Rút trích thông tin trong văn bản (Information Extraction)

Nguyễn Trường Sơn
ntson@fit.hcmus.edu.vn

KHOA CÔNG NGHỆ THÔNG TIN
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

Nội dung

- Giới thiệu rút trích thông tin và ứng dụng
- Các cách tiếp cận
- Demo:
  - Nhận diện thực thể có tên (NER)

Rút trích thông tin

- **Information extraction** (**IE**):
  - Tự động rút trích các thông tin có cấu trúc từ những tài liệu không có cấu trúc / bán cấu trúc
  - Rút trích thông tin = rút trích từ các tài liệu đa phương tiện (multimedia): **văn bản**, hình ảnh, âm thanh, video

Yesterday, New York based Foo Inc. announced their acquisition of Bar Corp."

MergeBetween(Foo Inc, Bar Corp, Yesterday)

Các dạng bài toán đặc trưng

- Nhận diện thực thể đã định danh:
  - Named Entity Recognization (NER)
- Nhận diện quan hệ giữa các thực thể
  - Relation Extraction
  - Event Extraction
- Nhận diện đồng tham chiếu (Co-reference)
- …

Named Entity Recognition and Coreference Resolution

Named Entity Recognition (NER):

Example:

The shiny red rocket was fired on Tuesday. It is the brainchild of Dr. Big Head.
Dr. Head is a staff scientist at We Build Rockets Inc.

→ <person> Dr. Big Head </person>

Output <person> Dr. Head </person>
<organization> We Build Rockets Inc </organization>
<time> Tuesday </time>

Coreference resolution (anaphor resolution):

• Connect pronous etc. to subject/object of previous sentence

Examples:

• The shiny red rocket was fired on Tuesday. It is the brainchild of Dr. Big Head.

Output → … on Tuesday. It <reference> The shiny red rocket </reference> is the …

• Alas, poor Yorick, I knew him Horatio.

Mục tiêu và tổng quan

- Mục tiêu:
  - gán nhãn(annotate) tài liệu văn bản hoặc trang Web (named entity recognition, html2xml, etc.)
  - trích sự kiện (extract facts) từ tài liệu văn bản hay trang Web (relation learning)
  - tìm sự kiện (find facts) trên Web (hoặc trên Wikipedia) để to xây dựng các kho dữ liệu về quan hệ (thesaurus/ontology relations)
  - làm giàu thông tin (information enrichment) (vd. Cho phân tích kinh doanh)
- Công nghệ:
  - NLP (PoS tagging, chunk parsing, etc.)
  - Lexicon lookups (name dictionaries, geo gazetteers, etc.)
  - Pattern matching & rule learning (regular expressions, FSAs)
  - Statistical learning (HMMs, CRFs, etc.)
  - Text mining in general, deep learning *

Tạo "ngữ nghĩa(Semantic)" cho dữ liệu

Hầu hết dữ liệu ở dạng HTML (hoặc PDF hoặc RSS hoặc ..
Hoặc từ các nguồn dữ liệu không có cấu trúc

Có thể truy xuất thông qua Wrappers hoặc Web Service)

→ Luật, FSAs (biểu thức chính qui), …→ HMMs, MRFs, ...

Nhận dạng dữ liệu „free-form" ?

Tạo "ngữ nghĩa(Semantic)" cho dữ liệu

Hầu hết dữ liệu ở dạng HTML (hoặc PDF hoặc RSS hoặc ..
Hoặc từ các nguồn dữ liệu không có cấu trúc

Tạo "ngữ nghĩa(Semantic)" cho dữ liệu

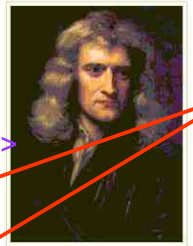
Hầu hết dữ liệu ở dạng HTML (hoặc PDF hoặc RSS hoặc ..
Hoặc từ các nguồn dữ liệu không có cấu trúc

Isaac Newton

From Wikipedia, the free encyclopedia.

<TimePeriod>
<Person>
<Scientist>
<Scientist>
<Publication>
<Painter>
<Person>

Sir Isaac Newton (25 December 1642 – 20 March 1727 by the Julian calendar in use in England at the time; or 4 January 1643 – 31 March 1727 by the Gregorian calendar) was an English physicist, mathematician, astronomer, philosopher, and alchemist; who wrote the *Philosophiae Naturalis Principia Mathematica* (published 5 July 1687)[1], where he described universal gravitation and, via his laws of motion, laid the groundwork for classical mechanics. Newton also shares credit with Gottfried Wilhelm Leibniz for the development of differential calculus. However, their work was not a collaboration; they both discovered calculus separately but nearly contemporaneously.



Sir Isaac Newton in Kneller's portrait of 1689.

NLP-based IE từ Web

ANNIE Output for http://en.wikipedia.org/wiki/Che_Guevarra

Annotation Key:
Person Location Organization Date Address Money Percent

Che Guevara

From Wikipedia, the free encyclopedia.

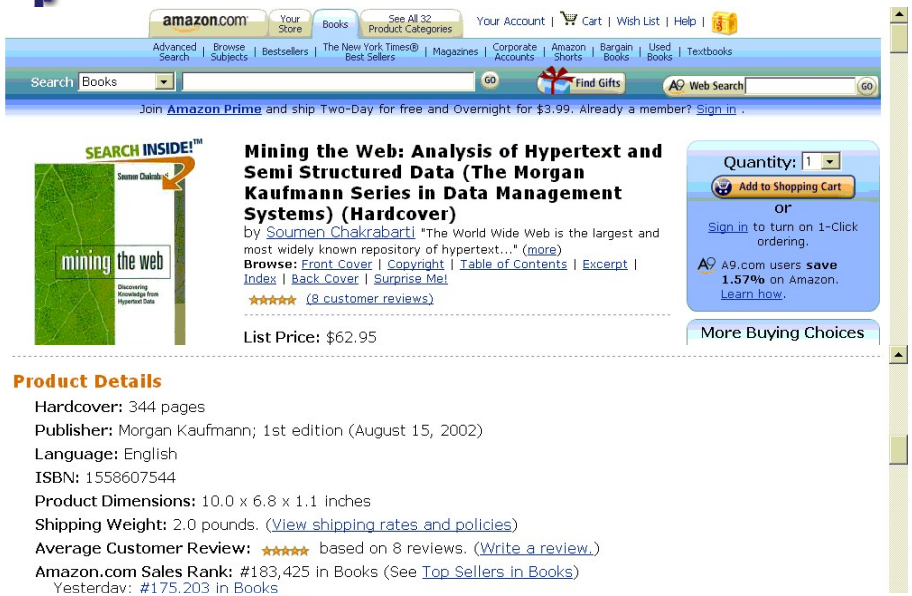
(Redirected from Che Guevarra)
Jump to: navigation , search

Che Guevara

Ernesto Rafael Guevara de la Serna (June 14 , 1928 [1] ? October 9 , 1967), commonly known as Che Guevara or el Che, was an Argentine -born Marxist revolutionary and Cuban guerrilla leader. Guevara was a member of Fidel Castro 's " 26th of July Movement " that seized power in Cuba in 1959 . After serving in various important posts in the new government, Guevara left Cuba in 1965 with the hope of fomenting revolutions in other countries, first in the Congo-Kinshasa (currently the Democratic Republic of the Congo) and later in Bolivia , where he was captured in a CIA -organized military operation. It is believed by some that the CIA wished to keep Guevara alive for interrogation but, after his capture in the Yuro ravine, he died at the hands of the Bolivian Army in La Higuera near Vallegrande on October 9 , 1967 . Testimony by various individuals who were participants in, or

Leading open-source tool: GATE/ANNIE
http://www.gate.ac.uk/annie/

Rút trích các mẫu tin có cấu trúc từ Web (1)



Rút trích các mẫu tin có cấu trúc từ Web (1l)

```
<div class="buying"><b class="sans">Mining the Web: Analysis of Hypertext and
Semi Structured Data (The Morgan Kaufmann Series in Data Management Systems)
(Hardcover)</b><br />by

<a href="/exec/obidos/search-handle-url/index=books&field-author-exact=Soumen%20Ch

<div class="buying" id="priceBlock">
<style type="text/css">
  td.productLabel { font-weight: bold; text-align: right; white-space: nowrap; vertical-align: t
  table.product   { border: 0px; padding: 0px; border-collapse: collapse; }
</style>
<table class="product">
  <tr>
    <td class="productLabel">List Price:</td>
    <td>$62.95</td>
  </tr>
  <tr>
    <td class="productLabel">Price:</td>
    <td><b class="price">$62.95</b>
& this item ships for <b>FREE with Super Saver Shipping</b>.
```

Rút trích các mẫu tin có cấu trúc từ Web (1II)

```
<a name="productDetails" id="productDetails"></a>
<hr noshade="noshade" size="1" class="bucketDivider
<table cellpadding="0" cellspacing="0" border="0">
  <tr>
    <td class="bucket">
<b class="h1">Product Details</b><br />
  <div class="content">
<ul>
<li><b>Hardcover:</b> 344 pages</li>
<li><b>Publisher:</b> Morgan Kaufmann; 1st edition (
<li><b>Language:</b> English</li>
<li><b>ISBN:</b> 1558607544</li>
<li><b>Product Dimensions:</b> 10.0 x 6.8 x 1.1 inche
<li><b>Shipping Weight:</b> 2.0 pounds. (<a href="ht
5490548?%5Fencoding=UTF8&seller=ATVPDKIKX0D
<li><b>Average Customer Review:</b> <img src="http
height="12" border="0" /> based on 8 reviews.
 (<a href="http://www.amazon.com/gp/customer-reviews/write-a-review.html/102-8395894
<li>
<b>Amazon.com Sales Rank:</b>  #183,425 in Books (See <a href="/exec/obidos/tg/new
Books</a>)
```

extract record:

Title: Mining the Web: Analysis of Hypertext and Semi Struct
Author: Soumen Chakrabarti,
Hardcover: 344 pages,
Publisher: Morgan Kaufmann,
Language: English,
ISBN: 1558607544.
...
AverageCustomerReview: 4
NumberOfReviews: 8,
SalesRank: 183425
...

Sample dataset: VLSP NER

```
<ENAMEX TYPE="LOCATION">Mỹ</ENAMEX> kêu gọi <ENAMEX TYPE="LOCATION">Nga</ENAMEX>, <ENAMEX
TYPE="LOCATION">Trung</ENAMEX> dừng phổ biến vũ khí hạt nhân

Ngoại trưởng <ENAMEX TYPE="LOCATION">Mỹ</ENAMEX> <ENAMEX TYPE="PERSON">Rex Tillerson</ENAMEX> kêu
gọi cộng đồng quốc tế ngăn chặn các nước sở hữu vũ khí hạt nhân.

Trong lời kêu gọi dừng các hoạt động tạo điều kiện để vũ khí hạt nhân phổ biến, Ngoại trưởng
<ENAMEX TYPE="LOCATION">Mỹ</ENAMEX> <ENAMEX TYPE="PERSON">Rex Tillerson</ENAMEX> nêu đích danh
<ENAMEX TYPE="LOCATION">Nga</ENAMEX>, <ENAMEX TYPE="LOCATION">Trung Quốc</ENAMEX> – ám chỉ các
nước này nên gia tăng áp lực nhằm buộc <ENAMEX TYPE="LOCATION">Triều Tiên</ENAMEX> phải từ bỏ tham
vọng vũ khí hạt nhân.
Ông <ENAMEX TYPE="PERSON">Tillerson</ENAMEX> lấy <ENAMEX TYPE="LOCATION">Triều Tiên</ENAMEX> làm
bài học điển hình cho việc thất bại trong hoạt động ngăn chặn các nước sở hữu vũ khí tiêu diệt
hàng loạt.
Ngoại trưởng <ENAMEX TYPE="LOCATION">Mỹ</ENAMEX> đưa ra phát ngôn trên trong cuộc họp cấp Bộ
trưởng tại <ENAMEX TYPE="ORGANIZATION">Hội đồng Bảo an Liên Hợp Quốc</ENAMEX> do <ENAMEX
TYPE="LOCATION">Mỹ</ENAMEX> kêu gọi tổ chức nhằm giải quyết mối đe doạ từ các hoạt động phát triển
các vũ khí sinh học, hoá học, hạt nhân.
Ông <ENAMEX TYPE="PERSON">Tillerson</ENAMEX> nhắc lại, <ENAMEX TYPE="LOCATION">Washington</ENAMEX>
và <ENAMEX TYPE="LOCATION">Moscow</ENAMEX> từng hợp tác rất tốt trong thời kỳ Xô-viết về các biện
pháp giải trừ vũ khí, dù là đối thủ Chiến tranh Lạnh với nhau. Vậy nên, bây giờ "hai nước nên hợp
tác như vậy một lần nữa".
Hiện nay, thế giới đang nóng vấn đề vũ khí hạt nhân tên lửa khi <ENAMEX TYPE="LOCATION">Triều
Tiên</ENAMEX> không ngừng thực hiện các vụ thử tên lửa, hạt nhân với sức mạnh không ngừng phát
triển bất chấp các lệnh trừng phạt từ cộng đồng quốc tế
```

NER for COVID-19

Example 1:

Bệnh nhân " 669 " là bác sĩ làm việc tại Bệnh viện Đa khoa Đồng Nai
Patient " 669 " is a doctor working at Dong Nai General Hospital

Example 2:

Bệnh viện Bệnh Nhiệt đới TP HCM xét nghiệm dương tính lần một đêm 12/3 .
Ho Chi Minh City Hospital for Tropical Diseases returns a positive test result in the evening of 12/3 .

Example 3:

Hai nữ điều dưỡng Bệnh viện Bạch Mai lây từ bên ngoài và lây nhiễm cho nhau.
Two nurses of Bach Mai Hospital got infected from external source and then infected each other.

Example 4:

Bệnh nhân tử vong tại Bệnh viện Phổi Đà Nẵng với chẩn đoán viêm phổi nặng , suy đa tạng không hồi phục , trên bệnh nhân suy thận mạn giai đoạn cuối .
The patient died at Da Nang Lung Hospital , diagnosed with severe pneumonia with history of unrecoverable multiorgan dysfunction syndrome , terminal chronic kidney failure .

NER for COVID-19

Label	Definition
PATIENT_ID	Unique identifier of a COVID-19 patient in Vietnam. An PATIENT_ID annotation over "X" refers to as the X^{th} patient having COVID-19 in Vietnam.
PERSON_NAME	Name of a patient or person who comes into contact with a patient.
AGE	Age of a patient or person who comes into contact with a patient.
GENDER	Gender of a patient or person who comes into contact with a patient.
OCCUPATION	Job of a patient or person who comes into contact with a patient.
LOCATION	Locations/places that a patient was presented at.
ORGANIZATION	Organizations related to a patient, e.g. company, government organization, and the like, with structures and their own functions.
SYMPTOM&DISEASE	Symptoms that a patient experiences, and diseases that a patient had prior to COVID-19 or complications that usually appear in death reports.
TRANSPORTATION	Means of transportation that a patient used. Here, we only tag the specific identifier of vehicles, e.g. flight numbers and bus/car plates.
DATE	Any date that appears in the sentence.

Table 1: Definitions of entity types in our annotation guidelines. We do not annotate nested entities.

Các ứng dụng IE

- So sánh giá và tư vấn mua sắm portals
  ví dụ: consumer electronics, used cars, real estate, pharmacy, etc.
- Phân tích kinh doanh dựa trên hồ sơ khách hàng, báo cáo tài c
  e.g.: How was company X (the market Y) performing in the last 5 years?
- Thị trường/khách hàng, ứng ụng PR , phân tích truyền thông
  e.g.: How are our products perceived by teenagers (girls)?
  How good (and positive?) is the press coverage of X vs. Y?
  Who are the stakeholders in a public dispute on a planned airport?
- Giới thiệu việc làm (applications/resumes, job offers)
  e.g.: WHo well does the candidate match the desired profile?
- Quản trị tri thức trong các công ty tư vấn
  e.g.: Do we have experience and competence on X, Y, and Z in Brazil?
- Khai thác E-mail đã được lưu trữ
  e.g.: Who knew about the scandal on X before it became public?
- Trích tri thức từ văn bản khoa học
  e.g.: Which anti-HIV drugs have been found ineffective in recent papers?
- phám phá tri thức tổng quan
  Can we learn encyclopedic knowledge from text & Web corpora?

Các quan điểm và cách tiếp cận IE

IE là học các biểu thức chính qui (regular expressions)
(wrapping pages with common structure from Deep-Web source)

IE là học các quan hệ (relations)
(rules for identifying instances of n-ary relations)

IE là học các sự kiện (fact boundaries)

IE là học phân đoạn văm bản (segmentation) (HMMs etc.)

IE là học các mẫu ngữ cảnh (contextual patterns) (graph models

IE là phân tích ngôn ngữ tự nhiên (natural-language analysis) (N

IE là khai phá kho văn bản lớn để khám phá tri trức
(combination of tools incl. Web queries)

Đánh giá chất lượng của IE

Sử dụng các độ đo chuẩn của IR:
- precision
- Recall
- F1 measure

Các bộ dữ liệu chuẩn từ các cuộc thi, hội nghị:

Tiếng Anh và các ngôn ngữ khác:
- MUC (Message Understanding Conference)
- ACE (Automatic Content Extraction), http://www.nist.gov/speech/tests/ace/
- TREC Enterprise Track, http://trec.nist.gov/tracks.html
- Enron e-mail mining, http://www.cs.cmu.edu/~enron

Tiếng Việt:
- VLSP Shared Tasks: 2018, 2019, 2020

MUC Information Extraction Example

Bridgestone Sports Co. said Friday it had set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be supplied to Japan. The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and "metal wood" clubs a month.

JOINT-VENTURE-1
- Relationship: TIE-UP
- Entities: "Bridgestone Sport Co." , "a local concern", "a Japanese trading house"
- Joint Ent: "Bridgestone Sports Taiwan Co."
- Activity: ACTIVITY-1
- Amount: NT$20 000 000

ACTIVITY-1
- Activity: PRODUCTION
- Company: "Bridgestone Sports Taiwan Co."
- Product: "iron and 'metal wood' clubs"
- Start date: DURING: January 1990

Các thách thức trong IE

- Phụ thuộc lĩnh vực
- **Landscape of IE Tasks**
  - Nhiều cấp độ định dạng khác nhau
  - Sự đa dạng
  - Sự phức tạp
  - Cấp độ rút trích khác nhau

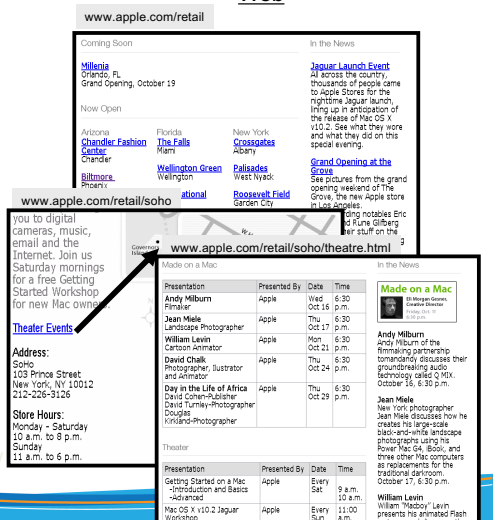
IE is different in different domains!

Example: on web there is less grammar, but more formatting & linking

Newswire

> Apple to Open Its First Retail Store in New York City
>
> MACWORLD EXPO, NEW YORK--July 17, 2002--Apple's first retail store in New York City will open in Manhattan's SoHo district on Thursday, July 18 at 8:00 a.m. EDT. The SoHo store will be Apple's largest retail store to date and is a stunning example of Apple's commitment to offering customers the world's best computer shopping experience.
>
> "Fourteen months after opening our first retail store, our 31 stores are attracting over 100,000 visitors each week," said Steve Jobs, Apple's CEO. "We hope our SoHo store will surprise and delight both Mac and PC users who want to see everything the Mac can do to enhance their digital lifestyles."

Web



The directory structure, link structure, formatting & layout of the Web is its own new grammar.

Landscape of IE Tasks (1/4): Degree of Formatting

Text paragraphs without formatting

> Astro Teller is the CEO and co-founder of BodyMedia. Astro holds a Ph.D. in Artificial Intelligence from Carnegie Mellon University, where he was inducted as a national Hertz fellow. His M.S. in symbolic and heuristic computation and B.S. in computer science are from Stanford University. His work in science, literature and business has appeared in international media from the New York Times to CNN to NPR.

Grammatical sentences and some formatting & links

> **Dr. Steven Minton** - Founder/CTO
> Dr. Minton is a fellow of the American Association of Artificial Intelligence and was the founder of the Journal of Artificial Intelligence Research. Prior to founding Fetch, Minton was a faculty member at USC and a project leader at USC's Information Sciences Institute. A graduate of Yale University and Carnegie Mellon University, Minton has been a Principal Investigator at NASA Ames and taught at Stanford, UC Berkeley and USC.
>
> **Frank Huybrechts** - COO
> Mr. Huybrechts has over 20 years of
>
> • Press
> **Contact**
> • General information
> • Directions maps

Non-grammatical snippets, rich formatting & links

Barto, Andrew G.	(413) 545-2109	barto@cs.umass.edu	CS276
Professor. Computational neuroscience, reinforcement learning, adaptive motor control, artificial neural networks, adaptive and learning control, motor development.			
Berger, Emery D.	(413) 577-4211	emery@cs.umass.edu	CS344
Assistant Professor.			
Brock, Oliver	(413) 577-0334	oli@cs.umass.edu	CS246
Assistant Professor.			
Clarke, Lori A.	(413) 545-1328	clarke@cs.umass.edu	CS304
Professor. Software verification, testing, and analysis; software architecture and design.			
Cohen, Paul R.	(413) 545-3638	cohen@cs.umass.edu	CS278
Professor. Planning, simulation, natural language, agent-based systems, intelligent data analysis, intelligent user interfaces.			

Tables

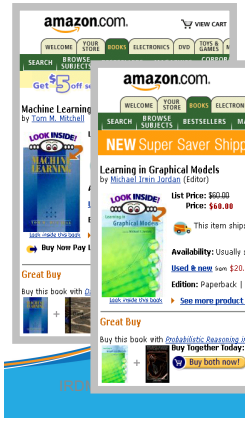
8:30 - 9:30 AM	Invited Talk: Plausibility Measures: A General Approach for Representing Uncertainty *Joseph Y. Halpern, Cornell University*				
9:30 - 10:00 AM	Coffee Break				
10:00 - 11:30 AM	Technical Paper Sessions:				
Cognitive Robotics	**Logic Programming**	**Natural Language Generation**	**Complexity Analysis**	**Neural Networks**	**Games**
739: A Logical Account of Causal and Topological Maps *Emilio Remolina and Benjamin Kuipers*	116: A-System: Problem Solving through Abduction *Marc Denecker, Antonis Kakas, and Bert Van Nuffelen*	758: Title Generation for Machine-Translated Documents *Rong Jin and Alexander G. Hauptmann*	417: Let's go Nuts: Complexity of Nested Circumscription and Abnormality Theories *Marco Cadoli, Thomas Eiter, and Georg Gottlob*	179: Knowledge Extraction and Comparison from Local Function Networks *Kenneth McGarry, Stefan Wermter, and John MacIntyre*	71: Iterative Widening *Tristan Cazenave*
549: Online-Execution of ccGolog Plans *Henrik Grosskreutz*	131: A Comparative Study of Logic Programs with	246: Dealing with Dependencies between Content Planning and	470: A Perspective on Knowledge Compilation	258: Violation-Guided Learning for Constrained	353: Temporal Difference Learning Applied to a

Landscape of IE Tasks (2/4): Intended Breadth of Coverage

Web site specific

Formatting

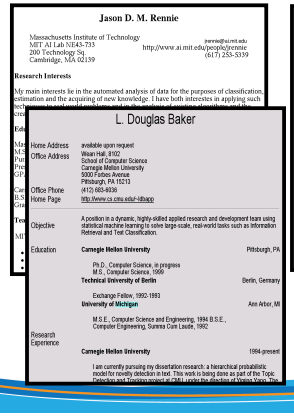
Amazon.com Book Pages



Genre specific

Layout

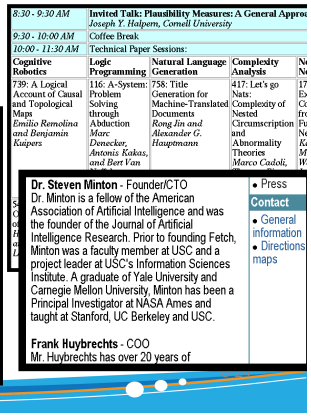
Resumes



Wide, non-specific

Language

University Names



Landscape of IE Tasks (3/4): Complexity

E.g. word patterns:

Closed set

U.S. states

He was born in Alabama…

The big Wyoming sky…

Regular set

U.S. phone numbers

Phone: (413) 545-1323

The CALD main office can be reached at 412-268-1299

Complex pattern

U.S. postal addresses

University of Arkansas
P.O. Box 140
Hope, AR 71802

Headquarters:
1128 Main Street, 4th Floor
Cincinnati, Ohio 45210

Ambiguous patterns, needing context and many sources of evidence

Person names

…was among the six houses sold by Hope Feldman that year.

Pawel Opalinski, Software Engineer at WhizBang Labs.

Landscape of IE Tasks (4/4): Single Field vs. Record

Jack Welch will retire as CEO of General Electric tomorrow. The top role at the Connecticut company will be filled by Jeffrey Immelt.

Single entity

Person: Jack Welch

Person: Jeffrey Immelt

Location: Connecticut

Binary relationship

Relation: Person-Title
Person: Jack Welch
Title: CEO

Relation: Company-Location
Company: General Electric
Location: Connecticut

N-ary record

Relation: Succession
Company: General Electric
Title: CEO
Out: Jack Welsh
In: Jeffrey Immelt

"Named entity" extraction

Relation extraction

Các mô hình sử dụng trong IE

Lexicons

Abraham Lincoln was born in Kentucky.

member?

Alabama
Alaska
…
Wisconsin
Wyoming

Classify Pre-segmented Candidates

Abraham Lincoln was born in Kentucky.

Classifier

which class?

Sliding Window

Abraham Lincoln was born in Kentucky.

Classifier

which class?

Try alternate window sizes:

Boundary Models

Abraham Lincoln was born in Kentu

BEGIN

Classifier

which class?

BEGIN END BEGIN END

Finite State Machines

Abraham Lincoln was born in Kentucky.

Most likely state sequence?

Context Free Grammars

Abraham Lincoln was born in Kentucky.

NNP NNP V V P NP

PP

VP

NP

VP

S

Most likely parse?

Any of these models can be used to capture words, formatting or both.

Học các biểu thức chính qui

Input: Các mẫu biểu thức chính qui được gán nhãn thủ công
learn: (restricted) biểu thức chính qui
hoặc một mạng trạng thái hữu hạn dùng để đọc một câu và xuất ra các mẫu mong muốn

Ví dụ:
This appartment has 3 bedrooms.
 The monthly rent is $ 995.
This appartment has 3 bedrooms.
 The monthly rent is $ 995.
The number of bedrooms is 2.
 The rent is $ 675 per month.

Learned pattern: * *Digit* „
“ * „$“ *Number* *
Input sentence: There are 2 bedrooms.
 The price is $ 500 for one month.
Output tokens: Bedrooms: 1, Price: 500

but: grammar inference for full-fledged regular languages is hard
→ focus on restricted fragments of the class of regular languages

implemented in WHISK (Soderland 1999) and a few other systems

IE như bài toán phân loại các giới hạn

Ý tưởng chính:

Bộ phân loại (ví dụ: SVMs) nhận dạng ra token bắt đầu và token kết thúc của một đối tượng được xét đến

Tổ hợp nhiều bộ phân loại để nâng cao hiệu năng

There will be a talk by Alan Turing at the CS Department at 4 PM .

Prof. Dr. James D. Watson will speak on DNA at MPI on Thursday, Jan 12.

The lecture by Sir Francis Crick will be in the Institute of Informatics this week.

perso
place
time

Bộ phân loại kiểm tra mỗi token (với nhãn từ loại (PoS), các token xung quanh ... như các đặc trưng cho việc phân thành hai lớp: begin-fact, end-fact)

Implemented in ELIE system (Finn/Kushmerick)

IE như là bài toán gán nhãn từng token trong chuỗi– Sequence Labeling

Ý tưởng chính:

Bộ phân loại xác định mỗi token thuộc về một **loại đặc biệt** thể hiện cho loại đối tượng (dựa vào các đặc trưng của token đang xét)

There will be a talk by Alan Turing at the CS Department at 4 PM .
O O O O O B-PER I-PER O O B-PLACE I-PLACE O B-TIME I-ITIME O

Prof. Dr. James D. Watson will speak on DNA at MPI on Thursday, Jan 12.

The lecture by Sir Francis Crick will be in the Institute of Informatics this week.

IOB annotation: một số biến thể: **IOBE**, **IOBES**, **IOE**

IOB Notation / CONLL format

ORG MISC
EU rejects **German** call to boycott **British** lamb .
MISC

PER
Peter Blackburn

LOC
BRUSSELS 1996-08-22

	token	postag	label	
0	EU	NNP	B-ORG	ORGANIZAION:EU
1	rejects	VBZ	O	
2	German	JJ	B-MISC	MISC:German
3	call	NN	O	
4	to	TO	O	
5	boycott	VB	O	
6	British	JJ	B-MISC	MISC: British
7	lamb	NN	O	
8	.	.	O	
9				
10	Peter	NNP	B-PER	PERSON: Peter Blackburn
11	Blackburn	NNP	I-PER	
12				
13	BRUSSELS	NNP	B-LOC	LOCATION: BRUSSELS
14	1996-08-22	CD	O	
15				

IOB annotation: một số biến thể: **IOBE**, **IOBES**, **IOE**

Relation Extraction: a task in IE

- Tìm mối quan hệ giữa các thực thể:

May 19 1995, Atlanta -- The Centers for Disease Control and Prevention, which is in the front line of the world's response to the deadly Ebola epidemic in Zaire , is finding itself hard pressed to cope with the crisis...

Information Extraction System

Date	*Disease Name*	*Location*
Jan. 1995	Malaria	Ethiopia
July 1995	Mad Cow Disease	U.K.
Feb. 1995	Pneumonia	U.S.

Relation Extraction: Method

- Manually engineered rules:
  - Rules defined over words/ Rules defined over words/entites entites:
    - "<company> located in located in <location>"
  - Rules defined over parsed text:
    - ((Obj) (Verb located) (*) (Subj))
- Machine Learning-based
  - Binary Relation Association as Binary Classification

Christos Faloutsos conferred with Ted Senator, the KDD 2003 General Chair.

Person Person Role

Person-Role (Christos Faloutsos, KDD 2003 General Chair) → NO

Person-Role (Ted Senator, KDD 2003 General Chair) → YES

Các cách tiếp cận

- Dựa trên luật (Rule-based)
- Các mô hình cho sequence labeling sử dụng học máy:
  - HMM
  - **CRF** *
  - RNN / LSTM
  - LSTM-CRF *
  - **BERT** *
  - Other deep learning models
- Kết hợp các trên tri thức ngôn ngữ học:
  - Đặc trưng từ loại (POS tagger), ngữ nghĩa (tuỳ theo bài toán)

Nhận diện thực thể có tên bằng CRF

Huấn luyện

Sử dụng và đánh giá

CRF: J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proc. of ICML, pp.282-289, 2001

Nhận diện thực thể có tên bằng CRF

- Các tập dữ liệu:
  - CONLL 2002 (Spain)
  - CONLL 2003 (English)
  - VLSP 2018 (Vietnamese)
  - Covid NER
    - https://github.com/VinAIResearch/PhoNER_COVID19
  - NER and slot filling:
    - https://github.com/VinAIResearch/JointIDSF
  - Nhãn ORG, PER, LOC, MISC, (DATE, NUM), …

Nhận diện thực thể có tên bằng CRF

- Các công cụ thư viện cần thiết:
  - CRF: CRF++ (C++),
  - Công cụ rút trích đặc trưng:
    - nltk : Tiếng Anh và các ngôn ngữ khác
    - Vncorenlp: Tiếng Việt
  - Tip : Nên biểu diễn dữ liệu đầu vào dưới dạng CONLL format

Các đặc trưng thường dùng cho NER

Bà **Hà** là một trong hai hộ nuôi bò_sữa đầu_tiên tại xã **Hoà_Bình**.

Extract features for all token/words of the input sentence

Feature Dict

{
Current word: "**Hà**",
Previous: "Bà",
Next word: "là",
…,
…
}

{
Current word: "Hoà_Bình",
Previous: "xã",
Next word: ".",
…,
…
}

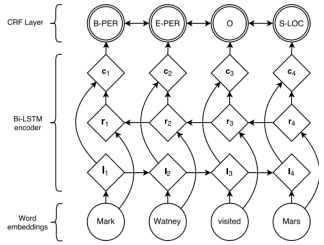
Các mô hình NER khác

- Sử dụng các mô hình học sâu:
  - Cho kết quả tốt vượt trội
- Một số mô hình tiêu biểu:
  - RNN / LSTM
  - LSTM-CRF *
  - BiLSTM-CRF *
  - BERT *
- Google: Tài liệu Hướng dẫn khá đầy đủ
  - https://github.com/guillaumegenthial/tf_ner
  - https://github.com/google-research/bert

Các mô hình NER khác

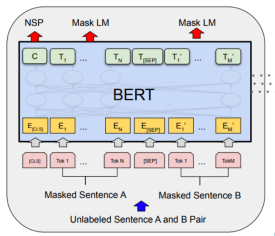
Neural Architectures for Named Entity Recognition

Guillaume Lample♠ Miguel Ballesteros♣♠
Sandeep Subramanian♠ Kazuya Kawakami♠ Chris Dyer♠
♠Carnegie Mellon University ♣NLP Group, Pompeu Fabra University
{glample,sandeeps,kkawakam,cdyer}@cs.cmu.edu,
miguel.ballesteros@upf.edu

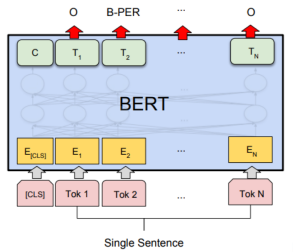


BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com



Pre-training



Fine-tuning

Tham khảo

- https://web.stanford.edu/class/cs124/lec/Information_Extraction_and_Named_Entity_Recognition.pdf
- https://www.slideshare.net/SHUBH177/text-categorization-72584299
- https://www.slideshare.net/kanimozhiu/text-datamining-txtcat
- https://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html

- https://www.coursera.org/learn/text-mining-analytics/lecture/QmxDT/2-1-description-of-possible-project-ideas
- https://www.coursera.org/learn/text-mining-analytics/home/week/3