoạn 1980-1990
- Chuyển thể thành văn bản từ các bản tin radio
- Được sử dụng làm tập dữ liệu chuẩn để đánh giá các hệ thống phân tích sự kiện

# Tập dữ liệu MUC-4 (tiếp)

- Tập sự kiện
  - TERRORIST{Attack, Bombing, Kidnapping, Arson}
- Tập các thành phần
  - Perpetrator
  - Victim
  - Instrument
  - Target



# Tập dữ liệu MUC-4 (tiếp)

- Hạn chế:
  - Không có thông tin dư thừa
  - Các sự kiện khác nhau có cùng các thành phần



XÂY DỰNG TẬP DỮ LIỆU ASTRE

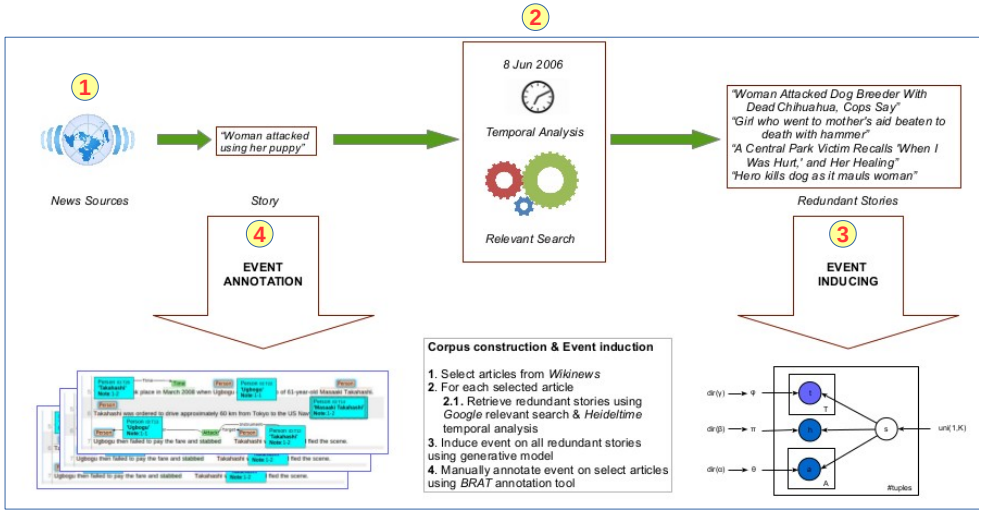
*1. LỰA CHỌN NGUỒN*

*2. THU THẬP SỰ KIỆN TRÙNG LẶP*

*3. HỌC CẤU TRÚC SỰ KIỆN*

*4. GÁN NHÃN SỰ KIỆN*





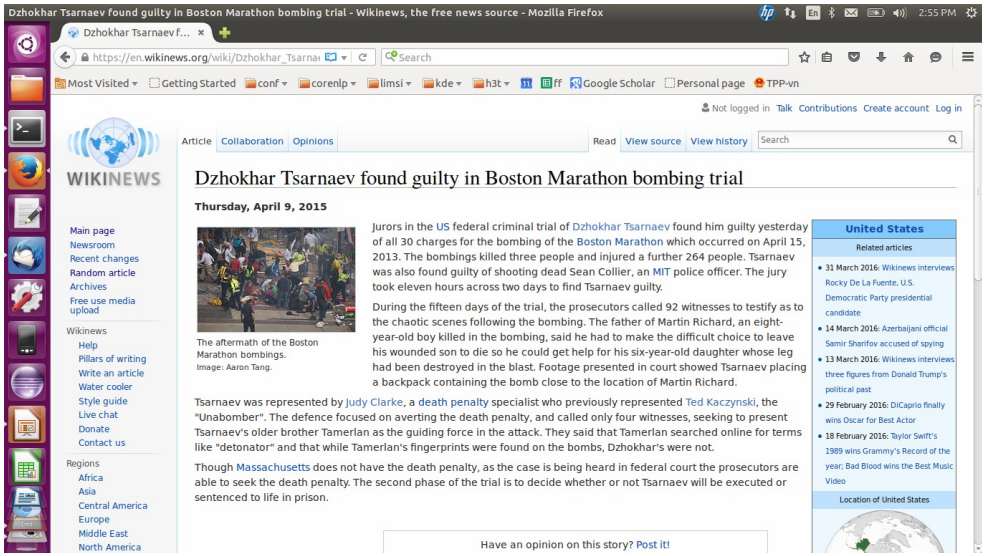
# 1-Lựa chọn nguồn

- Thể loại tin tức
- Đa ngôn ngữ (trước hết là tiếng Anh)
- Cập nhật, đa dạng

  → Wikinews



https://en.wikinews.org/wiki/Dzhokhar_Tsarnaev_found_guilty_in_Boston_Marathon_bombing_trial






# 2-Thu thập sự kiện trùng lặp

2.1 Trích xuất ra ngày xuất bản (DCT) của bài báo miêu tả sự kiện

2.2 Trích xuất ra tiêu đề của bài báo

2.3 Đưa tiêu đề bài báo làm câu truy vấn tới một bộ máy tìm kiếm Web

2.4 Trích xuất ra ngày xuất bản của các kết quả tìm kiếm.

2.5 Lấy các kết quả tìm kiếm nằm trong khoảng [DCT-delta, DCT+delta]



# 3-Học cấu trúc sự kiện

- Biểu diễn các sự kiện con dưới dạng (sự kiện, đối tượng tham gia)
- Sử dụng công cụ phân tích cú pháp phụ thuộc trong NLP để trích xuất ra cách biểu diễn này
- Sử dụng hiểu biết trong WordNet để lọc các thành phần cú pháp không phải là sự kiện

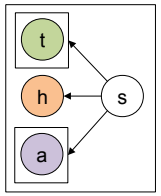
# 3-Học cấu trúc sự kiện (tiếp)

- Sử dụng mô hình sinh để học cấu trúc sự kiện
- Học theo lô sử dụng phương pháp lấy mẫu Gibbs



# Học cấu trúc sự kiện (tiếp)

Model




Induced slot: **ATTACK_victim**

| Attributes pr(a\|s) | Head *pr(h\|s)* | Triggers *pr(t\|s)* |
|---|---|---|
| *innocent:amod* | *citizen* | *murder:dobj (+)* |
| *wounded:amod* | *woman* | *kill:dobj (-)* |
| *U.N:nn* | *police* | *die:nsubj* |
| *young:amod* | *victim (+)* | *die:prep_of* |
| *official:nn* | *civilian (+)* | *assassinate:dobj (+)* |

Slot Assignment

Municipal **official** **Sergio Horna** was seriously **wounded**.

Two **extremist** **terrorists** were reportedly **killed** by national officers.



# 4-Gán nhãn sự kiện

- Xác định trước cấu trúc sự kiện dựa trên cuộc thi TAC KBP 2014
- Sử dụng công cụ gán nhãn BRAT
- Gán nhãn các thực thể là các thành phần tham gia sự kiện
- Gán nhãn đồng tham chiếu (nếu có) của các thực thể này



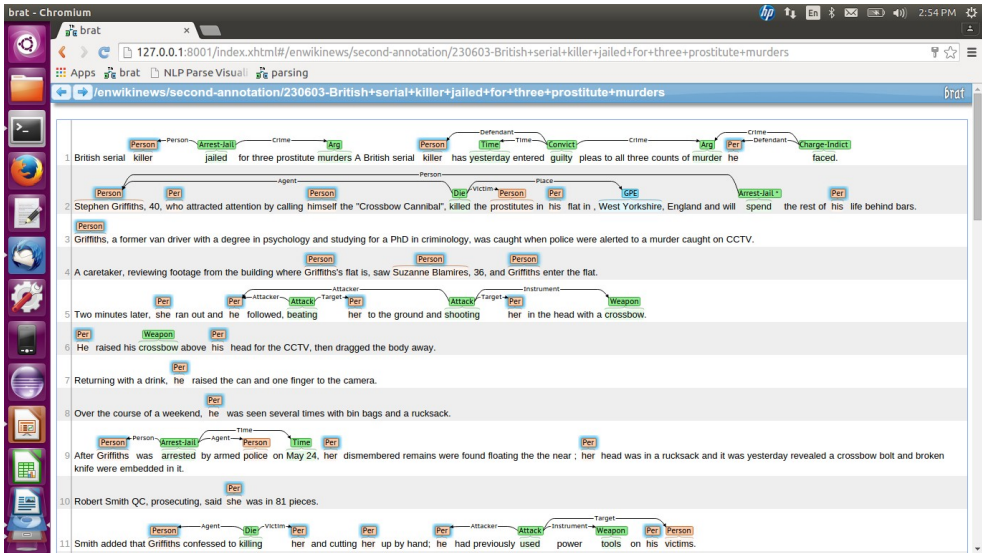
# 4-Gán nhãn sự kiện (tiếp)

**Law & Justice Events:**

- LIFE {Injure, Die}
  → Agent, victim, instrument
- CONFLICT {Attack}
  → Attacker, target, instrument
- JUSTICE {Charge-Indict, Arrest-Jail, Release-Parole, Sentence, Convict, Appeal, Acquit, Execute, Extradite}
  → Agent, person, crime, adjudicator, defendant, sentence

- Theme
- Event
- Role

http://127.0.0.1:8001/index.xhtml#/enwikinews/second-annotation/230603-British+serial+killer+jailed+for+three+prostitute+murders






THÍ NGHIỆM ĐÁNH GIÁ



- Vấn đề:
  - Học cấu trúc sự kiện
  - Gán nhãn thành phần sự kiện
- Tập dự liệu:
  - MUC-4: 1,300 tài liệu (1980-1990)
  - ASTRE: 1,083 tài liệu (2000-hiện nay)
- Phương pháp sử dụng mô hình sinh:
  - Chambers (2013)
  - Nguyen et al (2015)



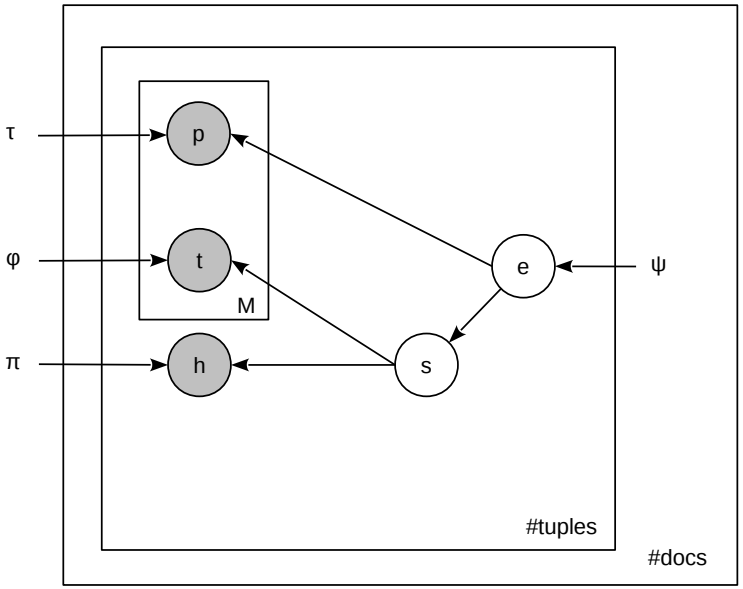



Figure: Probabilistic schema induction in Chambers (2013)

BOMBING_instrument

| Attributes | Heads | Triggers |
|---|---|---|
| car:nn | bomb | explode:nsubj |
| powerful:amod | fire | hear:dobj |
| explosive:amod | explosion | place:dobj |
| dynamite:nn | blow | cause:nsubj |
| heavy:amod | charge | set:dobj |

KIDNAPPING_victim

| Attributes | Heads | Triggers |
|---|---|---|
| several:amod | people | arrest:dobj |
| other:amod | person | kidnap:dobj |
| responsible:amod | man | release:dobj |
| military:amod | member | kill:dobj |
| young:amod | leader | identify:prep_as |

Figure 4: Attribute, head and trigger distributions learned by the model *HT+A* for learned slots that were mapped to BOMBING_instrument and KIDNAPPING_victim.



- Phép đo: Precision (độ chính xác), recall (độ bao phủ), F-score
- Tính kết qủa trung bình của 10 lần lấy mẫu
- Sử dụng sourcode của Nguyen et al (2015) và cài đặt lại mô hình của Chambers (2013)

| System | Dev score | | | Test score | | |
|---|---|---|---|---|---|---|
| | P | R | F | P | R | F |
| (Chambers, 2013) | 36 | 36 | 36 | 16 | 29 | 20 |
| (Nguyen et al., 2015) | 42 | 30 | 35 | 21 | 25 | 23 |



## KẾT LUẬN:

- tăng tính đa dạng của sự kiện
- bổ sung thông tin dư thừa
- có thể sử dụng để đánh giá việc tìm kiếm thông tin
- chưa giải quyết được việc mapping tự động giữa chủ đề học được với thành phần đã được gán nhãn



# Tài liệu


- Tập dữ liệu ASTRE, mã nguồn phương pháp suy diễn, các bài báo liên quan:

  Trang cá nhân: http://is.hust.edu.vn/~hieunk

  Email: *hieunk@soict.hust.edu.vn*
- Kiem-Hieu Nguyen et al. *A Dataset for Open Event Extraction in English*. LREC 2016
- Kiem-Hieu Nguyen et al. *Generative Event Schema Induction with Entity Disambiguation*. ACL 2015

# Một vài hướng nghiên cứu liên quan

- Thiết kế mô hình chủ đề hướng thời gian
  - Áp dụng vào hệ thống tóm tắt đa văn bản hướng câu truy vấn
  - Áp dụng vào hệ thống xây dựng dòng sự kiện
- Hệ thống phân tích quan điểm hướng thực thể dựa trên xử lý ngôn ngữ tự nhiên + mô hình chủ đề



HỎI - ĐÁP