RÚT TRÍCH THÔNG TIN

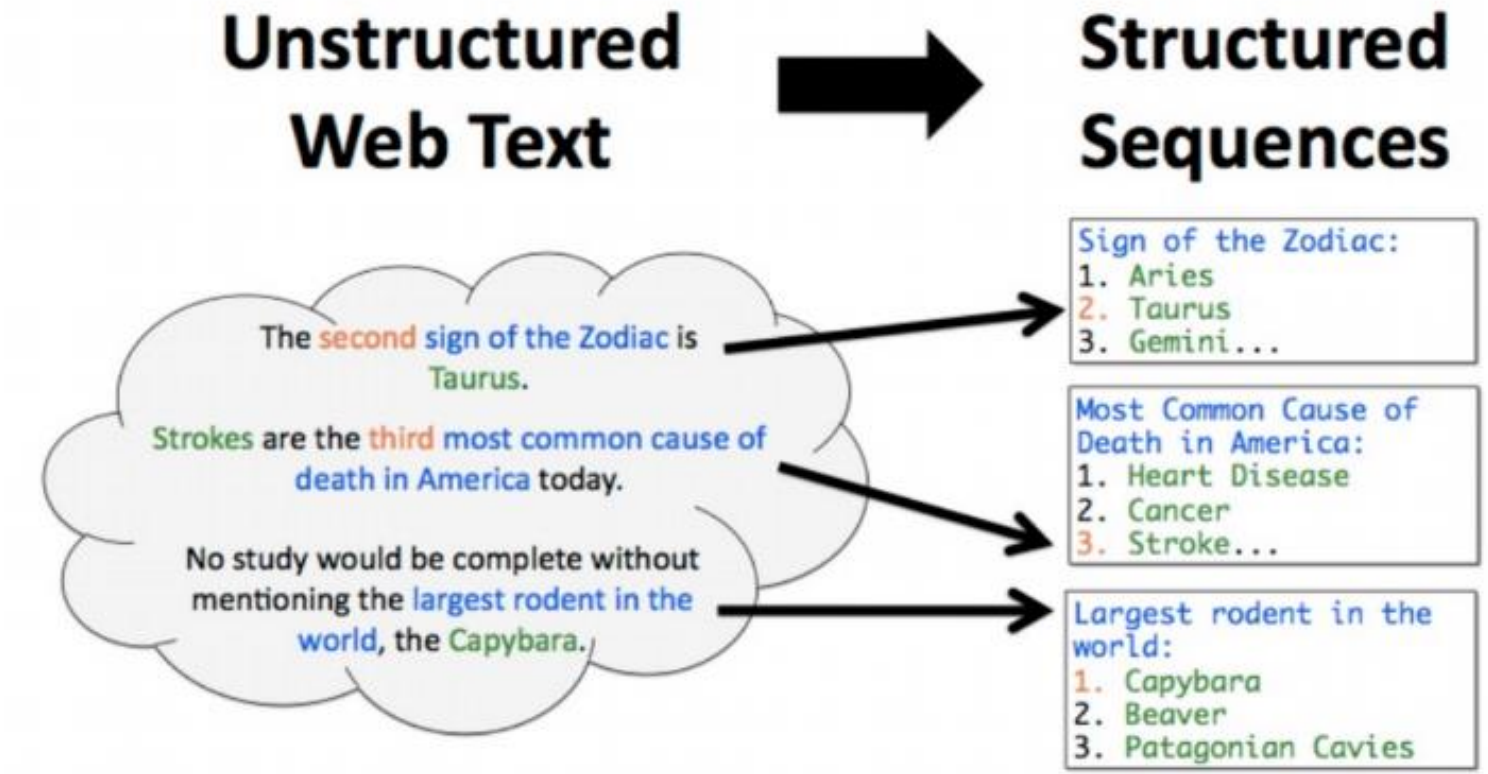
Information Extraction

NỘI DUNG

1. Rút trích thông tin
2. Các tác vụ sử dụng rút trích thông tin
3. Phương pháp tiếp cận đối với rút trích thông tin
4. Rút trích quan hệ (relation extraction)
5. Rút trích thông tin thời gian (time extraction)

Định nghĩa

- Rút trích thông tin là quá trình chuyển các thông tin có trong **dữ liệu phi cấu trúc** (unstructured data) thành **dữ liệu có cấu trúc** (structured data).



Ví dụ về rút trích thông tin

To: dbworld@…

There will be a summer school "Communication Technology and Data Analytics for Future Energy Systems" from September 11 to September 15, 2017, at the University of Passau, Germany.

Type = summer_school

Title = “Comm…Systems”

Start_date = 2017-09-11

End_date = 2017-09-15

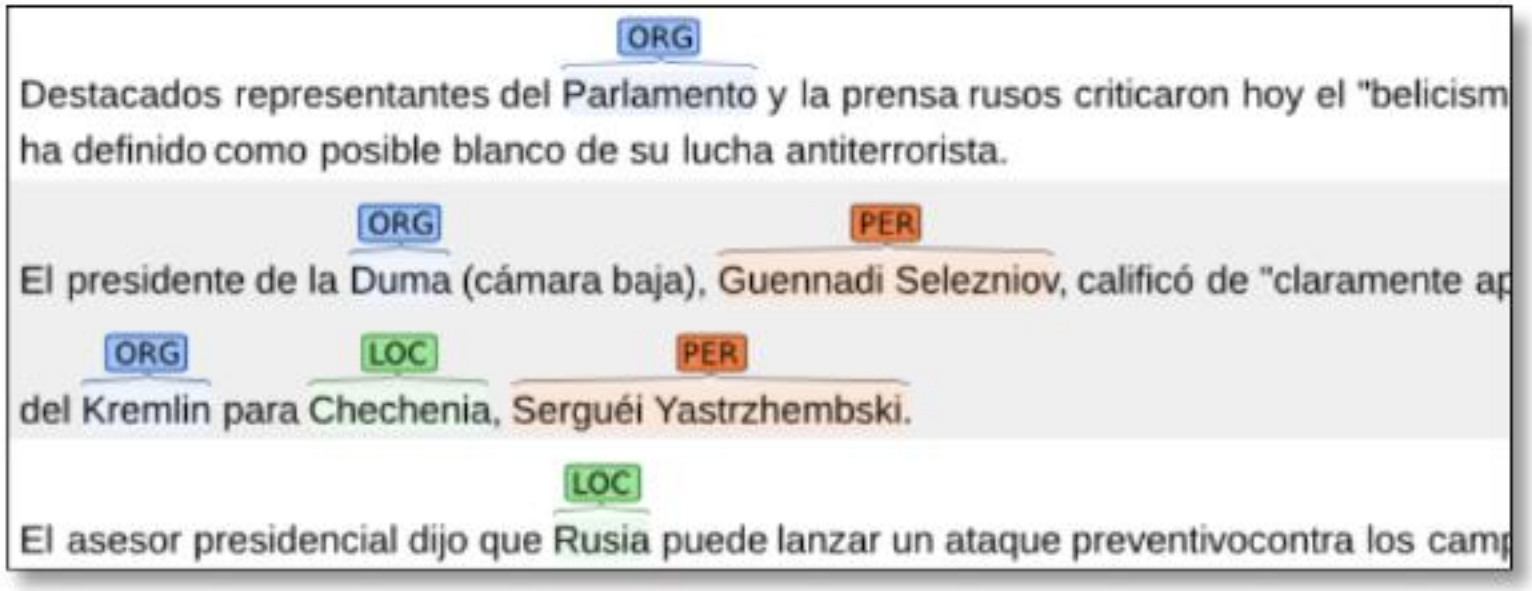
Institution = Univ…Passau

Location = Germany

Các tác vụ về rút trích thông tin

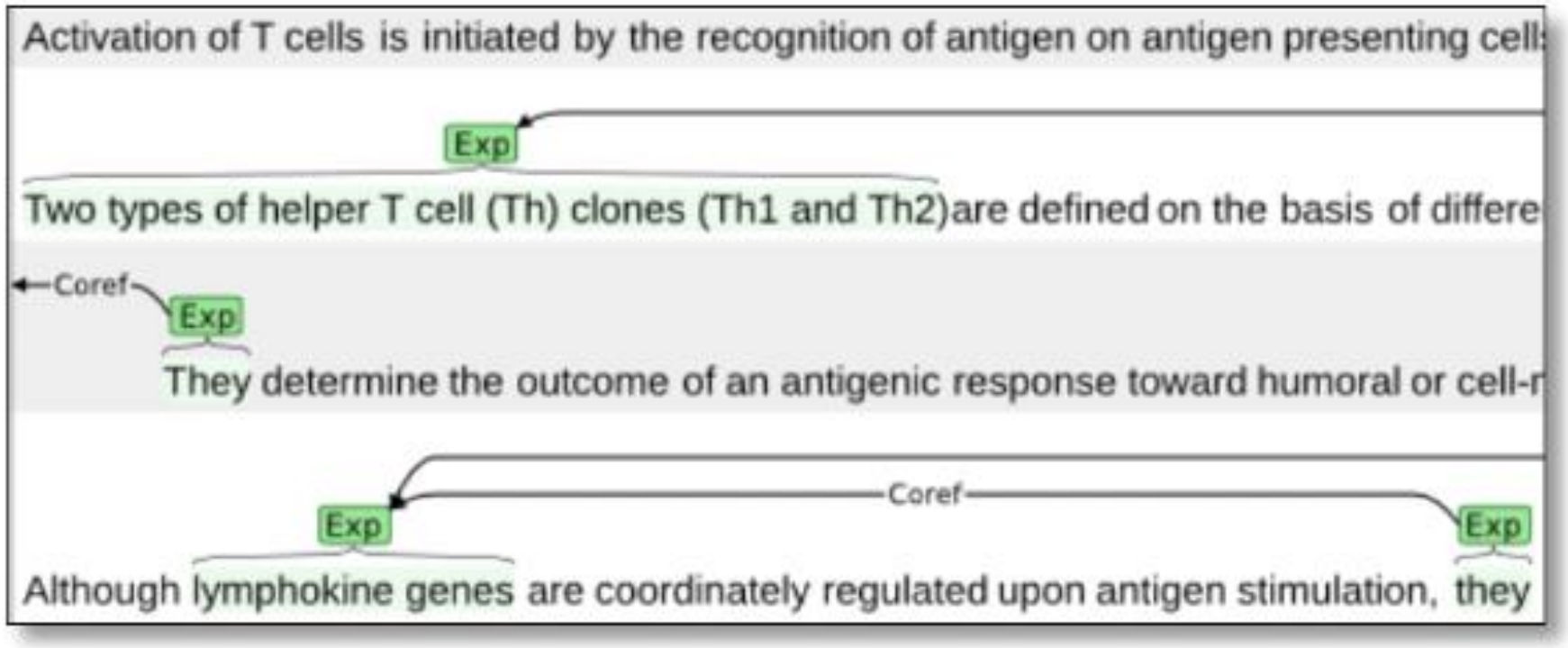
- Rút trích quan hệ (Relationship extraction)
- Rút trích thời gian (Time extraction)
- Nhận diện tên riêng (Named entity recognition)
- Rút trích mẫu thông tin có cấu trúc (Structured record extraction)
- Phân giải đồng tham chiếu (Co-reference resolution)

Đề cập đến thực thể (Entity mention)



Tjong Kim Sang, Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition (2002)

Phân giải đồng tham chiếu (Co-reference resolution)



Nguyen et al., Overview of the Protein Coreference task in BioNLP Shared Task 2011 (2011)

Cách tiếp cận rút trích thông tin

- Dựa trên luật (rule-based): bao gồm biểu thức chính quy (regular expression)
- Dựa trên học máy (machine learning)

Ví dụ: Rút trích email

Biểu thức chính quy: ***([\w\.\-_]+)?\w+@[\w-_]+(\.\w+){1,}***

```
Welcome to RegExr v2.0 by gskinner.com!
¬
Edit the Expression & Text to see matches. Roll over matches or the expression for details. Undo mistakes
with ctrl-z. Save & Share expressions with friends or the Community. A full Reference & Help is available
in the Library, or watch the video Tutorial.
¬
Sample text for testing:¬
abcdefghijklmnopqrstuvwxyz ABCDEFGHIJKLMNOPQRSTUVWXYZ¬
0123456789 +-.,!@#$%^&*();\\/|<>\"\'¬
12345 -98.7 3.141 .6180 9,000 +42¬
555.123.4567⟶+1-(800)-555-2468¬
fo-o@j_demo.net⟶bar.ba@test.co.uk.com.hk.finish.this.is.not.the.end.of.this.domain¬
www.demo.com⟶http://foo.co.uk/¬
http://regexr.com/foo.html?q=bar
```

Thông tin rút trích được:

fo-o@j_demo.net

bar.ba@test.co.uk.com.hk.finish.this.is.not.the.end.of.this.domain

Rút trích thông tin ngày tháng từ văn bản

Cho đoạn văn bản sau, hãy rút trích ra thông tin về các ngày tháng trong đoạn văn bản trên bằng cách sử dụng RegExp:

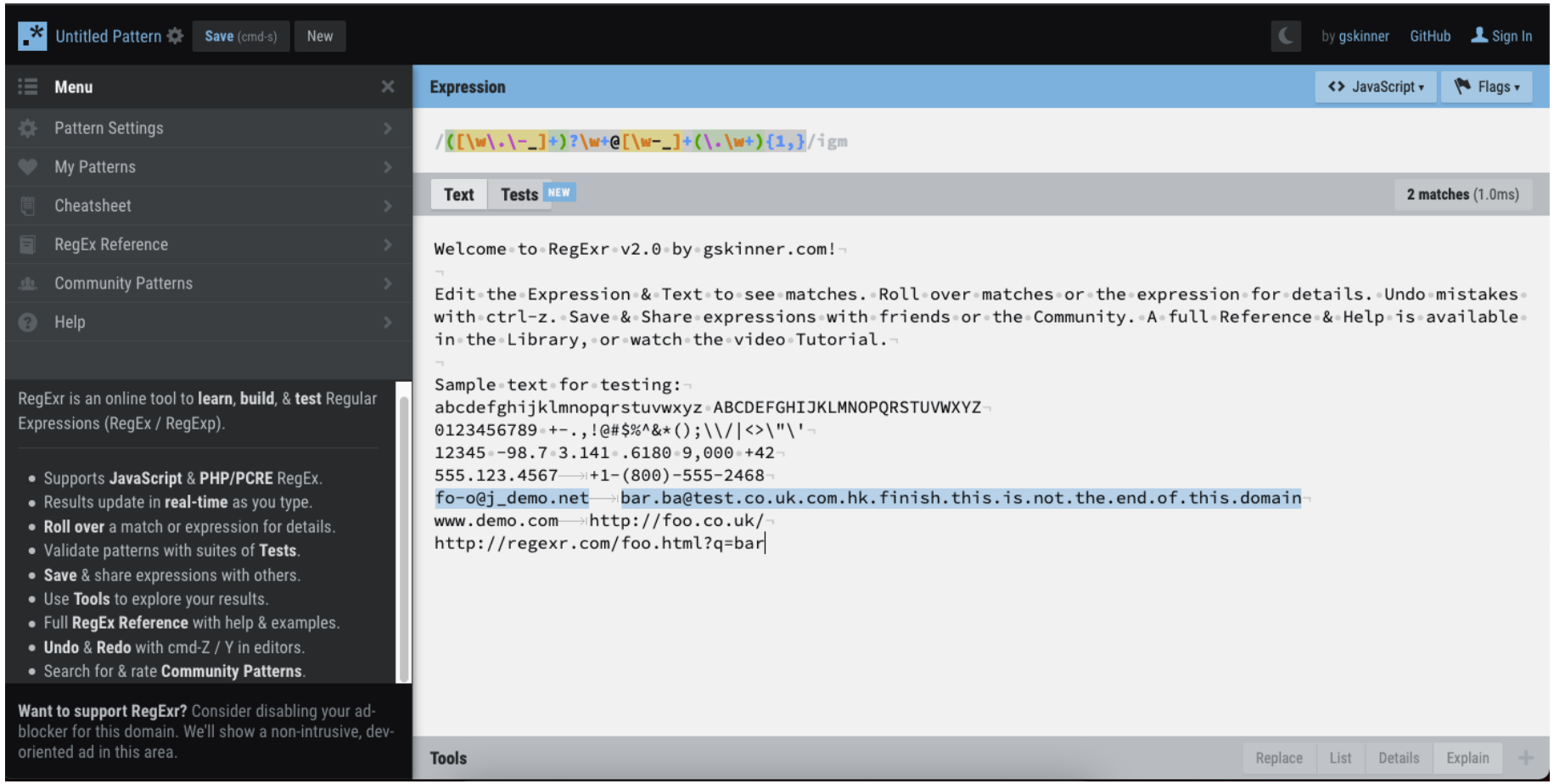
Trước diễn biến phức tạp của dịch bệnh covid 19, thư viện Trung tâm Đại học Quốc Gia TP Hồ Chí Minh vừa ra thông báo tạm ngưng hoạt động.

Theo đó, chi nhánh KTX B tạm ngưng hoạt động từ ngày 11/5/2021 đến khi có thông báo mới. Thư viện Trung tâm cơ sở chính hoạt động bình thường, vẫn phục vụ truy cập trực tuyến và hỗ trợ đọc giả từ xa qua điện thoại, email, facebook.

Bên cạnh đó, tất cả sách có hạn trả từ 10/5/2021 sẽ được gia hạn đến ngày 30/6/2021. Thư viện nhận trả sách ở Thư viện Trung tâm trụ sở chính, khu phố 6, Linh Trung, Thủ Đức.

https://www.uit.edu.vn/thu-vien-trung-tam-dai-hoc-quoc-gia-tp-ho-chi-minh-thong-bao-tam-ngung-hoat-dong

RegExr: Công cụ sử dụng để xây dựng biểu thức chính quy



https://regexr.com/

Rule-based – Xpath

- Là ngôn ngữ dùng để truy vấn trên tài liệu XML và HTML.
- Có thể dùng Xpath để trích xuất các thông tin từ các tài liệu XML và HTML

Ví dụ: trích xuất thông tin từ XML

```
<MiBibliotecaMP3>
 <archivo>
  <canción>Hangar 18</canción>
  <artista cantante="Dave Mustaine">Megadeth</artista>
  <disco discográfica="capitol" año="1990">Rust in Peace</disco>
  <puntuación>9</puntuación>
 </archivo>
 <archivo>
  <canción>Peace Sells</canción>
  <artista cantante="Dave Mustaine">Megadeth</artista>
  <disco discográfica="capitol" año="1986">Peace Sells...But Who's Buying</disco>
  <puntuación>9</puntuación>
 </archivo>
 <archivo>
  <canción>Master of Puppets</canción>
  <artista cantante="James Hetfield">Metallica</artista>
  <disco discográfica="Elektra" año="1986">Master of Puppets</disco>
  <puntuacion>10</puntuacion>
 </archivo>
 <archivo>
  <canción>Among The Living</canción>
  <artista cantante="Joey Belladonna">Anthrax</artista>
  <disco discográfica="Megaforce" año="1987">Among The Living</disco>
  <puntuacion>8</puntuacion>
 </archivo>
</MiBibliotecaMP3>
```

/MiBibliotecaMP3/archivo/disco

<disco discográfica=”capitol” año=”1990”>Rust in Peace</disco>

Rút trích quan hệ (Relation extraction)

Rút trích quan hệ (Relation extraction)

- Rút trích quan hệ (relation extraction): nhằm tìm ra và phân loại các *quan hệ ngữ nghĩa* (semantic relation) giữa các thực thể trong văn bản.
  - **Input**: văn bản
  - **Output**: thực thể và mối quan hệ giữa chúng.
- Các quan hệ này thường là các quan hệ 2 ngôi như:
  - child-of
  - employment
  - part-whole
  - geospatial relation (là quan hệ kết hợp giữa thông tin vị trí địa lý, sự kiện, thời gian: động đất, …)
- **Knowledge graphs** (tạm dịch là đồ thị tri thức) là công cụ dùng để mô tả các tri thức có mối quan hệ cấu trúc với nhau.
  - Cấu trúc ở đây có thể là cấu trúc đồ thị (graph-structured) hoặc là dạng topology.

Một số quan hệ khác trong văn bản

Relations	Types	Examples
Physical-Located	PER-GPE	**He** was in **Tennessee**
Part-Whole-Subsidiary	ORG-ORG	**XYZ**, the parent company of **ABC**
Person-Social-Family	PER-PER	**Yoko**'s husband **John**
Org-AFF-Founder	PER-ORG	**Steve Jobs**, co-founder of **Apple**...

VD: rút trích mối quan hệ giữa các thực thể có tên trong văn bản

Citing high fuel prices, [ORG **United Airlines**] said [TIME **Friday**] it has increased fares by [MONEY **$6**] per round trip on flights to some cities also served by lower-cost carriers. [ORG **American Airlines**], a unit of [ORG **AMR Corp.**], immediately matched the move, spokesman [PER **Tim Wagner**] said. [ORG **United**], a unit of [ORG **UAL Corp.**], said the increase took effect [TIME **Thursday**] and applies to most routes where it competes against discount carriers, such as [LOC **Chicago**] to [LOC **Dallas**] and [LOC **Denver**] to [LOC **San Francisco**].

Quan hệ 1: ***Tim Wagner*** *is a spokesman for* ***American Airlines***
Quan hệ 2: ***United*** *is a unit of* ***UAL Corp***
Quan hệ 3: ***American*** *is a unit of* ***AMR***

Mô hình hoá việc rút trích thông tin từ văn bản

- Mô hình gồm 3 thành phần: ***(Domain, Classes, Relations)***
  - *Domain*: Tập hợp các thực thể trong văn bản
  - *Classes*: Các loại thực thể
  - *Relations*: Các loại quan hệ

Mô hình hoá việc rút trích thông tin từ văn bản (Ví dụ)

Citing high fuel prices, [ORG **United Airlines**] said [TIME **Friday**] it has increased fares by [MONEY **$6**] per round trip on flights to some cities also served by lower-cost carriers. [ORG **American Airlines**], a unit of [ORG **AMR Corp.**], immediately matched the move, spokesman [PER **Tim Wagner**] said. [ORG **United**], a unit of [ORG **UAL Corp.**], said the increase took effect [TIME **Thursday**] and applies to most routes where it competes against discount carriers, such as [LOC **Chicago**] to [LOC **Dallas**] and [LOC **Denver**] to [LOC **San Francisco**].

Domain	$\mathcal{D} = \{a,b,c,d,e,f,g,h,i\}$
United, UAL, American Airlines, AMR	a,b,c,d
Tim Wagner	e
Chicago, Dallas, Denver, and San Francisco	f,g,h,i
Classes	
United, UAL, American, and AMR are organizations	$Org = \{a,b,c,d\}$
Tim Wagner is a person	$Pers = \{e\}$
Chicago, Dallas, Denver, and San Francisco are places	$Loc = \{f,g,h,i\}$
Relations	
United is a unit of UAL	$PartOf = \{\langle a,b\rangle, \langle c,d\rangle\}$
American is a unit of AMR	
Tim Wagner works for American Airlines	$OrgAff = \{\langle c,e\rangle\}$
United serves Chicago, Dallas, Denver, and San Francisco	$Serves = \{\langle a,f\rangle, \langle a,g\rangle, \langle a,h\rangle, \langle a,i\rangle\}$

Một số bộ dữ liệu để nghiên cứu bài toán Rút trích quan hệ

- The TAC Relation Extraction Dataset (TACRED Dataset).
  - https://nlp.stanford.edu/projects/tacred/
- SemEval-2010 Task 8 Dataset.
  - https://www.aclweb.org/anthology/S10-1006/

Ví dụ: bộ dữ liệu TACRED

Example	Entity Types & Label
Carey will succeed Cathleen P. Black, who held the position for 15 years and will take on a new role as chairwoman of Hearst Magazines, the company said.	PERSON/TITLE Relation: *per:title*
Irene Morgan Kirkaldy, who was born and reared in Baltimore, lived on Long Island and ran a child-care center in Queens with her second husband, Stanley Kirkaldy.	PERSON/CITY Relation: *per:city_of_birth*
Baldwin declined further comment, and said JetBlue chief executive Dave Barger was unavailable.	Types: PERSON/TITLE Relation: *no_relation*

https://nlp.stanford.edu/pubs/zhang2017tacred.pdf

Cách tiếp cận (approach) bài toán rút trích quan hệ

- Dựa theo mẫu (pattern)
- Học có giám sát
- Học bán giám sát
- Distant supervision (kết hợp học có giám sát và bán giám sát)
- Cách tiếp cận theo học không giám sát

Tiếp cận dựa theo mẫu (pattern)

Hearst patterns: sử dụng **lexico-syntactic pattern để rút quan hệ hyponym (hạ vị)**

VD: rút trích thông tin về mối quan hệ của "*Gelidium*" trong câu sau:

Agar is a substance prepared from a mixture of red algae, such as Gelidium, *for laboratory or industrial use*

Sử dụng **lexico-syntactic pattern như sau:**

NP_0 such as {NP_1, NP_2..., (and|or) NP_i}, $i \geq 1$

$\forall NP_i$, $i \geq 1$, *hyponym*(NP_i,NP_0)

hyponym(Gelidium, red algae)

hyponym of

Kết quả:

Agar is a substance prepared from a mixture of red algae, such as Gelidium, *for laboratory or industrial use*

https://www.aclweb.org/anthology/P18-2057/

Một số ví dụ về lexico-syntactic pattern (Hearst patterns)

NP {, NP}* {,} (and|or) other NP_H	temples, treasuries, and other important civic buildings
NP_H such as {NP,}* {(or|and)} NP	red algae such as Gelidium
such NP_H as {NP,}* {(or|and)} NP	such authors as Herrick, Goldsmith, and Shakespeare
NP_H {,} including {NP,}* {(or|and)} NP	common-law countries, including Canada and England
NP_H {,} especially {NP}* {(or|and)} NP	European countries, especially France, England, and Spain

https://www.aclweb.org/anthology/P18-2057/

Lexico-syntactic pattern với các thực thể có tên

- VD 1: PER, POSITION of ORG

  George Marshall, Secretary of State of the United States

- VD 2: PER (named|appointed|chose|etc.) PER Prep? POSITION

  Truman appointed Marshall Secretary of State

- VD 3: PER [be]? (named|appointed|etc.) Prep? ORG POSITION

  George Marshall was named US Secretary of State

Tiếp cận theo học có giám sát

- Phương pháp tiếp cận:

  (1) Tìm các cặp thực thể tên (pair of named entities)

  (2) Áp dụng bộ phân lớp các mối quan hệ giữa các thực thể trên từng cặp thực thể.

    - Các phương pháp phân lớp thường gặp: Logistic Regression, RNN, Random forest, transformers, ...

Mô tả thuật toán

```
function FINDRELATIONS(words) returns relations

    relations ← nil
    entities ← FINDENTITIES(words)
    forall entity pairs ⟨e1, e2⟩ in entities do
        if RELATED?(e1, e2)
            relations ← relations+CLASSIFYRELATION(e1, e2)
```

Những loại đặc trưng sử dụng

- Đặc trưng từ (Word features)
- Đặc trưng thực thể (Named entities features)
- Đặc trưng ngữ pháp (Syntactic features)

Đặc trưng từ

Câu ví dụ:

American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said.

- **Feature 1**: Từ chính (headword) của **M1** và **M2** và kết hợp chúng

  Airlines (M1) *Wagner(M2)* *Airlines-Wagner*

- **Feature 2**: Bag-of-words và bigrams có trong M1 và M2

  American, Airlines, Tim, Wagner, American Airlines, Tim Wagner

- **Feature 3**: Từ hoặc bigram ở vị trí nào đó

  M2: -1 spokesman

  M2: +1 said

- **Feature 4**: Bag-of-words và bigrams xuất hiện giữa M1 và M2

  a, AMR, of, immediately, matched, move, spokesman, the, unit

Đặc trưng thực thể

Câu ví dụ:

American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said.

- **Feature 1**: Loại thực thể và kết hợp chúng.

  M1: ORG, M2: PER,

  M1M2: ORG-PER

- **Feature 2**: Mức độ thực thể của M1 và M2 (là một trong 3 loại: NAME, NOMINAL, PRONOUN).

  M1: NAME [it or he would be PRONOUN]

  M2: NAME [the company would be NOMINAL]

- **Feature 3**: Số lượng thực thể xuất hiện giữa M1 và M2

Đặc trưng ngữ pháp

- Đường đi giữa M1 và M2 trong cây ngữ pháp thành tố (Constituent paths between M1 and M2):

  NP↑NP↑S↑S↓NP

- Đường đi từ giữa M1 và M2 theo ngữ pháp phụ thuộc:

  Airlines $\leftarrow_{sub\ j}$ matched $\leftarrow_{com\ p}$ said $\rightarrow_{sub\ j}$ Wagner

Bộ phân lớp quan hệ dùng mạng nơ ron (Neural supervised relation classifiers)

- **Input**: Câu (sentence) và 2 spans:
  - Subject: thực thể tên người, tên tổ chức.
  - Object: thực thể còn lại.
- **Output**: mối quan hệ giữa 2 thực thể đầu vào.
  - 42 mối quan hệ (theo bộ dữ liệu TACRED).
  - Không có quan hệ nào.

Zhang et al., Position-aware Attention and Supervised Data Improve Slot Filling, EMNLP, 2017

Ví dụ

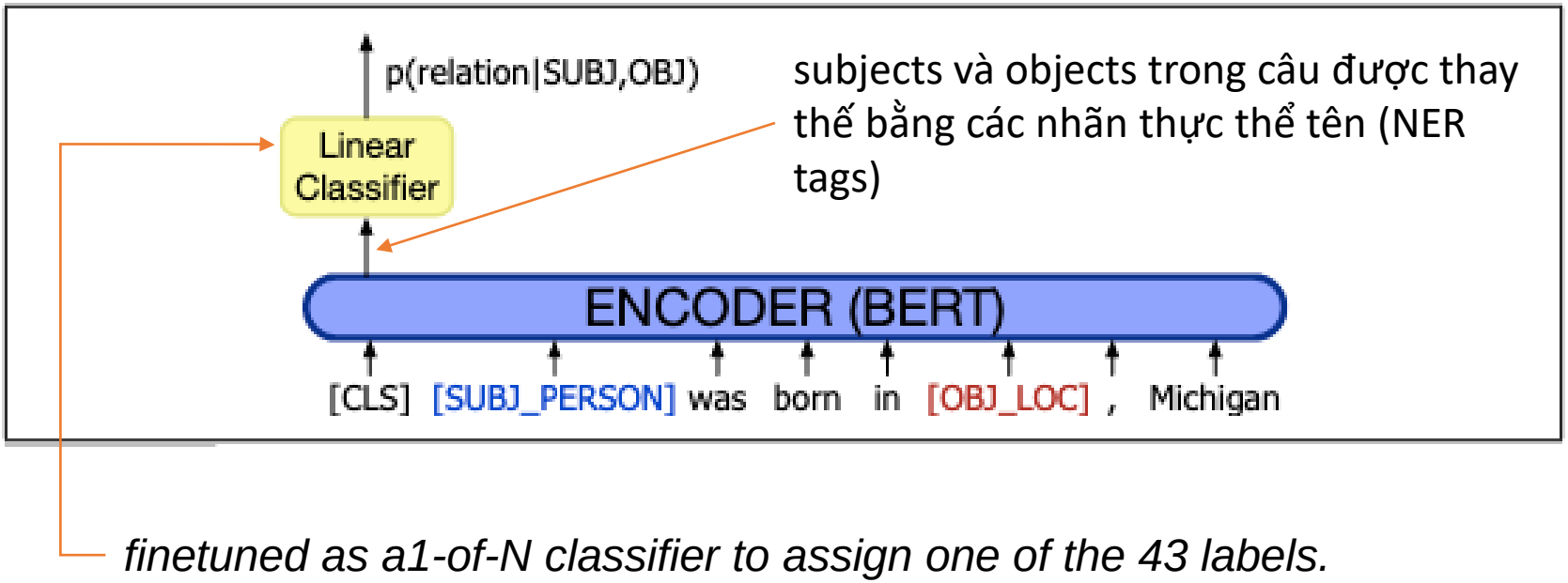
- Input: Penner is survived by his brother, John, a copy editor at the Times, and his former wife, Times sportswriter Lisa Dillman.
- Output:

Subject	Relation	Object
Mike Penner	per:spouse	Lisa Dillman
Mike Penner	per:siblings	John Penner
Lisa Dillman	per:title	Sportswriter
Lisa Dillman	per:employee_of	Los Angeles Times
John Penner	per:title	Copy Editor
John Penner	per:employee_of	Los Angeles Times

https://www.aclweb.org/anthology/D17-1004.pdf

Một số mô hình tiếp cận theo transformers

- RoBERTa (Liu et al., 2019).
- SPANbert (Joshiet al., 2020)



Tiếp cận theo học bán giám sát

- Sử dụng khi: Dữ liệu có nhãn không nhiều, trong khi các mô hình học có giám sát cần nhiều dữ liệu để huấn luyện.
- Các giải quyết: **Bootstrapping**
  - Sử dụng **seed pattern**, và **seed tuple** (những **lexico-syntactic pattern** trước đó) nhằm trích xuất ra các thực thể mới.

  => tạo ra dữ liệu mới từ dữ liệu hiện có.

Thuật toán Bootstrap

function BOOTSTRAP(*Relation R*) **returns** *new relation tuples*

tuples ← Gather a set of seed tuples that have relation *R*
iterate
sentences ← find sentences that contain entities in *tuples*
patterns ← generalize the context between and around entities in *sentences*
newpairs ← use *patterns* to grep for more tuples
newpairs ← *newpairs* with high confidence
tuples ← *tuples* + *newpairs*
return tuples

Ví dụ

Giả sử, cần tìm ra mối quan hệ giữa **airline** và hub.

Dữ liệu chưa gán nhãn:

(17.6) Budget airline Ryanair, which uses Charleroi as a hub, scrapped all weekend flights out of the airport.

(17.7) All flights in and out of Ryanair's hub at Charleroi airport were grounded on Friday...

(17.8) A spokesman at Charleroi, a main hub for Ryanair, estimated that 8000 passengers had already been affected.

- Sử dụng các **seed pattern** sau:
  / [ORG], which uses [LOC] as a hub /
  / [ORG]'s hub at [LOC] /
  / [LOC], a main hub for [ORG] /

- Kết quả thu được: < ***Ryanair ,Charleroi***>

  (17.6) Budget airline [Ryanair, which uses Charleroi as a hub], scrapped all weekend flights out of the airport.

  (17.7) All flights in and out of [Ryanair's hub at Charleroi] airport were grounded on Friday...

  (17.8) A spokesman at [Charleroi, a main hub for Ryanair], estimated that 8000 passengers had already been affected.

Semantic drift

- Là lỗi xảy ra do các pattern so khớp không chính xác, dẫn đến việc trích xuất ra các kết quả sai.
- VD: Sydney has a ferry hub at Circular Quay.

Pattern sử dụng:

/ [ORG], which uses [LOC] as a hub /

/ [ORG]'s hub at [LOC] /

/ [LOC], a main hub for [ORG] /

➔Trích xuất được bộ <***Sydney, Circular Quay*** >

➔***Sydney không phải là airlines***

Khắc phục semantic drift

- Phương pháp sử dụng: Khoảng tin cậy (Confidence values).
- Cho **D** là tập tài liệu (document), **T** là một bộ các thực thể, và pattern **p**.
- Công thức tính khoảng tin cậy khi trích xuất ra bộ T từ pattern p như sau:

$$Conf_{RlogF}(p) = \frac{|hits(p)|}{|finds(p)|} log(|finds(p)|)$$

- Trong đó:
  - **hits(p)**: tập các bộ tìm được trong T khi tìm trong D bằng pattern p.
  - **find(p)**: tổng các bộ tìm được trong D bằng pattern p.

Tiếp cận Distant Supervision

- Phương pháp này kết hợp điểm mạnh giữa phương pháp **học có giám sát** và **boostraping**.
- Ý tưởng:
  - Tạo ra các seed example từ các CSDL lớn (DPedia, Wikipedia).
  - Sinh các pattern từ các example, bao gồm cả các noisy pattern.
  - Kết hợp các pattern lại với nhau thông qua một bộ phân lớp (classifier).

Ví dụ

- B1. Trích xuất quan hệ place-of-birth giữa thực thể người và thực thể thành phố từ các database có sẵn: **tuples** place-of-birth (<Edwin Hubble, Marshfield>)
- B2. Áp dụng một bộ gán nhãn thực thể trên tập dữ liệu văn bản lớn như Wikipedia nhằm gán nhãn cho các thực thể tên.
- B3. Trích xuất ra các câu (sentence) trong tập dữ liệu có chứa các tuples (B1).

VD:

..**Hubble** was born in **Marshfield**...

...**Einstein**, born (1879), **Ulm**...

...**Hubble's** birthplace in **Marshfield**...

Ví dụ (tt)

- B4. Rút ra các training examples từ dữ liệu trên để huấn luyện bộ phân lớp có giám sát.
  - ❖Do dữ liệu huấn luyện lớn nên có thể dùng nhiều đặc trưng giàu thông tin (rich features).

VD:

<born-in, Edwin Hubble, Marshfield>

<born-in, Albert Einstein, Ulm>

<born-year, Albert Einstein, 1879>

Distant Supervision

function DISTANT SUPERVISION(*Database D, Text T*) **returns** *relation classifier C*

foreach relation *R*
  foreach tuple (*e1,e2*) of entities with relation *R* in *D*
    sentences ← Sentences in *T* that contain *e1* and *e2*
    f ← Frequent features in *sentences*
    observations ← observations + new training tuple (*e1, e2, f, R*)
C ← Train supervised classifier on *observations*
return *C*

A neural classifier would skip the feature set f

Rich features learning for supervised classifier

Sentence: ***American Airlines***, *a unit of AMR, immediately matched the move, spokesman* ***Tim Wagner*** *said*

Rich features:

M1 = ORG & M2 = PER & nextword=“said”& path=NP↑NP↑S↑S↓NP

Vai trò của rich features: liên kết các features độc lập lại với nhau thành một features có khả năng thể hiện được quan hệ giữa các thực thể trong câu.

Tiếp cận theo học không giám sát

- Mục tiêu: trích xuất được thông tin cần thiết khi không có dữ liệu huấn luyện được gán nhãn sẵn (unlabeled data).
  - Tác vụ này thường được gọi là open information extraction (Open IE).
- Trong tác vụ Open IE, các quan hệ (relation) thường là một chuỗi các từ (string of words), và thường bắt đầu là động từ (verb).
  - **Hệ thống tiêu biểu: ReVerb** system (**Fader et al., 2011**)

ReVerb system (Fader et al., 2011)

- Mục tiêu: trích xuất ra thông tin về mối quan hệ giữa các thực thể trong một câu s.
- Các bước thực hiện:
  - **Bước 1**: Run a **part-of-speech tagger** and entity chunker over s.
  - **Bước 2**: For each verb in s, find the **longest sequence of words** w that **start with a verb** and satisfy **syntactic and lexical constraints**, merging adjacent matches.
  - **Bước 3**: For each phrase w, find the **nearest noun phrase** x **to the left** which is **not a relative pronoun**, **wh-word** or **existential "there"**. Find the **nearest noun phrase** y **to the right**.
  - **Bước 4**: Assign **confidence** c to the relation **r = (x,w,y)** using a confidence classifier and return it

https://www.aclweb.org/anthology/D11-1142/

Ràng buộc từ vựng và cú pháp (Syntactic and lexical constraints)

V|VP|VW*P

V = verb particle? adv?

W = (noun|adj|adv|pron|det)

P = (prep|particle|inf. marker)

So sánh các phương pháp tiếp cận

Pattern	Supervised learning	Semi-supervised	Unsupervised

• **Điểm mạnh**:
  • Độ chính xác cao (high-precision).
• **Điểm yếu**:
  • Độ phủ thấp (low-recall)
  • Chỉ phù hợp cho một miền dữ liệu (domain) cụ thể.
	• **Điểm mạnh**:
	  • High accuracies.
	• **Điểm yếu**:
	  • Yêu cầu annotated data => gán nhãn khá tốn thời gian và công sức.
		• **Điểm mạnh**:
		  • Không yêu cầu nhiều dữ liệu gán nhãn.
		• **Điểm yếu**:
		  • Low precision.
		  • Cần nhiều data (trích từ các CSDL lớn như Wikipedia).
			• Điểm mạnh:
			  • Có thể trích xuất thông tin đầy đủ mà không cần phải huấn luyện trước.
			• Điểm yếu:
			  • Bỏ qua một số quan hệ ngữ nghĩa quan trọng.

Đánh giá mô hình trích xuất quan hệ

- Metric đánh giá cho phương pháp supervised learning: precision, recall và F1.
- Metric đánh giá cho các hệ thống semi-supervised và unsupervised: sử dụng **estimated precision $\hat{p}$**

$$\hat{P} = \frac{\text{\# of correctly extracted relation tuples in the sample}}{\text{total \# of extracted relation tuples in the sample.}}$$

Rút trích sự kiện và thời gian (Event and times extraction)

Trích xuất thông tin về sự kiện (events)

- Trích xuất thông tin về sự kiện (event extraction) là tác vụ đi tìm xem các **thực thể (entities)** đã tham gia vào **sự kiện (events)** nào.
  - Rút trích sự kiện bao gồm thời gian sự kiện đó diễn ra ➔ rút trích thông tin về thời gian (time extraction).
  - Rút trích ra hành động cụ thể xảy ra của các sự kiện ➔ rút trích thông tin sự kiện (event extraction).

Rút trích thông tin thời gian

- Thời gian diễn ra sự kiện được biểu diễn bởi cụm từ chỉ thời gian: **temporal expression**.
- Các thông tin về thời gian diễn ra sự kiện sau khi được rút trích ra phải được chuẩn hoá (**normalization**) để có thể suy diễn.
- Ứng dụng:
  - Các hệ thống hỏi đáp (QA).
  - Các hệ thống trợ lý ảo (assistant).

Cụm từ chỉ thời gian (Temporal expression)

- Cụm từ chỉ thời gian có 3 loại:
  - ***Thời gian tuyệt đối*** *(absolute time)*: thời gian cụ thể (năm/tháng/ngày/giờ/phút/giây/....)
  - ***Thời gian tương đối*** *(relative time)*: thời gian cụ thể ứng với một sự kiện nào đó (VD: thứ 4, học kỳ sắp tới, ...)
  - ***Khoảng thời gian*** *(duration)*: khoảng thời gian diễn ra sự kiện.

Ví dụ về Temporal expression

Absolute	Relative	Durations
April 24, 1916	yesterday	four hours
The summer of '77	next semester	three weeks
10:15 AM	two weeks from yesterday	six days
The 3rd quarter of 2006	last quarter	the last three quarters

Lexical trigger

- **Lexical trigger** là cấu trúc ngữ pháp (grammatical construction) dùng để biểu diễn cho các temporal expressions.
- **Lexical trigger** có thể là: danh từ (noun), danh từ riêng (proper noun), tính từ (adjective) và trạng từ (adverb).

Category	Examples
Noun	*morning, noon, night, winter, dusk, dawn*
Proper Noun	*January, Monday, Ides, Easter, Rosh Hashana, Ramadan, Tet*
Adjective	*recent, past, annual, former*
Adverb	*hourly, daily, monthly, yearly*

Chuẩn hóa thời gian (Temporal Normalization)

- Temporal Normalization là quá trình chuyển từ **temporal expression** thành **ngày giờ cụ thể**:
  - Absolute time ➔ ngày tháng năm theo định dạng ISO.
  - Duration ➔ khoảng thời gian hoặc là thời điểm bắt đầu hoặc kết thúc.

Unit	Pattern	Sample Value
Fully specified dates	YYYY-MM-DD	1991-09-28
Weeks	YYYY-Wnn	2007-W27
Weekends	PnWE	P1WE
24-hour clock times	HH:MM:SS	11:13:45
Dates and times	YYYY-MM-DDTHH:MM:SS	1991-09-28T11:00:00
Financial quarters	Qn	1999-Q3

Định dạng ngày tháng năm theo chuẩn ISO 8601

Thời điểm neo (Temporal anchor)

- **Thời điểm neo:** là một thông tin cơ sở để xác định quan hệ thời gian trong văn bản.

VD: nếu “today” là **Thời điểm neo** thì “yesterday” có thể xem là sự kiện xảy ra trước đó, còn “tomorrow” là sự kiện xảy ra sau đó.

Bài toán Nhận diện cụm từ thời gian (Temporal expression recognition)

- Input: Câu s.
- Output: **Những cụm từ chỉ thời gian.**

- Phương pháp tiếp cận:
  - **Rules-based**: sử dụng các pattern.
  - **Sequence labeling**: Sử dụng IOB tags.

Cách tiếp cận Nhận diện cụm từ thời gian

Rule-based

```
# yesterday/today/tomorrow
$string =" s/((($OT+the$CT+\s+)?$OT+day$CT+\s+$OT+(before|after)$CT+\s+)?$OT+$TERelDayExpr$CT+
(\s+$OT+(morning|afternoon|evening|night)$CT+)?)/<TIMEX$tever TYPE=\"DATE\">$1
<\/TIMEX$tever>/gio;

$string =" s/($OT+\w+$CT+\s+)<TIMEX$tever TYPE=\"DATE\"[^>]*>($OT+(Today|Tonight)$CT+)
<\/TIMEX$tever>/$1$4/gso;

# this (morning/afternoon/evening)
$string =" s/(($OT+(early|late)$CT+\s+)?$OT+this$CT+\s*$OT+(morning|afternoon|evening)$CT+)/
  <TIMEX$tever TYPE=\"DATE\">$1<\/TIMEX$tever>/gosi;
$string =" s/(($OT+(early|late)$CT+\s+)?$OT+last$CT+\s*$OT+night$CT+)/<TIMEX$tever
   TYPE=\"DATE\">$1<\/TIMEX$tever>/gsio;
```

GUTime temporal tagging system in Tarsqi (Verhagen et al., 2005)

Sequence labeling

A fare increase initiated last week by UAL Corp's...

O O O O B I O O O

Feature	Explanation
Token	The target token to be labeled
Tokens in window	Bag of tokens in the window around a target
Shape	Character shape features
POS	Parts of speech of target and window words
Chunk tags	Base phrase chunk tag for target and words in a window
Lexical triggers	Presence in a list of temporal terms

Các dạng feature dùng cho các bộ nhãn IOB

Rút trích thông tin về sự kiện

- **Mục tiêu**: Nhận diện các sự kiện (event) được đề cập trong văn bản.

VD:

[EVENT Citing] high fuel prices, United Airlines [EVENT said] Friday it has [EVENT increased] fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit of AMR Corp., immediately [EVENT matched] [EVENT the move], spokesman Tim Wagner [EVENT said]. United, a unit of UAL Corp., [EVENT said] [EVENT the increase] took effect Thursday and [EVENT applies] to most routes where it [EVENT competes] against discount carriers, such as Chicago to Dallas and Denver to San Francisco.

- Phân loại sự kiện:
  - Hành động (action)
  - Trạng thái (state)
  - Tường thuật (say, report, tell, explain)

Đặc trưng sử dụng

Feature	Explanation
Character affixes	Character-level prefixes and suffixes of target word
Nominalization suffix	Character-level suffixes for nominalizations (e.g., *-tion*)
Part of speech	Part of speech of the target word
Light verb	Binary feature indicating that the target is governed by a light verb
Subject syntactic category	Syntactic category of the subject of the sentence
Morphological stem	Stemmed version of the target word
Verb root	Root form of the verb basis for a nominalization
WordNet hypernyms	Hypernym set for the target

- Light verb: *do*, *give*, *have*, *make*, *get*, and *take*

Bài toán xác định thứ tự các sự kiện theo thời gian

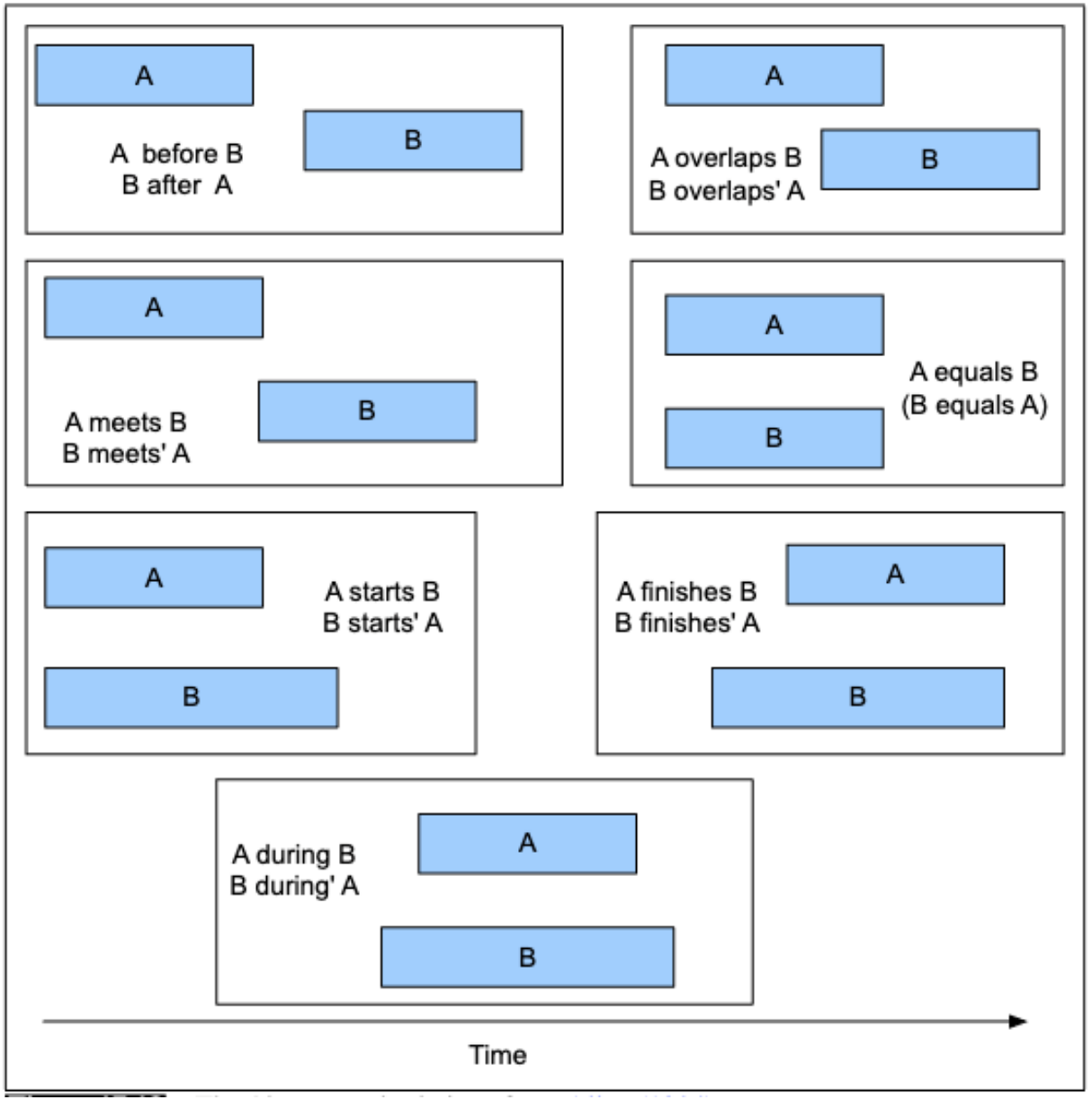
- Xác định thứ tự các sự kiện theo thời gian (Temporal Ordering of Events): nhằm mục tiêu xác định xem sự kiện nào xảy ra trước, sự kiện nào xảy ra sau.

VD:

> Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit of AMR Corp., immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL Corp., said the increase took effect Thursday and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Denver to San Francisco.

"fare increase by **American Airlines**" xảy ra sau "fare increase by **United Airlines**".

Allen relations

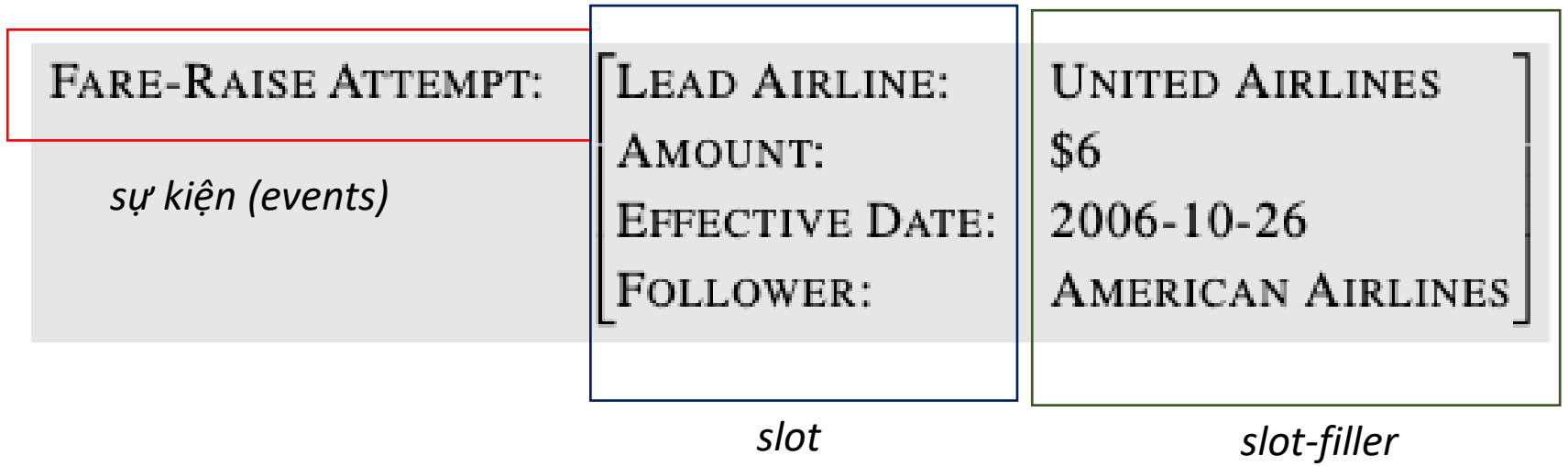


Allen, J, **Towards a general theory of action and time** (1984)

-Biểu diễn quan hệ về mặt thời gian của các sự kiện.
-Ứng dụng trong suy diễn ngữ nghĩa về thời gian (action reasoning, temporal reasoning)
-Ứng dụng trong biểu diễn ngữ nghĩa trong tiếng Anh.

Điền mẫu (Template Filling)

- **Template filling** là tác vụ tìm các đoạn thông tin được mô tả trong văn bản để điền vào mẫu (template) được định sẵn.
  - VD: Điền mẫu "Nâng giá vé máy bay"



Phương pháp tiếp cận

- Tiếp cận máy học (gồm 2 tác vụ chính).
  - *Tác vụ 1: template recognition*
  - *Tác vụ 2: role-filler extraction*

FARE-RAISE ATTEMPT:	LEAD AIRLINE:	UNITED AIRLINES
	AMOUNT:	$6
	EFFECTIVE DATE:	2006-10-26
	FOLLOWER:	AMERICAN AIRLINES

- Dựa trên luật: Regular expression + rules (VD: Hệ thống Finite-State Template-Filling - FST)
  - Tokenization, chunking, parsing
  - Xác định thực thể và sự kiện cho từng slot

```
NG(Company/ies) VG(Set-up) NG(Joint-Venture) with NG(Company/ies)
VG(Produce) NG(Product)
```

  - Phân giải đồng tham chiếu (coreference resolution)
  - Merge thông tin dựa trên kết quả phân giải đồng tham chiếu

Tổng kết

- Rút trích thông tin nhằm tìm ra các thông tin có giá trị từ dữ liệu văn bản (không có cấu trúc).
- Hai bài toán chính liên quan đến trích xuất thông tin:
  - Rút trích quan hệ (relation extraction).
  - Rút trích sự kiện và thời gian (events and time extraction).
- Các phương pháp tiếp cận: rule-based, sequence labeling, supervised learning, semi-supervised learning, unsupervised learning.
- Các độ đo đánh giá: precision, recall, accuracy, MUC.

Bộ dữ liệu

- Relation extraction:

  TACRED dataset (Zhang et al., 2017)

- Time and events extraction:

  TimeML corpus (Pustejovsky et al., 2003)

- Event:

  MUC datasets (https://catalog.ldc.upenn.edu/LDC2001T02)

  Message Understanding Conference

Q&A