

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@soict.hust.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 liệu 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.

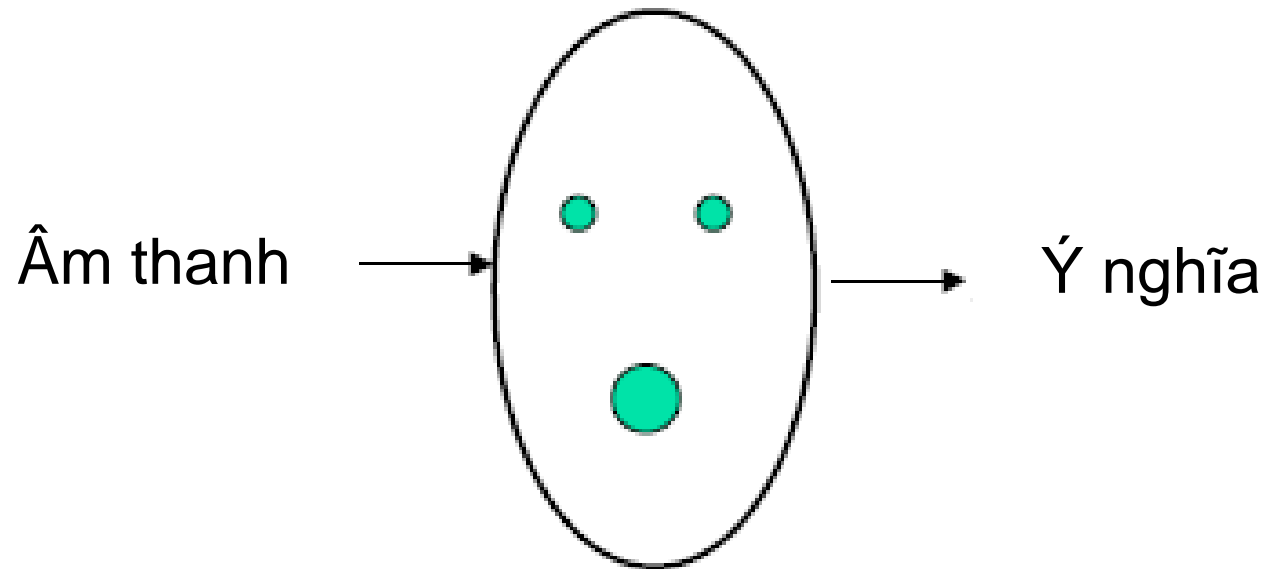
Thông tin chung



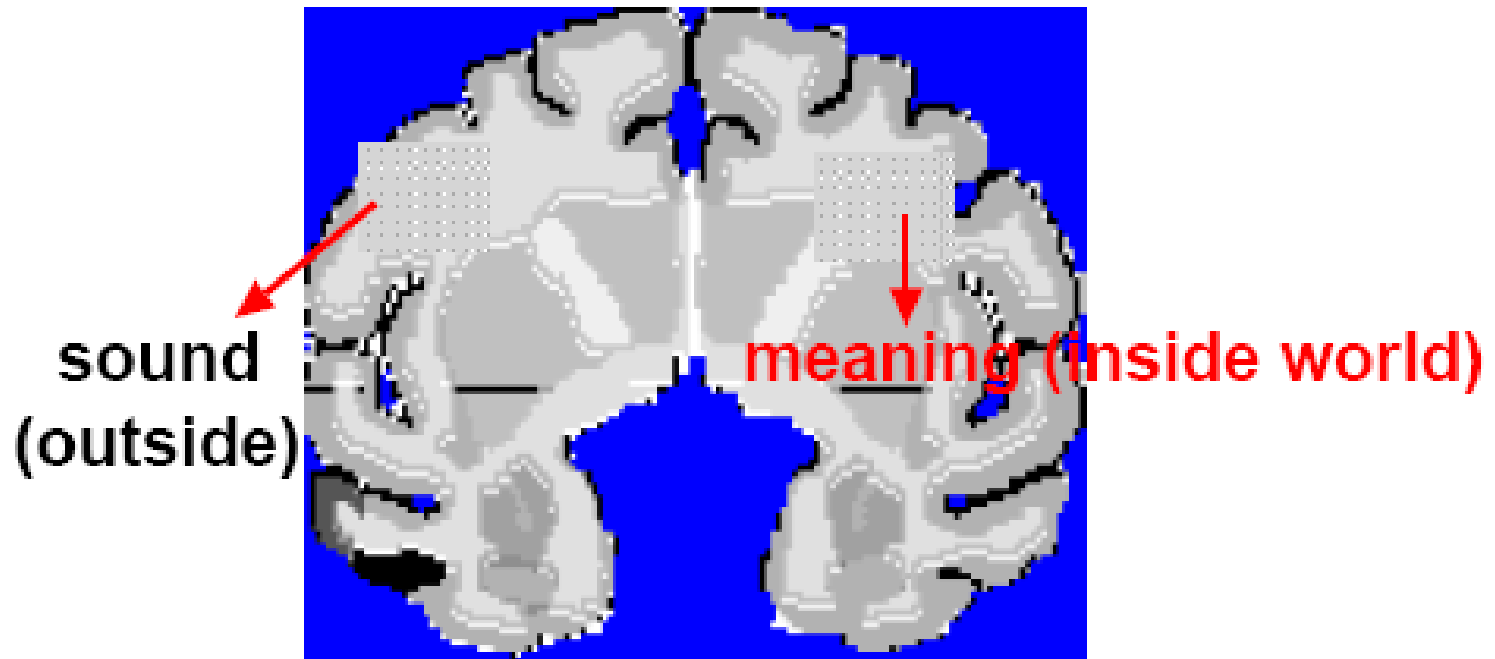
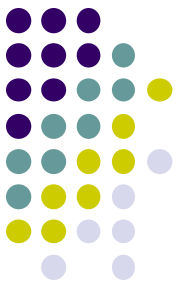
- **Đánh giá**
 - Giữa kỳ: 30%
 - Trung bình lên bảng: 15%
 - BTL: 15%
 - Cuối kỳ: thi viết 70%
 - Điều kiện được thi cuối kỳ:
 - Vắng mặt ít hơn 30% số lần điểm danh
 - Có điểm giữa kỳ
 - **Bài tập lớn:**
 - Viết tiểu luận hoặc cài đặt chương trình theo nhóm (≤ 4 sv)
 - Bảo vệ kết quả trong 2 tuần cuối của học kỳ
- **Website:** <https://users.soict.hust.edu.vn/huonglt/UNLP>



Xử lý NNTN là gì?



Xử lý NNTN = chuyển đổi âm thanh thành ngữ nghĩa



NNTN là trung tâm của trí tuệ con người

Xử lý NNTN là gì?



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



Các ứng dụng của XLNNTN



Language Tools

[About G](#)

Search across languages

Type a search phrase in your own language to easily find pages in another language. We'll translate the results for you to read.

Search for:

My language:

Search pages written in:

Tip: Use [advanced s](#)

Search by language and country without translating your search phrase.

Translate text

»

Translate a we

»

- Arabic
- Bulgarian
- Chinese (Simplified)
- Chinese (Traditional)
- Croatian
- Czech
- Danish
- Dutch
- English**
- Finnish
- French
- German
- Greek
- Hindi
- Italian
- Japanese
- Korean
- Norwegian
- Polish
- Portuguese
- Romanian
- Russian
- Spanish
- Swedish

Inside the USA » Blog Archive » April Fools - Microsoft Internet ...

File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites

Address <http://insidetheusa.net/2008/04/02/april-fools/> Go Links

Pennsylvanie Etats Unis
Visitez-vous Pennsylvanie? Comparez prix & critiques d'hôtels

Krankenversicherung USA
Unkomplizierter, hoher Kostenschutz vom US-Spezialisten! Div. Lösungen.

Announces Google

April Fools

par Jerome ITU ~ 02/04/2008, 09:22 . Classé dans : Humour, Politique US .

La journée de mardi a été riche en poissons de tout genre.

Dans la catégorie écolo, on nous a présenté le tout nouveau [Air Force One](#), un modèle hybride, "15 à 20% plus économique".

Dans la catégorie politique, Hillary fend l'armure et propose, au vu des récentes [performances](#) du sénateur, un défi [au bowling](#) à Obama pour décider du nominé démocrate. Ainsi "les américains sauront que si le téléphone sonne à 3 heures du matin, ils auront un président prêt à jouer au bowling dès le premier jour".

Dans la catégorie sport, c'est Chabal qui a fait les frais de l'humour du jour. Les sites spécialisés ont relayé son [départ dans la NFL](#) américaine, aux New England Patriots, pour un contrat de 15 millions de dollars pour 3 ans. On attend toujours la confirmation de l'homme qui soulève les foules en Nouvelle-Zélande.

Et, enfin, dans la catégorie blog, Superfrenchie a révélé un pan de sa [généalogie](#). Il serait apparenté à un certain... Bill O'Reilly. "My cousin Billy". Là, c'est gros quand même !

Merci pour cette imagination débordante, en tout cas.

De la part d'un internaute bloqué devant son écran toute la journée, la faute à de maudits troubles digestifs...

Internet

Translated version of <http://insidetheusa.net/2008/04/02/april...>

File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites

Address <http://translate.google.com/translate?u=http%3A%...> Go Links

Google™ This page was [automatically translated](#) from French. [View original web page](#) or mouse over text to view original I

Ads by GO

April Fools

by Jerome ITU ~ 02/04/2008, 09:22. Filed under: Funny, U.S. policy.

The day Tuesday was rich in fish of all kinds.

In the green category, we introduced the brand new [Air Force One](#), a hybrid "15 to 20% more economical."

In the political category, Hillary fend armor and offers, given the recent [performances](#) of the senator, a challenge [bowling](#) to Obama to decide the Democratic nominee. Thus "the Americans know that if the phone rings at 3 o'clock in the morning, they will have a president ready to play bowling from the first day."

In the sport category is Chabal who has borne the brunt of humour of the day. The specialized sites have relayed his [departure in the NFL](#) American, the New England Patriots, for a contract of 15 million dollars for 3 years. It is still awaiting confirmation from the man who raised the crowds in New Zealand.

And, finally, in the category blog, Superfrenchie revealed a pan of its [genealogy](#). It would be akin to a certain... Bill O'Reilly. "My cousin Billy." There is still big

Thank you for your imagination, anyway.

On the part of a visitor blocked in front of his screen all day, the fault of cursed digestive disorders...

Some anecdotes crispy, you who you are delivered to your workplace on Tuesday

Internet

Additional plugins are required to display all the media on this page.

Install Missing Plug

TH ̣ GI ̣I



Mỹ tiết lộ danh sách quà tặng của Obama

(Dân trí) - Theo Bộ Ngoại giao Mỹ, Tổng thống Obama và gia đình đã nhận hàng trăm ngàn đô la quà tặng từ các nhà lãnh đạo thế giới trong năm 2009, năm đầu tiên ông tại vị.

Xem tiếp

- Mỹ phát hiện balô chứa bom trên đường diễu hành
- Bò chết la liệt tại một trang trại ở Mỹ
- Nhật mua chiến đấu cơ tiêm kích tối tân của Mỹ

TH ̣ THAO



Hàn Quốc, Australia thắng tiến vào tứ kết

(Dân trí) - Hàn Quốc chứng tỏ tham vọng lên ngôi ở Asian Cup năm nay khi đè bẹp Ấn Độ 4-1. Tuy nhiên, Australia mới là đội giành ngôi đầu bảng C sau khi hạ Bahrain 1-0 nhờ hơn hiệu số so với Hàn Quốc (cùng được 7 điểm)...

Xem tiếp

- Denilson chỉ trích Fabregas, nội bộ Arsenal dậy sóng
- Cự lại "vua sân cỏ", sao MU có nguy cơ bị FA phạt nặng
- 11 ngôi sao sáng nhất vòng 23 Premier League

GI ̄O D ̄C - KHUY ̄N HỌC



Khâm phục nghị lực của cô giáo khuyết tật

(Dân trí) - Dù bị khuyết tật, Nguyễn Thị Hải Ly (28 tuổi, trú tại phường Trường An, TP Huế) vẫn vượt qua mặc cảm, học rất giỏi. Tốt nghiệp 2 trường đại học với tấm bằng loại ưu, Ly quyết định mang "ánh sáng tri thức" đến với trung tâm trẻ em khuyết tật Thủy Biều (TP Huế).

Xem tiếp



20°C

Tổng biên tập
PH ̄M HUY HO ̄N



Click here to download plugin.



Click here to download plu



Click here to download plugin.



Click here to download plugin.

quà tặng
H ̄P D ̄N

TRONG TH ̄NG KHAI TR ̄C



natural language processing



Tìm kiếm

Khoảng 55.700.000 kết quả (0,35 giây)

[Tìm kiếm nâng cao](#)

Mọi thứ

Hình ảnh

Video

Tin tức

Thảo luận

Sách

Blog

Nhiều hơn

Hà Nội

Thay đổi vị trí

Web

[Các trang viết bằng tiếng Việt](#)
[Các trang từ Việt Nam](#)
[Trang nước ngoài được dịch](#)

Mọi lúc

[2 tuần qua](#)

Kết quả chuẩn

[Trang web có hình ảnh](#)
[Natural language processing - Wikipedia, the free encyclopedia](#) - [[Dịch trang này](#)]

Natural language processing (NLP) is a field of computer science and linguistics concerned with the interactions between computers and human (natural) ...

History - NLP using machine learning - Major tasks in NLP - Statistical NLP

[en.wikipedia.org/.../Natural_language_processing](#) - [Đã lưu trong bộ nhớ cache](#) - [Tương tự](#)

[\[PDF\] NLP - Natural Language Processing INTRODUCTION Natural Language ...](#) - [[Dịch trang này](#)]

Định dạng tệp: PDF/Adobe Acrobat - [Xem Nhanh](#)

Natural Language Processing (NLP) is the computerized approach to ... Definition: **Natural Language Processing** is a theoretically motivated range of ...

[www.cnlp.org/publications/03nlp.lis.encyclopedia.pdf](#) - [Tương tự](#)

[The Stanford NLP \(Natural Language Processing\) Group](#) - [[Dịch trang này](#)]

Stanford **Natural Language Processing** and Computational Linguistics Group.

[nlp.stanford.edu/](#) - [Đã lưu trong bộ nhớ cache](#) - [Tương tự](#)

[Course Home - Stanford School of Engineering - Stanford ...](#) - [[Dịch trang này](#)]

This course is designed to introduce students to the fundamental concepts ...

[see.stanford.edu/.../courseinfo.aspx?...](#) - [Đã lưu trong bộ nhớ cache](#) - [Tương tự](#)

[+](#) [Hiển thị kết quả khác từ stanford.edu](#)

[Natural Language Processing - Microsoft Research](#) - [[Dịch trang này](#)]

Building a computer system that will analyze, understand, and generate **natural** languages.

[research.microsoft.com/.../nlp/](#) - [Đã lưu trong bộ nhớ cache](#)

[Natural Language Processing - AAAI](#) - [[Dịch trang này](#)]

What is **NLP**. From the **Natural Language Processing** Research Group at the University of Sheffield Department of Computer Science. ...

[www.aaai.org/aitopics/html/natlang.html](#) - [Tương tự](#)



"công nghệ thông tin"



Scholar

About 12,200 results (0.04 sec)

Articles

Tip: Search for **English** results only. You can specify your search language in [Scholar Settings](#).

Case law

Tội Phạm Trong Lĩnh Vực Công Nghệ Thông Tin

PV Lợi - Tội Phạm Trong Lĩnh Vực Công Nghệ Thông Tin, 2008 - lib.hpu.edu.vn

My library

Khái niệm, đặc điểm của tội phạm trong lĩnh vực **công nghệ thông tin**. Tình hình tội phạm và các quy định pháp luật về phòng chống tội phạm. Quan điểm và giải pháp đấu tranh phòng chống tội phạm trong lĩnh vực **công nghệ thông tin** ở nước ta. Công ước của hội

[Cite](#) [Save](#) [More](#)

Any time

Hệ thống thông tin quản lý của UPS trong chiến lược cạnh tranh cầu

AT Hoài - 2007 - 117.3.71.125

Since 2016

Since 2015

Since 2012

Custom range...

... Năm xuất bản: 2007. Nhà xuất bản: **Công nghệ thông tin**. Trích dẫn: Thông tin KHKT & Kinh tế Bưu điện. Tóm tắt: UPS (United Parcel Services) là một công ty chuyên phát bưu gửi đường bộ và đường không lớn nhất thế giới. Công ty này được thành lập vào năm 1907. ...

[Cite](#) [Save](#) [More](#)

Sort by relevance

Sort by date

[CITATION] Wireless power transfer: Principles and engineering explorations

KY Kim - 2012 - InTech

include patents

Cited by 25 [Related articles](#) [Cite](#) [Save](#)

Trích rút thông tin



Google™ [Advanced Search](#) [Preferences](#) [Language Tools](#) [Search Tips](#)
baker job opening

[Web](#) [Images](#) [Groups](#) [Directory](#) [News-Now!](#) Results

Searched the web for **baker job opening**

[Job Opening - Find ANY Job! - Search by Type, Industry & Geography](#)
[www.careerbuilder.com](#) Post Your RESUME Here to Reach Thousands of Employers - It's FREE!

[Job Opening At Flipdog.Com](#)
[www.FlipDog.com](#) Fetch your next **job** at FlipDog.com!

[Softimage::Community::Discussion Groups::ds.archive.0004](#)
... Le Rudulier; Drive space Ken Skaggs; Help about rendering denis.courtot; **JOB OPENING** ... Tony Cacciarelli; RE: ALE Karim Arbaoui; RE: omf to timeline Martin **Baker**; Re ...
[www.softimage.com/community/xsi/discuss/Archives/ds.archive.0004/default.htm](#) - 49k - [Cached](#) - [Similar pages](#)

[Softimage::Community::Discussion Groups::ds.archive.0004](#)
... Re: **JOB OPENING** Philip Herring - 2000/04/28 22:35. ... RE: omf to timeline Martin **Baker** - 2000/04/26 17:33. Re: omf to timeline adam - 2000/04/26 18:11. ...
[www.softimage.com/community/xsi/discuss/Archives/ds.archive.0004/ThreadIndex.htm](#) - 50k - [Cached](#) - [Similar pages](#)
[[More results from www.softimage.com](#)]

[CGI: Job Opening](#)
[www.genomics.cornell.edu/jobs/view_job.cfm?id=10](#) - 15k - [Cached](#) - [Similar pages](#)

[Information Activist Job Opening - May 2001](#)
[www.igc.org/datacenter/job.html](#) - 6k - [Cached](#) - [Similar pages](#)

[Post an Employee Benefits Job Opening \(Help Wanted\) Ad](#)
... edit the ad to add a new **job opening** ... as possible when it is emailed to 2,985 **job** ... jobs/posthelpwanted.shtml
Webmaster: [webmaster@BenefitsLink.com](#) (Dave **Baker** ...
[www.benefitslink.com/jobs/posthelpwanted.shtml](#) - 24k - [Cached](#) - [Similar pages](#)

[Post an Employee Benefits Job Opening \(Help Wanted\) Ad](#)
Employee Benefits Jobs! Brought to you by BenefitsLink (tm) and its EmployeeBenefitsJobs.com (tm) division.
[www.benefitslink.com/jobs/pricinginfo.shtml](#) - 7k - [Cached](#) - [Similar pages](#)
[[More results from www.benefitslink.com](#)]

Martin Baker, a person

Genomics job

Employers job posting form

Trích rút thông tin



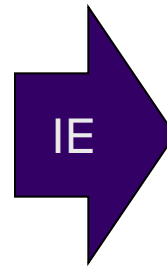
October 14, 2002, 4:00 a.m. PT

For years, [Microsoft Corporation](#) [CEO Bill Gates](#) railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said [Bill Veghte](#), a [Microsoft VP](#). "That's a super-important shift for us in terms of code access."

[Richard Stallman](#), [founder](#) of the [Free Software Foundation](#), countered saying...



NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft..



Dan Jurafsky



Information Extraction & Sentiment Analysis



Attributes:

zoom
affordability
size and weight
flash
ease of use

Size and weight

- ✓ • nice and compact to carry!
- since the camera is small and light, I won't need to carry around those heavy, bulky professional cameras either!
- the camera feels flimsy, is plastic and very light in weight you have to be very delicate in the handling of this camera

NewsInEssence [Radev & al. 01]



NewsInEssence: Web-based News Summarization - Microsoft Internet Explorer provided by AT&T WorldNet Service

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites History Print

Links Customize Links Free Hotmail Windows Media Windows Like Music - Try AOL WorldNet Customer Care WorldNet Home Page WorldNet Member Services

Address http://www.newsinesence.com/nie.cgi

...www..NewsInEssence..com...

Interactive Multi-source News Summarization

Home
Current Clusters
Create Cluster
Summarize Cluster
Track Cluster
User Cluster Archive
CIDR Cluster Archive
Google Cluster Archive

Help
About News InEssence
Contact Us

CLAIR
MEAD
summarization.com

4 Killed In Florida Fireworks Blast July 2, 2003 19:10:48

4 Killed In Florida Fireworks Blast July 2, 2003 19:10:48. BONITA SPRINGS, Fla., July 2, 2003 Investigators and firefighters gather at the scene of a tractor-trailer that exploded as workers were unloading fireworks in Bonita Springs, Fla., Wednesday, July 2, 2003. Kevin McKenzie was mowing a strip of grass at Lover's Key about 300 feet from the tractor trailer when the explosion happened at 2:10 p.m., shooting flames and fireworks from the truck.

[8 Articles from 7 Sources] [4 Summaries]

Recent User Clusters (more)

- 'Liberia's Taylor bans church radio station'
11 articles, 3 summaries: 07/02, 9:57 PM
- 'Knesset backs Sharon on roadmap'
7 articles, 3 summaries: 07/01, 11:48 AM
- 'Israel pulls out of Bethlehem'
5 articles, 4 summaries: 07/01, 11:25 AM

Recent CIDR Clusters (more)

- 'Bush challenge to Iraq attackers: Bring them on'
25 articles, 4 summaries: 07/02, 7:40 PM
- 'Bill sparks massive Hong Kong protest'
14 articles, 4 summaries: 07/02, 7:40 PM
- 'Edinburgh Evening News - Top Stories - Palestinian police back in Bethlehem'
13 articles, 4 summaries: 07/02, 7:40 PM

NIE Headlines

Build your own cluster of articles.

NewsTroll from URL:
URL must be from CNN, Yahoo!, MSNBC, BBC, or USA Today.

NewsTroll from query:

[Advanced Options](#)

User Clusters (Archive)

- 'Liberia's Taylor bans church radio station'
11 articles, 3 summaries: 07/02, 9:57 PM
- 'Knesset backs Sharon on roadmap'
7 articles, 3 summaries: 07/01, 11:48 AM
- 'Israel pulls out of Bethlehem'
5 articles, 4 summaries: 07/01, 11:25 AM
- 'India cool on Pakistan offer'
1 article, 3 summaries: 06/25, 10:33 AM

4 Killed In Florida Fireworks Blast July 2, 2003 19:10:48

produced on 07/02, 7:40 PM

2% Summary

4 Killed In Florida Fireworks Blast July 2, 2003 19:10:48 (4:1) BONITA SPRINGS, Fla., July 2, 2003 Investigators and firefighters gather at the scene of a tractor-trailer that exploded as workers were unloading fireworks in Bonita Springs, Fla., Wednesday, July 2, 2003. (4:2)

Done Internet

Birmingham, England, according to police in Vaesteraas, 60 miles northwest of the capital, Stockholm. (Z:6) Security officers at Vaesteraas airport found the weapon in a toiletries bag when they scanned the man s hand luggage on Thursday, police spokesman Ulf Palm said. (Z:Z)

Summaries of all documents: [10%] [20%]

Cluster Documents

Included	Index	Title	Source	Publication Date
<input checked="" type="checkbox"/>	1	Hijack suspect 'denies having gun' [Use As Seed] http://news.bbc.co.uk/1/hi/world/europe/2224395.stm	news.bbc.co.uk	08/30, 5:23 PM
<input checked="" type="checkbox"/>	2	Swedish airport security praised [Use As Seed] http://news.bbc.co.uk/1/hi/world/europe/2225741.stm	news.bbc.co.uk	08/30, 12:34 PM
<input checked="" type="checkbox"/>	3	'It can't get more scary than this' [Use As Seed] http://news.bbc.co.uk/1/hi/world/europe/2225342.stm	news.bbc.co.uk	08/30, 11:10 AM
<input checked="" type="checkbox"/>	4	Hijack suspect 'not attending conference' [Use As Seed] http://news.bbc.co.uk/1/hi/england/2225318.stm	news.bbc.co.uk	08/30, 8:54 AM
<input checked="" type="checkbox"/>	5	Terror experts quiz hijack suspect [Use As Seed] http://www.cnn.com/2002/WORLD/europe/08/30/stockholm.gun/index.html	www.cnn.com	08/30, 5:57 AM
<input checked="" type="checkbox"/>	6	Swede charged with plans to hijack plane [Use As Seed] http://www.msnbc.com/news/801304.asp?cp1=1	www.msnbc.com	08/30, 12:00 AM
<input checked="" type="checkbox"/>	7	Swede faces attempt hijack charge [Use As Seed] http://www.msnbc.com/news/801297.asp	www.msnbc.com	08/29, 12:00 AM

Compression:

Track This Topic: Receive an update on this topic via email

Email: Time:

Google News [02]



Google News - Microsoft Internet Explorer provided by AT&T WorldNet Service

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites History Print Copy Paste

Links Customize Links Free Hotmail Windows Media Windows Like Music - Try AOL WorldNet Customer Care WorldNet Home Page WorldNet Member Services

Address http://news.google.com/ Go

Google News BETA

Web Images Groups Directory News

Search News Search the Web

Google named best News service by Webby Awards

Top Stories Auto-generated 7 minutes ago

World

U.S.

Business

Sci/Tech

Sports

Entertainment

Health

Make Google News Your Homepage

Text Version

About Google News

Top Stories

Palestinians Resume Control of Bethlehem
Washington Post - 40 minutes ago
BETHLEHEM, West Bank, July 2 -- Israeli troops pulled out of this biblical West Bank town today and turned over control to Palestinian security forces, who raised flags and patrolled the city's historic Manger Square. The hand-over was the latest step ...
[US Praises Bethlehem Handover by Israel](#) Reuters
[Israel releases eight Palestinian prisoners](#) SABC News
[Forward - Guardian - Christian Science Monitor - International Herald Tribune - and 2037 related »](#)

Pentagon readies plans for peace mission in Liberia
Minneapolis Star Tribune - 1 hour ago
WASHINGTON, DC -- The Pentagon has ordered military planners to prepare detailed options for US troops to join an international peacekeeping force in Liberia, two senior military officials said Wednesday.
[Bush May Send Troops To Liberia](#) WCCO
[US to send troops to Liberia](#) Guardian
[CNN - Men's News Daily - CBC News - ABC News - and 758 related »](#)

WorldCom Offers New Stock With Settlement
Washington Post - [and 83 related »](#)

3 Top Shuttle Managers Replaced
Washington Post - [and 306 related »](#)

Attorney says unconfirmed statements about missing Baylor player feeding media frenzy
News 24 Houston - [and 1502 related »](#)

Schwarzenegger takes "T3" act to troops in Gulf
Boston.com - [and 247 related »](#)

Food giant maps steps to fight rising obesity
International Herald Tribune - [and 108 related »](#)

In the News

[New Jersey](#) [San Diego](#)
[Homeland Security](#) [Defence Secretary](#)
[Los Angeles](#) [Civil War](#)
[North Carolina](#) [Coca Cola](#)
[US Supreme Court](#) [West Bank](#)

World »

Kenyan Women Accuse British Troops of Rape
New York Times - 3 hours ago
LONDON, July 2. A group of 650 Kenyan women who say they

U.S. »

Low-Profile Suburb Braces for Onslaught
Washington Post - 40 minutes ago
CHESAPEAKE, Va., July 2. At first a Fairfax County judge

Done Internet



TEXTMAP
THE ENTITY SEARCH ENGINE

Monitoring the World So You Don't Have To ...

ENTITIES

SOL

Search!

[TextMap](#) : [TextMed](#) : [Textblq](#) : [TextBiz](#) : [Make homepage!](#) : [Link to us](#) : [Help?](#)

Question Answering

Wednesda

in what year did John Lennon die?

Answer: 1980

[[The Beatles Anthology](#) 02/28/2006 [wiki](#)]

ELIZA



- <http://en.wikipedia.org/wiki/ELIZA>
- <http://www.cs.bham.ac.uk/research/projects/cogaff/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?



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?

Các kết quả đạt được



Dan Jurafsky

Language Technology



making good progress

mostly solved

Spam detection

Let's go to Agra! ✓

Buy V1AGRA ... ✗

Part-of-speech (POS) tagging

ADJ ADJ NOUN VERB ADV

Colorless green ideas sleep furiously.

Named entity recognition (NER)

PERSON ORG LOC

Einstein met with UN officials in Princeton

Sentiment analysis

Best roast chicken in San Francisco! 👍

The waiter ignored us for 20 minutes. 👎

Coreference resolution

Carter told Mubarak he shouldn't run again.

Word sense disambiguation (WSD)

I need new batteries for my *mouse*.

Parsing

I can see Alcatraz from the window!

Machine translation (MT)

第13届上海国际电影节开幕...

The 13th Shanghai International Film Festival...

Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30

Party
May 27
add

still really hard

Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

Summarization

The Dow Jones is up

The S&P500 jumped

Housing prices rose

Economy is good

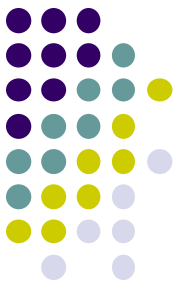
Dialog

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?



- Một số ứng dụng đang được quan tâm
 - Phân tích nhu cầu người dùng (user intend) trong thương mại điện tử
 - Phân tích quan điểm người dùng
 - Phát hiện sự kiện
 - Tóm tắt đơn/đa văn bản
 - Trích rút thông tin
- Xu hướng:
 - Deep learning
 - Word embedding



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, phân tích cú pháp, phân giải nhập nhằng từ, phân tích 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, ...



Các mức phân tích

- **Morphology (hình thái học):** cách từ được xây dựng, các tiền tố và hậu 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, cụm từ, và cách diễn đạt
- **Discourse (diễn ngôn):** quan hệ giữa các ý hoặc các câu
- **Pragmatic (thực chứng):** mục đích phát ngôn, cách sử dụng ngôn ngữ trong giao tiếp
- **World Knowledge (tri thức thế giới):** các tri thức về thế giới, các tri thức ngầm



Hình thái học

Tiếng Anh: ngôn ngữ biến hình, đa âm tiết

- kick, kicks, kicked, kicking
- sit, sits, sat, sitting
- murder, murders

v: nhồi nhét; n: những cái đã ăn, hẻm núi

Nhưng không phải lúc **rực rỡ** là xóa đuôi.

- gorge, gorgeous
- arm, army

Cánh tay

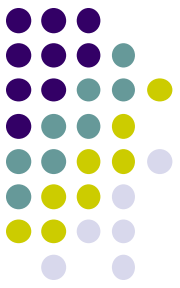
Quân đội

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



Tách từ

- 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

NN – noun,

danh từ, số ít hoặc số nhiều

VBD – verb, past tense

động từ, quá khứ

IN – preposition

giới 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/DT.
- 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 IN (giới từ so sánh)

Time flies // like an arrow.
 NNS VBP

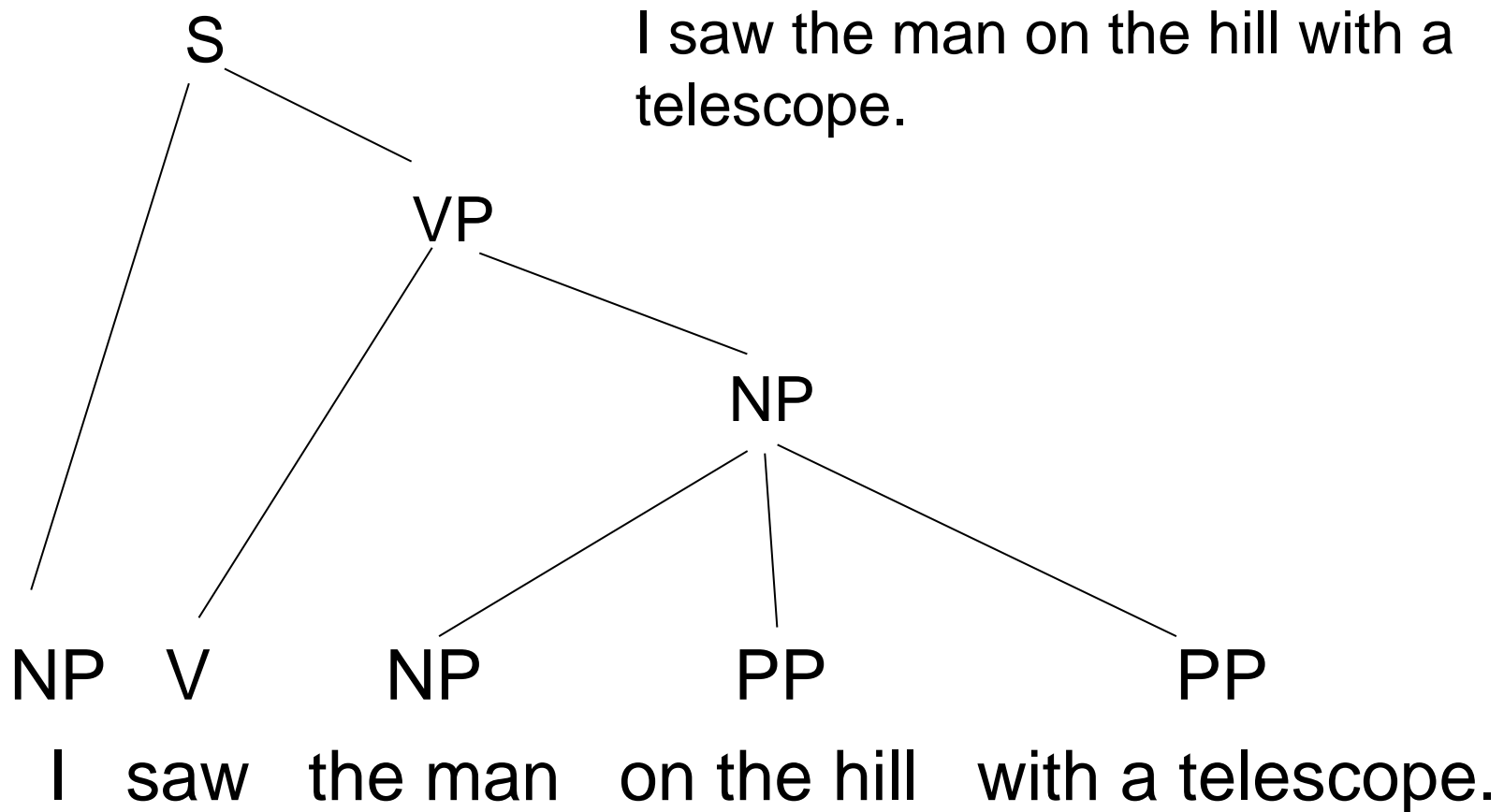
Ngữ pháp: nhập nhằng cấu trúc (từ loại)



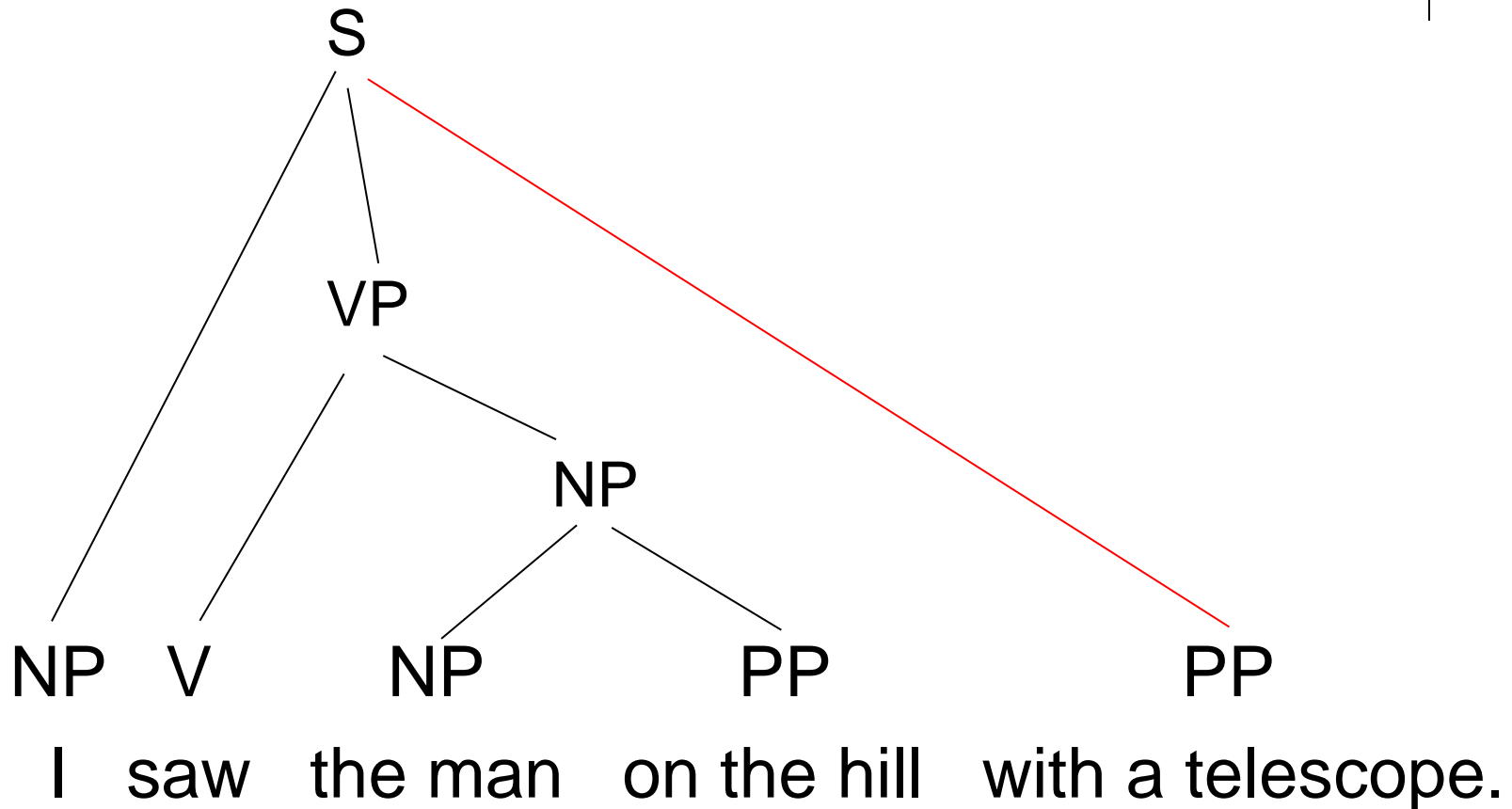
Ông già // đi nhanh quá.

Ông // già đi nhanh quá.

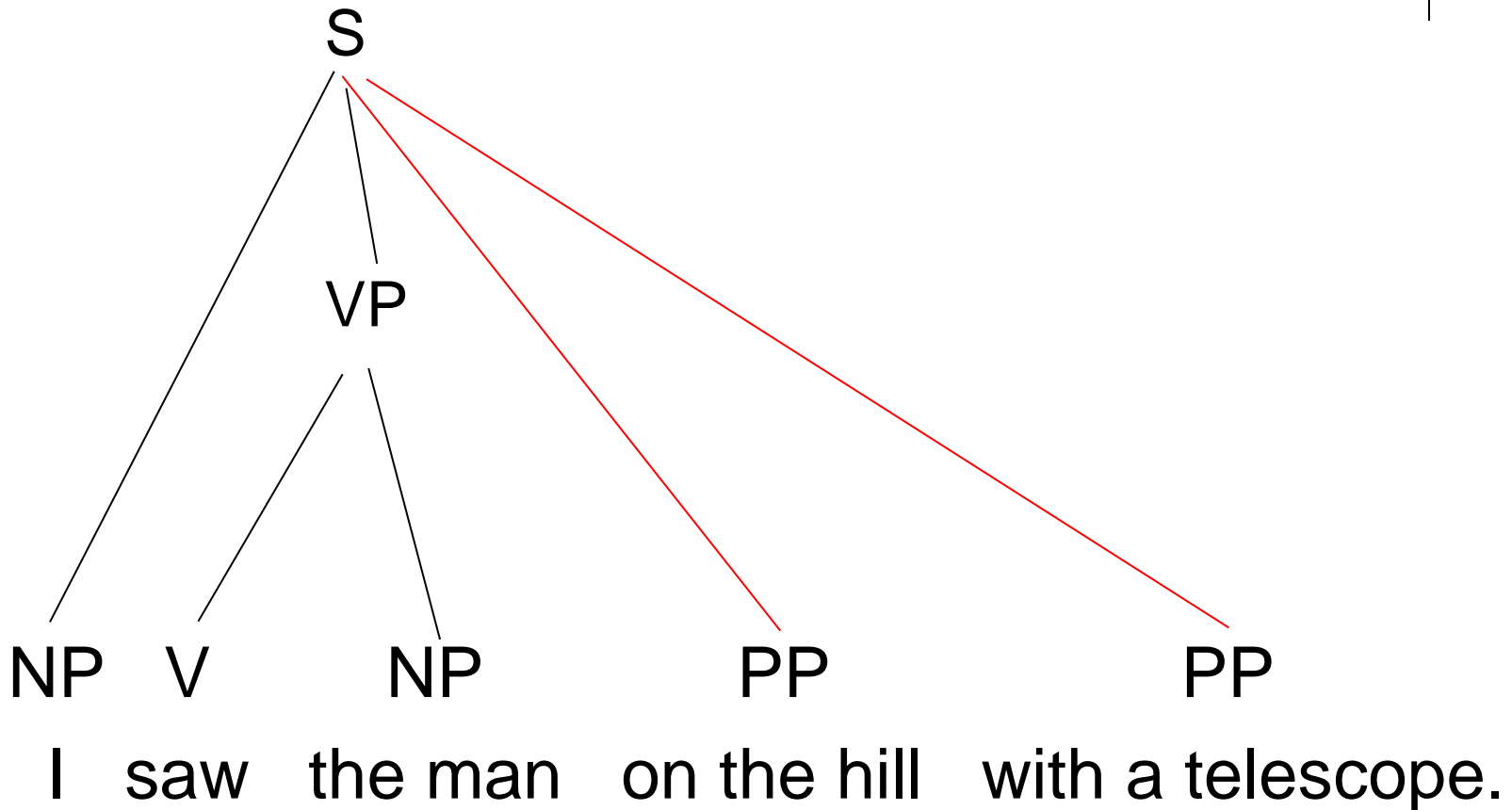
Ngữ pháp: nhập nhằng cấu trúc (liên kết)



Ngữ pháp: nhập nhằng cấu trúc (liên kết)



Ngữ pháp: nhập nhằng cấu trúc (liên kết)



Nhưng ngữ pháp không nói lên nhiều điều...



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

Ngữ nghĩa: nhập nhằng mức từ vựng



- I walked to the bank ...
of the river.
to get money.
- The bug in the room ...
was planted by spies.
flew out the window.
- I work for John Hancock ...
and he is a good boss.
which is a good company.

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.



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 lọ muối được không?

Nói kèm theo diễn tả

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

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?

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 kem
- Các bộ phận 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)

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



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

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 phải trong một phòng.
 - Nếu tượng E và F trong cùng phòng thì không có tượng nào khác trong phòng đó
 - Có ít 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
 - C. Tượng G trong phòng 2
 - D. Tượng C và H trong cùng phòng
 - E. Tượng G và F trong cùng phòng

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)



Đặc trưng của ngôn ngữ

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



Ngoài bộ nhớ, ta cần gì?

Số nhiều trong tiếng Anh:

- Toy+s -> toyz ; add z
- Book+s -> books ; add s
- Box+s-> boxes ; add es

➤ ***Cần có hệ thống luật để sinh/xử lý các trường hợp này***

Đặc điểm XLNNTN



NNTN:

- Nhập nhằng tại mọi mức
- Liên quan lập luận về thế giới



Giải pháp

- Ta cần các công cụ 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
 - $P(\text{"maison"} \rightarrow \text{"house"})$ **cao**
 - $P(\text{"L'avocat general"} \rightarrow \text{"the general avocado"})$ **thấp**

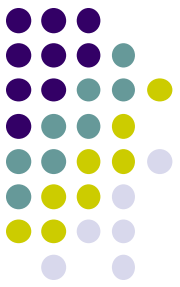


Nhắc lại các bài toán cơ bản trong XLNNTN



Phân tích hình thái từ

- Vào: chuỗi ký tự
- Ra: các cặp (gốc từ, thể 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.



Tách từ

- Vào: chuỗi ký tự
- Ra: các khả năng tách từ của xâu đầu vào
- Các vấn đề:
 - Các dạng đặc biệt không phải từ
 - Nhập nhằng từ



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



Ngữ nghĩa

- Vào: cấu trúc ngữ pháp của câu
- Ra: cấu trúc ngữ nghĩa của câu
- Vấn 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