Xử lý ngôn ngữ tự nhiên (Natural Language Processing)

Lê Thanh Hương Bộ môn Hệ thống Thông tin Viện CNTT &TT – Trường ĐHBKHN Email: huonglt-fit@mail.hut.edu.vn



Mục đích môn học



- Hiểu các nguyên tắc cơ bản và các cách tiếp cận trong XLNNTN
- Học các kỹ thuật và công cụ có thể dùng để phát triển các hệ thống hiểu văn bản hoặc nói chuyện với con người
- Thu được một số ý tưởng về các vấn đề mở trong XLNN

Tài liêu tham khảo

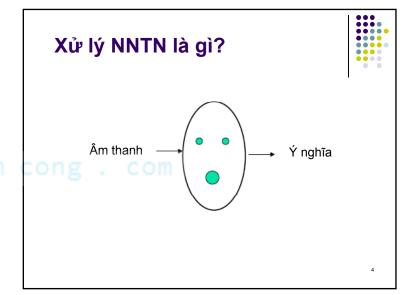
- Christopher Manning and Hinrich Schütze. 1999. Foundations of Statistical Natural Language Processing. The MIT Press.
- Dan Jurafsky and James Martin. 2000. Speech and Language Processing. PrenticeHall.
- James Allen. 1994. Natural Language Understanding. The Benjamins/Cummings Publishing Company Inc.

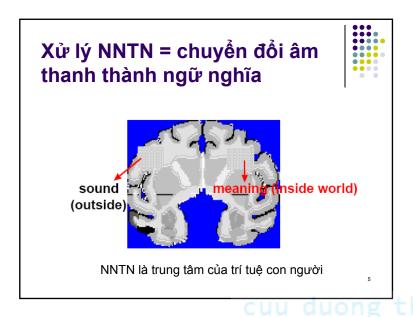
2

Thông tin chung



- Giữa kỳ: 30%
 - Điểm danh:10%
 - Đóng góp trong giờ học:10%
 - BTL: 10%
- Cuối kỳ: thi viết 70%
- Điều kiện được thi cuối kỳ:
 - Vắng mặt ít hơn 30% số lần điểm danh
 - Có tất cả các loại điểm giữa kỳ
- Bài tập lớn:
- Viết tiểu luận hoặc cài đặt chương trình theo nhóm (<= 4sv)
- Bảo vệ kết quả trong 2 tuần cuối của học kỳ
- Website: http://is.hut.vn/~huonglt/UNLP
- http://is.hut.edu.vn/~huonalt/UNLP





Xử lý NNTN là gì?







- Mục đích: hiểu được nhiều ngôn ngữ
- Không chỉ đơn giản là xử lý xâu hoặc so khớp từ khoá

6

ELIZA



- http://en.wikipedia.org/wiki/ELIZA
- http://www.cs.bham.ac.uk/research/projects/c ogaff/eliza/

Nói chuyện với Eliza



- E: how are you today?
- H: not so good
- E: why do you say not so good?
- H: i suffer from allergies?
- E: that's interesting
- H: i come out in spots
- E: i'm not sure i know what you're talking about
- . H: spots on my face
- E: your face
- H: that's right
- E: is it really that right?

8

2

Nói chuyện với Eliza



- H: you're a program, aren't you?
- E: does it please you to believe i am a program, am not i?
- H: it sure does
- E: i see
- H: yup
- E: a while back you talk about your face
- . H: what about your face?
- . E: what is it you really want to know?
- . H: are you my kind of girl?
- E: perhaps in your fantasies i am your kind of girl
- H: in my fantasies you are blonde

Nói chuyện với Eliza



- E: what makes you think i am blonde?
- H: i don't know
- E: don't you really know?
- H: no
- E: are you saying no just to be negative?
- H: no
- E: you are being a bit negative
- H: no i'm not
- E: how long have you been not?

1

Tại sao nghiên cứu XLNNTN



- Nghiên cứu cách con người xác định từ
- Nghiên cứu cách con người phân tích câu
- Nghiên cứu cách con người học một ngôn ngữ
- Nghiên cứu cách ngôn ngữ tiến hóa

Các chủ đề trong XLNNTN



- Mức phân tích: cú pháp, ngữ nghĩa, diễn ngôn, thực chứng, ...
- Các bài toán con: gán nhãn từ loại, PTCP, phân giải nhập nhằng từ, phânt ích cấu trúc diễn ngôn, ...
- Thuật toán và phương pháp: dựa trên tập ngữ liệu, dựa trên tri thức, ...
- Các ứng dụng: trích rút thông tin, phản hồi thông tin, dịch máy, hỏi đáp, hiểu ngôn ngữ tự nhiên, ...

12

Các mức phân tích



- Morphology (hình thái hoc): cách từ được xây dựng, các tiền tố và hâu tố của từ
- Syntax (cú pháp): mối liên hệ về cấu trúc ngữ pháp giữa các từ và ngữ
- Semantics (ngữ nghĩa): nghĩa của từ, cum từ, và cách diễn đàt
- Discourse (diễn ngôn): quan hệ giữa các ý hoặc các câu
- Pragmatic (thực chứng): mục đích phát ngôn, cách sử dụng ngồn ngữ trong giao tiếp
- World Knowledge (tri thức thế giới): các tri thức về thế giới, các tri thức ngầm

Hình thái học



- kick, kicks, kicked, kicking
- sit, sits, sat, sitting
- murder, murders v: nhồi nhét; n: những cái đã ăn, hẻm núi

Nhưng không phải lư rưc rõ à xóa đuôi.

- gorge, gorgeous-
- arm, army Quân đôi Cánh tay

Tiêng Việt: ngôn ngữ không biến hình, đơn âm tiết → cần tách từ

. . . .

Tách từ



15

- Một câu có thể có n khả năng tách từ, nhưng chỉ 1 trong chúng là đúng
- Giải pháp đơn giản: lấy chuỗi âm tiết dài nhất bắt đầu từ vị trí hiện tại và có trong từ điển từ
- Vấn đề: chồng chéo từ
 - Học sinh | học sinh | học.
 - Học sinh | học | sinh học.
- Liệt kê tất cả các khả năng có thể và thiết kế một giải pháp để lựa chọn cái tốt nhất

Gán nhãn từ loại



The boy threw a ball to the brown dog.

 The/DT boy/NN threw/VBD a/DT ball/NN to/IN the/DT brown/JJ dog/NN./.

DT – determiner từ chỉ định

danh từ, số ít hoặc số nhiều NN - noun

VBD - verb, past tense động từ, quá khứ

IN – preposition qiới từ JJ - adjective tính từ

. – dấu chấm câu













Gán nhãn từ loại



Con ngựa đá con ngựa đá.

- Con ngựa/DT đá/ĐgT con ngựa/DT đá/TT.
- Ông/ĐaT già/TT đi/Phó từ nhanh/TT quá/trạng từ.
- Ông già/DT đi/ĐgT nhanh/TT quá/trạng_từ.

Ngữ pháp: nhập nhằng cấu trúc (từ loại)



Time flies like an arrow.

Time // flies like an arrow.

VBZ giới từ so sánh (IN)

Time flies // like an arrow.

NNS VBP

cuu duong

Ngữ pháp: nhập nhằng cấu trúc (từ loai)

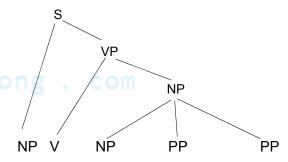


Ông già // đi nhanh quá.

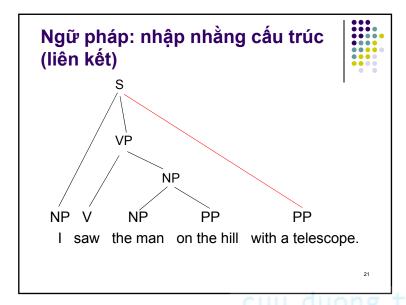
Ông // già đi nhanh quá.

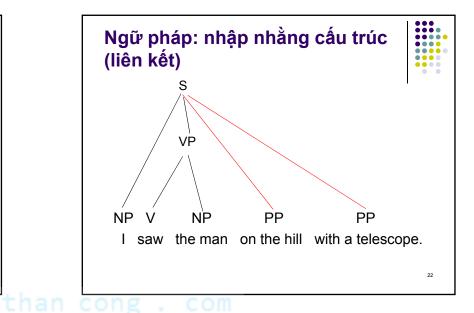
Ngữ pháp: nhập nhằng cấu trúc (liên kết)





saw the man on the hill with a telescope.





Nhưng ngữ pháp không nói lên nhiều điều...



- Colorless green ideas sleep furiously. [Chomsky]
- fire match arson hotel
- plastic cat food can cover CUU duong

I work for John Hancock ...

Ngữ nghĩa: nhập nhằng mức từ vưng



I walked to the bank ...

of the river.

to get money.

• The bug in the room ...

was planted by spies.

flew out the window.

and he is a good boss. which is a good company.

24

23

Diễn ngôn: đồng tham chiếu



President John F. Kennedy was assassinated.

The president was shot yesterday.

Relatives said that John was a good father.

JFK was the youngest president in history.

His family will bury him tomorrow.

Friends of the Massachusetts native will hold a candlelight service in Mr. Kennedy's home town.

25

Thực chứng



Bạn rút ra điều gì từ những điều tôi nói? Bạn phản ứng thế nào?

Luật hội thoại

- Bạn ơi mấy giờ rồi?
- Anh đưa cho em lo muối được không?

Nói kèm theo diễn tả

 Tôi cá với bạn 500.000 là đội Việt Nam sẽ thắng.

26

Tri thức thế giới



Mai đi ăn tối. Cô ấy gọi món bít tết. Cô ấy để lại tiền boa và về nhà.

- Mai ăn gì vào bữa tối?
- Ai mang bữa tối đến cho Mai?
- Ai làm bít tết?
- Mai có trả tiền không?

ciiaii

Tri thức về ngôn ngữ: Chúng ta biết gì về câu này?



- Các từ phải xuất hiện theo một trình tự nhất định:
- a. Chó kem ăn.b. Chó ăn kemCác bộ phân cấu thành câu:
 - chó = chủ ngữ (subject); ăn kem = vị ngữ (predicate)
- Ai làm gì cho ai:

chủ thể(chó), hành động(ăn), đối tượng(kem)

28

Các vấn đề khác?



- Hai câu "Mai nói chó ăn kem" và "Mai phủ nhận chó ăn kem" không logic với nhau
- Câu và thế giới: biết 1 câu là đúng hay sai có thể trong một vài trường hợp cụ thể nó đúng.
- "Tôi uống cà phê espresso sáng nay, nhưng Mai thông minh" không hợp lý

29

Tri thức ẩn



- 1. I want to solve the problem
- I wanna solve the problem
- 2. I understand these students
- These students I understand
- I want these students to solve the problem
- These students I want [x] to solve the problem
 - [x]=these students

30

Đặc trưng của ngôn ngữ



- Một số có thể nhớ được:
 - $\bullet \; \; \text{Singing} \rightarrow \text{Sing+ing}; \; \text{Bringing} \rightarrow \text{bring+ing}$
- Duckling → ?? Duckl +ing
- Cần phải biết duckl không phải là từ
- Nhưng không thể nhớ tất cả vì quá nhiều

31

Ngoài bộ nhớ, ta cần gì?



Số nhiều trong tiếng Anh:

- Toy+s -> toyz ; add z
- Book+s -> books; add s
- Church+s -> churchiz; add iz
- Box+s-> boxiz ; add iz
- Cần có hệ thống luật để sinh/xử lý các trường hợp này

32

8

"Phân tích" = gắn bề ngoài với cách biểu diễn trong của nó



- Vì sao XLNNTN khó: What makes NLP hard: không có tương ứng 1-1 với bất kỳ cách biểu diễn nào.
- Ta cần biết cấu trúc dữ liệu và thuật toán để thực hiện, mặc dù có thể xảy ra bùng nổ tổ hợp ở bất cứ công đoan xử lý nào

33

Phân tích câu hỏi LSAT / (former) GRE



- Sáu tượng điêu khắc C, D, E, F, G, H được triển lãm trong các phòng 1, 2, 3 của một triển lãm.
 - Tượng C và E có thể không trong cùng phòng.
 - Tượng D và G phía trong một phòng.
 - Nếu tượng E và F trong cùng phòng thì không có tượng nào khác trong phòng đó
 - Có íta nhất 1 tượng triển lãm trong một phòng, không có nhiều hơn 3 tượng trong bất cứ phòng nào
- Nếu tượng D được triển lãm trong phòng 3 và các tượng E, F trong phòng 1, trong các phát biểu dưới đây, phát biểu nào đúng:
 - A. Tượng C trong phòng 1
 - B. Tượng H trong phòng 1
 - Tượng G trong phòng 2
 - D. Tượng C và H trong cùng phòng
 - . Tượng G và F trong cùng phòng

34

Giải quyết đồng tham chiếu



U: A Bug's Life được chiếu tại chỗ nào của Mountain View?

S: A Bug's Life được chiếu ở rạp Summit.

U: Khi nào nó được chiếu ở đó?

S: Nó được chiếu lúc 2pm, 5pm, và 8pm.

U: Tôi muốn 1 người lớn, 2 trẻ con cho buổi chiếu đầu tiên. Nó giá bao nhiêu?

- Các nguồn tri thức:
 - Tri thức miền (Domain knowledge)
 - Tri thức về diễn ngôn (Discourse knowledge)
 - Tri thức thế giới (World knowledge)

35

Tại sao XLNNTN lại khó?



NNTN:

- Nhập nhằng tại mọi mức
- Phức tạp và mờ
- Liên quan lập luận về thế giới

36

9

Giải pháp

- Ta cần các công cu nào?
 - Tri thức về ngôn ngữ
 - Tri thức về thế giới
 - Cách kết hợp các tri thức
- Giải pháp tiềm năng:
 - Các mô hình xác suất xây dựng từ dữ liệu
 - P("maison" → "house") cao
 - P("L'avocat general" → "the general avocado") thấp

37

Nhắc lại các bài toán trong XLNNTN



- Vào: chuỗi ký tự
- Ra: các cặp (gốc từ, thẻ hình thái từ)
- Các vấn đề:
 - Kết hợp các thành phần cấu tạo nên từ
 - Loại hình thái từ (từ biến tố, từ phái sinh, từ ghép)
 - Ví dụ: quotations ~ quote/V + -ation(der.V->N) + NNS.

38

Phân tích cú pháp



- Vào: chuỗi các cặp (từ/từ loại)
- Ra: cấu trúc ngữ pháp của câu với các nút được gán nhãn (từ, từ loại, vai trò ngữ pháp)
- Vấn đề:
 - Quan hệ giữa từ, từ loại, và cấu trúc câu
 - Sử dụng nhãn cú pháp (Chủ ngữ, vị ngữ, bổ ngữ,)
 - Ví dụ: Tôi/ĐaT nhìn thấy/ĐgT Mai/DT
 - → ((Tôi/ĐaT)CN ((nhìn thấy/ĐgT) (Mai/DT)OBJ)VN)C

9

Ngữ nghĩa



- Vào: cấu trúc ngữ pháp của câu
- Ra: cấu trúc ngữ nghĩa của câu
- Ván đề:
 - Quan hệ giữa các đối tượng như chủ thể (Subject), đối tượng (Object), tác nhân (Agent), hậu quả (Effect) và các loại khác

((Học sinh/DT)CN ((học/ĐgT sinh học/DT)ĐgN)VN)C (Học sinh/DT)Sbj (học/ĐgT)action (sinh học/DT)Obj

40

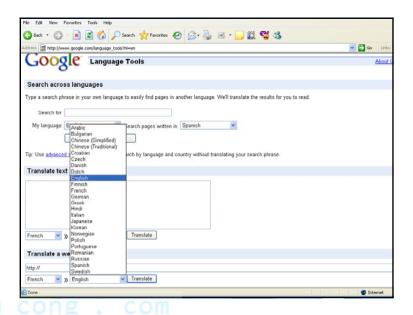
10

Các ứng dụng của XLNNTN

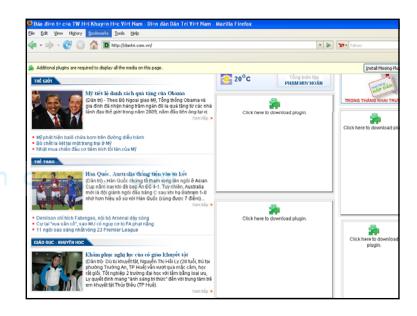
đáp (question answering)

- Khó: xử lý tiếng nói (speech processing), dich máy (machine translation), trích rúť thông tin (information extraction), giao diên
- Ứng dụng hiện nay: sửa lỗi chính tả, phân loai văn bản, ...

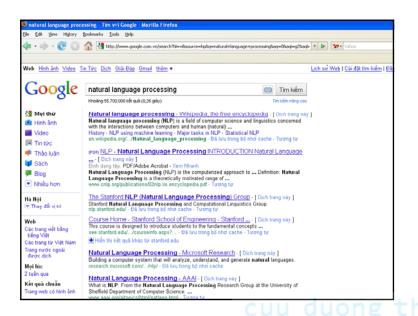
hôi thoai = NNTN (dialog interface), hỏi



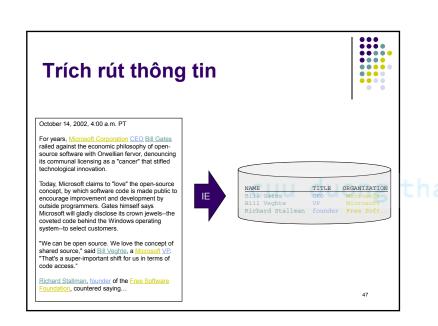
Inside the USA - Blog Archive - April Fools - Microsoft Internet ... 3 Back - 5 - 2 2 6 50 Search of Favorites Google This page was <u>automatically translated</u> from Fro April Fools April Fools r Jerome ITU = 02/04/2000, 09:22 , Classé dans : Humour, Politique US . osme ITU = 02/04/2008, 09:22. Filed under Fusky, U.S. pality La journée de mardi a été riche en poissons de tout genre The day Tuesday was rich in fish of all kinds. Dans la catégorie écolo, on nous a présenté le tout nouveau Air Force One, un In the green category, we introduced the brand new Air Force One, a hybr modèle hybride, "15 à 20% plus économique". 15 to 20% more economical." Dans la catégorie politique, Hillary fend l'armure et propose, au vu des récentes n the political category, Hillary fend armor and offers, given the recent per performances du sénateur, un défi au bowling à Obama pour décider du nominé of the senator, a challenge bowling to Obama to decide the Democratic no démocrate. Ainsi "les américains sauront que si le téléphone sonne à 3 heures du Thus 'the Americans know that if the phone rings at 3 o'clock in the morn matin, ils auront un président prêt à jouer au bowling des le premier jour". will have a president ready to play bowling from the first day." Dans la catégorie sport, c'est Chabal qui a fait les frais de l'humour du jour. Les In the sport category is Chabal who has borne the brunt of humour of the sites spécialisés ont relayé son départ dans la NFL américaine, aux New England pecialized sites have relayed his departure in the NFL American, the Ne Patriots, pour un contrat de 15 millions de dollars pour 3 ans. On attend toujours la Patriots, for a contract of 15 million dollars for 3 years. It is still awaiting onfirmation de l'homme qui soulève les foules en Nouvelle-Zélande. onfirmation from the man who raised the crowds in New Zealand. Et, enfin, dans la catégorie blog, Superfrenchie a révélé un pan de sa généalogie. Il And, finally, in the category blog, Superfrenchie revealed a pan of its gen serait apparenté à un certain... Bill O'Reilly. "My cousin Billy". Là, c'est gros quand rould be akin to a certain... Bill O'Reilly. "My cousin Billy." There is still hank you for your imagination, anyway. Merci pour cette imagination débordante, en tout cas. On the part of a visitor blocked in front of his screen all day, the fault of cu De la part d'un internaute bloqué devant son écran toute la journée, la faute à de maudits troubles digestifs. me anecdotes crispy, you who you are delivered to your workplace on

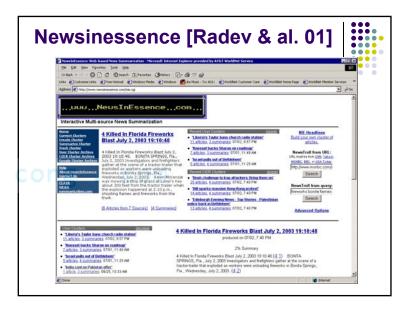


11

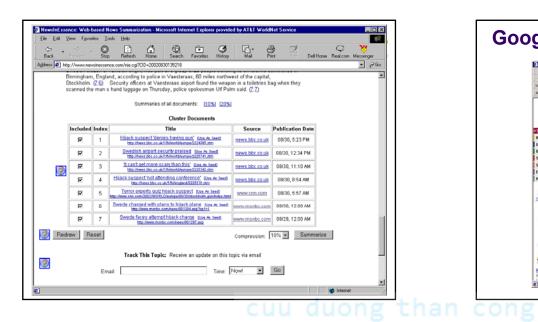








12





13

