

Natural Language Processing

Lê Thanh Hương

School of Information and Communication Technology - HUST

Email: huonglt@soict.hust.edu.vn

Course Goals

- Learn the **basic principles** and **theoretical approaches** underlying natural language processing
- Learn **techniques** and **tools** which can be used to develop practical, robust systems that can (partly) understand text or communicate with users in one or more languages
- Gain insight into many of the **open research problems** in natural language

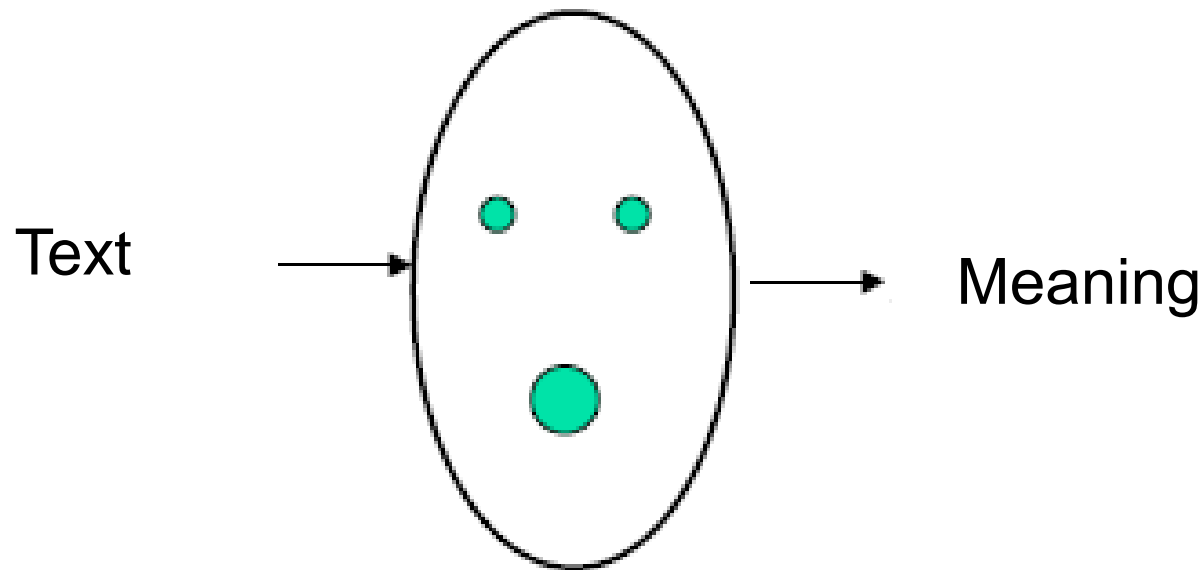
References

- 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.

- **Evaluate**

- Midterm: 40%
 - Continuous assessments: 20%
 - Group project: 20%
- Final Exam: 60%
- Group project :
 - Research papers (≤ 2 students) or implement an NLP tool (≤ 4 students)
 - Defend the project at the last three weeks of the semester

What is NLP?



What is NLP?







- Target: understand multi languages
- Not simple as text matching or keyword matching


Applications of NLP

Machine Translation

Language Tools - Microsoft Internet Explorer

File Edit View Favorites Tools Help

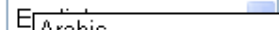

Address  http://www.google.com/language_tools?hl=en  Go  Links 

 **Language Tools** [About Google](#)

Search across languages

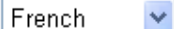

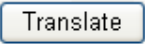
Type a search phrase in your own language to easily find pages in another language. We'll translate the results for you to read.

Search for:

My language:  Search pages written in:  Spanish

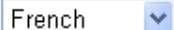


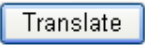
Tip: Use [advanced search](#) to search by language and country without translating your search phrase.

Translate text

 French  

Translate a web page

http://

 French   English 

- 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

Machine Translation

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

File Edit View Favorites Tools Help

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 nommé 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

Address <http://translate.google.com/translate?u=http%3A%2F%2Finsidetheusa.net%2F2008%2F04%2Fapril...> Go Links

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

Ads by Google

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 model, "15 to 20% more economical."

In the political category, Hillary fend armor and offers, given the recent [performance](#) 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

Text Categorization

Báo điện tử của TW Hội Khuy ến Học Vi ệt Nam - Di ền đ ầu Dân Trí Vi ệt Nam - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://dantri.com.vn/

Y Yahoo

TH Ế GI ỚI

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 Ể 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
Cựu lại "vua sân cỏ", sao MU có nguy cơ bị FA phạt nặng
11 ngôi sao sáng nhất vòng 23 Premier League

GI ẢO D ỤC - KHUY ẾN H ỌC

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
PHẠM HUY HOÀN

Click here to download plugin.

Click here to download plugin.

Click here to download plugin.

Click here to download plugin.

Click here to download plugin.

Click here to download plugin.

Information Retrieval

natural language processing - Tìm với Google - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.google.com.vn/search?hl=vi&source=hp&q=natural+language+processing&aq=0&aq=g2&aq=

Web Hình ảnh Video Tin tức Dịch Giải Đáp Gmail thêm

Lịch sử Web Cài đặt tìm kiếm Đăng nhập

Google

natural language processing

Khoảng 55.700.000 kết quả (0,35 giây)

Tim kiếm

Tim 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/D3nlp.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.aaii.org/aitonine/html/natlang.html](#) - Tương tự

Information Retrieval

Information Retrieval

"công nghệ thông tin" - Google Scholar - Mozilla Firefox

"công nghệ thông tin" - G... X +

← ⓘ 🔒 https://scholar.google.com.vn/scholar?hl=en&q="công+nghệ+thông+tin"&btn Search ☆ 📁 ⬇

Web Images More... huong

Google "công nghệ thông tin" 🔍

Scholar About 12,200 results (0.04 sec) ✎

Articles

Case law

My library

Any time

Since 2016

Since 2015

Since 2012

Custom range...

Sort by relevance

Sort by date

☒ include patents

Tip: Search for **English** results only. You can specify your search language in **Scholar Settings**.

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
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

Hệ thống thông tin quản lý của UPS trong chiến lược cạnh tranh cầu
AT Hoài - 2007 - 117.3.71.125
... 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 chuyển phát bưu gửi đường bộ và đường không lớn nhất thế giới. Công ty này được thành lập vào năm 1907. ...
Cite Save More

[CITATION] Wireless power transfer: Principles and engineering explorations
KY Kim - 2012 - InTech
Cited by 25 Related articles Cite Save

11

Information Extraction

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

[Web](#) [Images](#) [Groups](#) [Directory](#) [News-Now!](#)
Searched the web for **baker job opening** Results

[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

Information Extraction

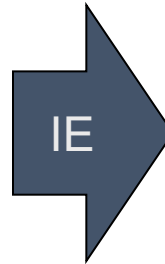
October 14, 2002, 4:00 a.m. PT

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

Text Summarization

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

Address <http://www.newsinesence.com/nie.cgi> Go

...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 In Essence](#)
[Contact Us](#)

[CLAIR](#)
[HEAD](#)
[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. [\(4:2\)](#)

Text Summarization

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

File Edit View Favorites Tools Help

Address  http://www.newsinessence.com/nie.cgi?CID=20020830135218

Birmingham, England, according to police in Vaesteraas, 60 miles northwest of the capital, Stockholm. (7:6) 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. (7:7)

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/WWWORLD/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



Redraw

Reset

Compression:

Summarize

Text Summarization

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

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

World

U.S.

Business

Sci/Tech

Sports

Entertainment

Health

Make Google News Your Homepage

Text Version

Auto-generated 7 minutes ago

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](#)

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 least a Fairfax County judge's

Done Internet

17

Question Answer

TextMap - The Entity Search Engine - Newspaper Analy...

File Edit View Favorites Tools Help

Address http://www.textmap.com/cgi-bin/qa_main.cgi?question=In+wt Go Links

 **TEXTMAP** Monitoring the World So You Don't Have To ...
THE ENTITY SEARCH ENGINE

ENTITIES SOL

Search!

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

Question Answering Wednesday

in what year did John Lennon die?

Answer: 1980

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

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

Talk with 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?

Talk with 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

Talk with 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?

Current achievements of NLP

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



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?

Some interested applications

- Analyze user opinion
- Event detection
- Single/multi-document summarization
- Information extraction
- Machine translation
- Question answering
- Current techniques:
 - Deep learning
 - Word embedding

Levels of Analysis

- Morphology: how words are constructed; prefixes & suffixes
- Syntax: structural relationships between words
- Semantics: meanings of words, phrases, and expressions
- Discourse: relationships across different sentences or thoughts; contextual effects
- Pragmatic: the purpose of a statement; how we use language to communicate
- World Knowledge: facts about the world at large; common sense

Morphology

English: metamorphic, polysyllabic language

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

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

But it's not just **rực rỡ** as adding and deleting endings..

- gorge, gorgeous
- arm, army

Cánh tay

Quân đội

Vietnamese: untransforming, monosyllabic language → need word segmentation

Word Segmentation

- A sentence can have n possibilities of word segmentation, but only one of them is correct.
- Simple solution: get the longest syllable chain from the current position. The chain is in the dictionary.
- Problem: overlapping
 - Học sinh | học sinh | học.
 - Học sinh | học | sinh học.
- ☞ List all possibilities and propose a method to select the best possibility.

Past of Speech Tagging

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

Past of Speech Tagging

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.

Syntax: structural ambiguity (part of speech)

Time flies like an arrow.

Time // flies like an arrow.
 VBZ IN (giới từ so sánh)

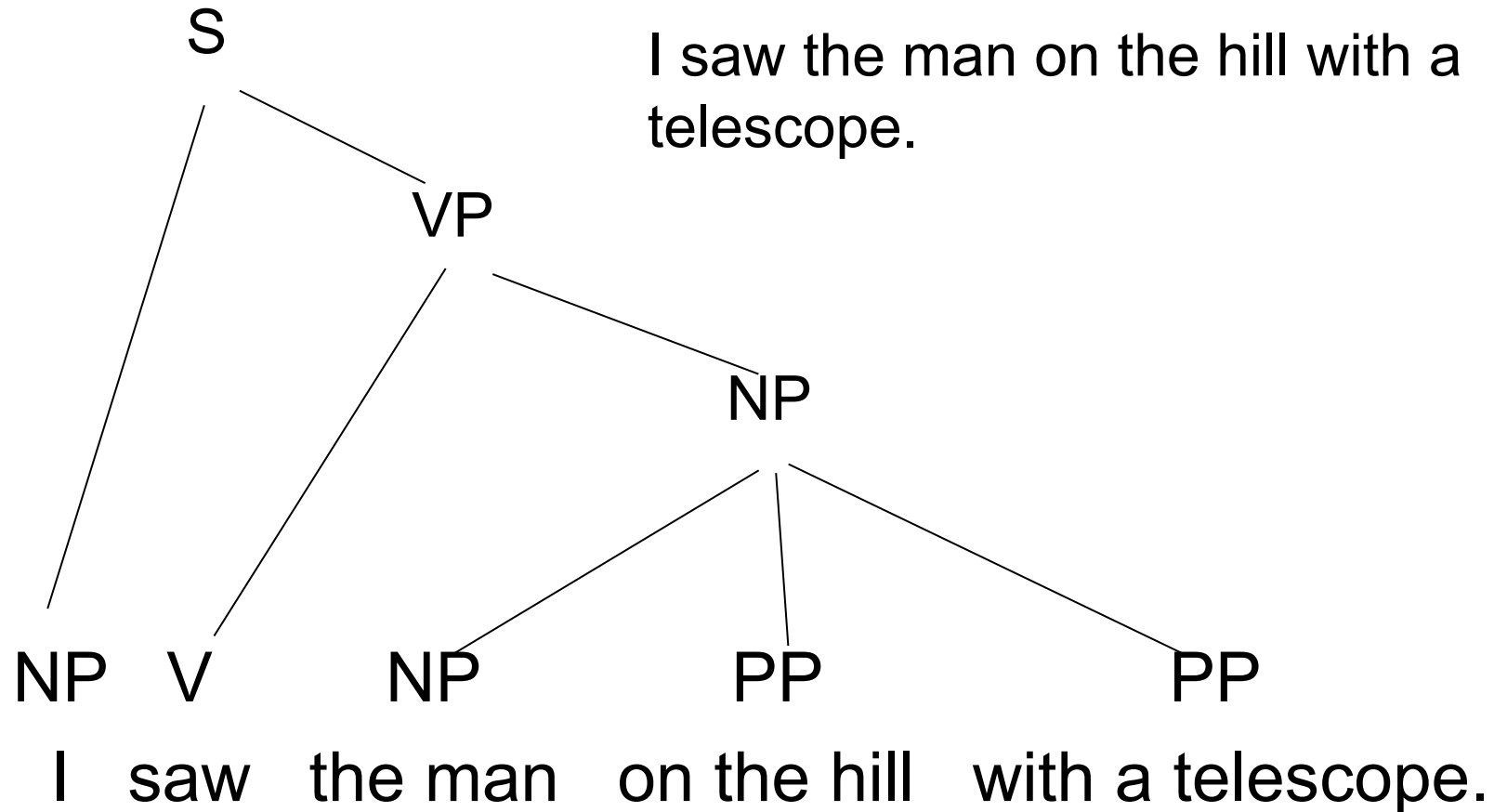
Time flies // like an arrow.
 NNS VBP

Syntax: structural ambiguity (part of speech)

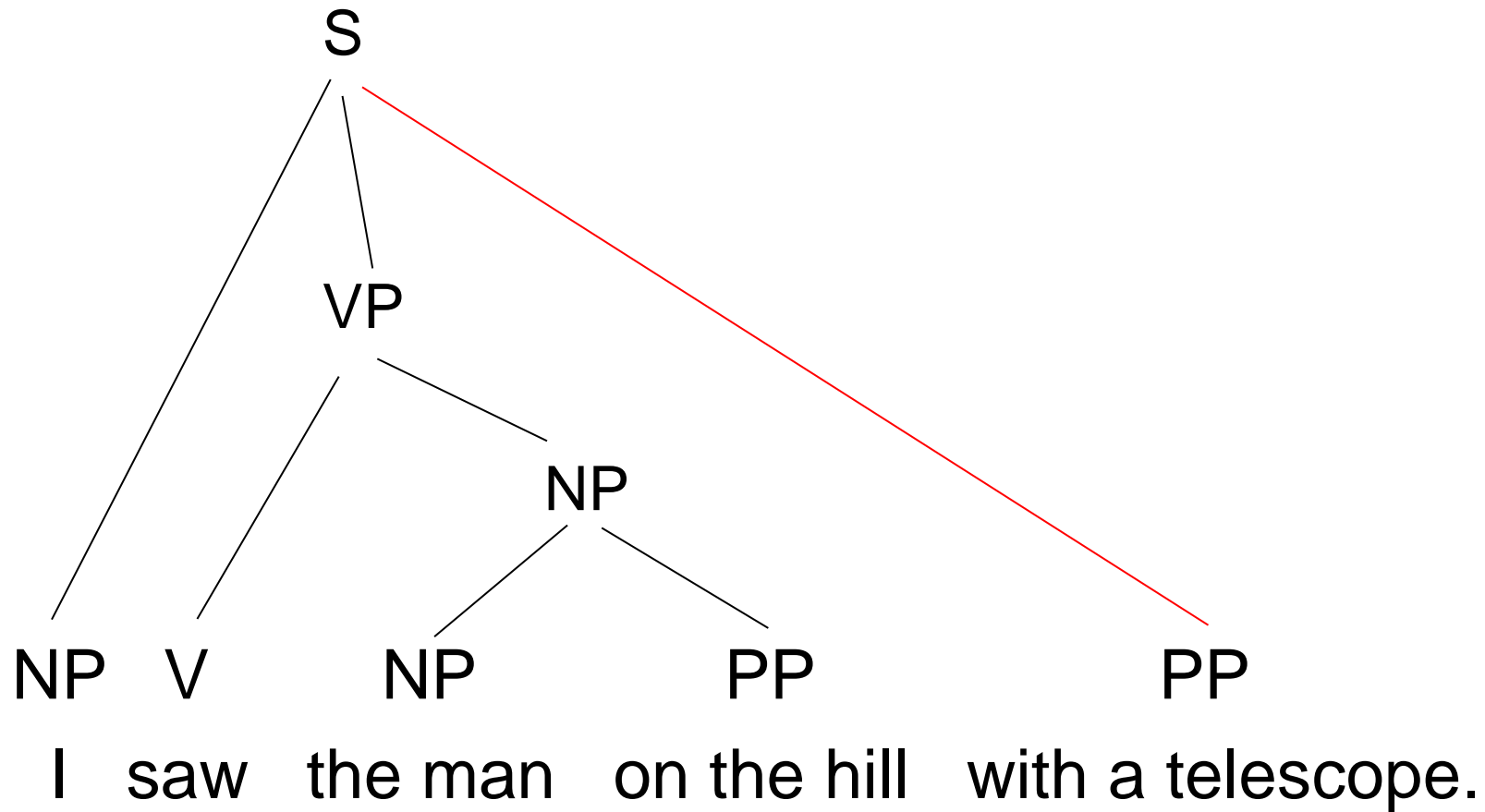
Ông già // đi nhanh quá.

Ông // già đi nhanh quá.

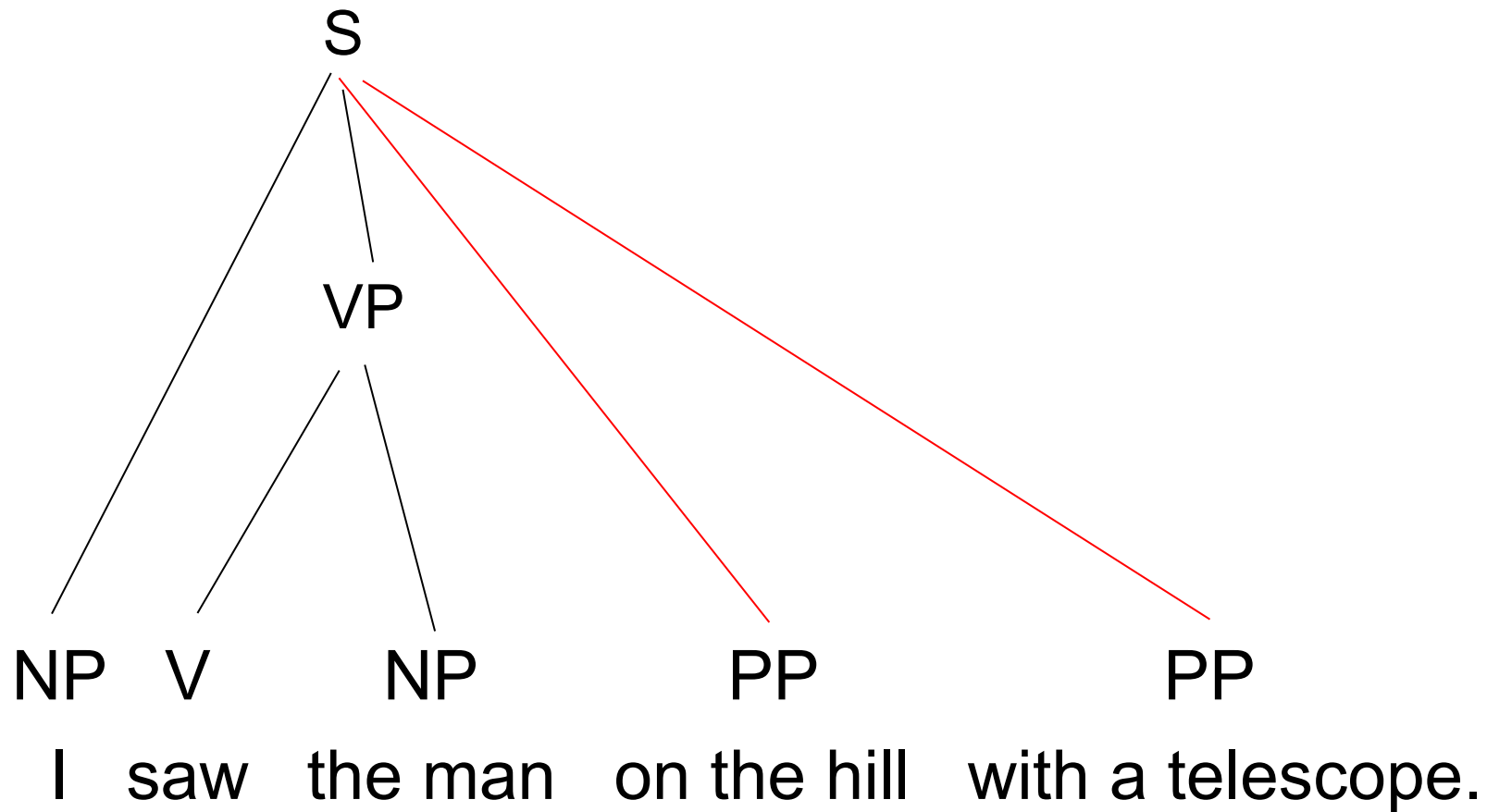
Syntax: structural ambiguity (attachment)



Syntax: structural ambiguity (attachment)



Syntax: structural ambiguity (attachment)



But syntax doesn't tell us much about meaning

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

Semantics: lexical ambiguity

- 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.

Discourse: coreference

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.

How do you react from what I said?

Rules of Conversation

- Can you tell me what time it is?
- Could I please have the salt?

Speech Acts

I bet you \$50 that the Jazz will win.

Mai went to the diner. She ordered a steak. She left a tip and went home.

- What did Mai eat for dinner?
- Who brought Mai her food?
- Who cooked the steak?
- Did Mai pay her bill?

Knowledge about language: What do we know about this sentence?

- Words must be appeared at a specific order:
 - a. Chó kem ăn. b. Chó ăn kem
- Các bộ phận cấu thành câu:
 - chó = chủ ngữ (subject); ăn kem = vị ngữ (predicate)
- Who did what to whom?
 - chủ thể(chó), hành động(ăn), đối tượng(kem)

Hidden knowledge

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

LSAT / (former) GRE Analytic Section Questions

- Six sculptures – C, D, E, F, G, H – are to be exhibited in rooms 1, 2, and 3 of an art gallery.
 - Sculptures C and E may not be exhibited in the same room.
 - Sculptures D and G must be exhibited in the same room.
 - If sculptures E and F are exhibited in the same room, no other sculpture may be exhibited in that room.
 - At least one sculpture must be exhibited in each room, and no more than three sculptures may be exhibited in any room.
- If sculpture D is exhibited in room 3 and sculptures E and F are exhibited in room 1, which of the following may be true?
 - A. Sculpture C is exhibited in room 1
 - B. Sculpture H is exhibited in room 1
 - C. Sculpture G is exhibited in room 2
 - D. Sculptures C and H are exhibited in the same room
 - E. Sculptures G and F are exhibited in the same room

Reference Resolution

- Knowledge sources:
 - Domain knowledge
 - Discourse knowledge
 - World knowledge

U: Where is **A Bug's Life** playing in **Mountain View**?

S: A Bug's Life is playing at the **Summit theater**.

U: When is **it** playing **there**?

S: It's playing at 2pm, 5pm, and 8pm.

U: I'd like 1 **adult** and 2 **children** for **the first show**.

How much would **that** cost?

What is the character of this knowledge?

- Some of it must be memorized :
 - Singing → Sing+ing; Bringing → bring+ing
- *Duckling* → ?? *Duckl +ing*
- So, must know duckl is not a word
- But it can't all be memorized

Besides memory, what else do we need?

English plural:

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

➤ ***Need a rule system to generate/analyze such cases***

Characteristics of NLP

- Ambiguous at all levels
- Involve reasoning about the world

Solutions?

- What do we need?
 - Linguistic knowledge
 - World knowledge
 - Combining all types of knowledge
- Potential solution:
 - Probabilistic models constructing from text corpus:
 - $P(\text{"maison"} \rightarrow \text{"house"})$ high
 - $P(\text{"L'avocat general"} \rightarrow \text{"the general avocado"})$ low