

NLP

Introduction to Natural Language Processing

Introduction

Introduction

- **Instructor:**
 - Dragomir Radev (PhD, Computer Science, Columbia University)
- **Course duration:**
 - 20 hours: 12 weeks x 1.5–2 hours per week
- **Intended audience:**
 - Primarily undergraduate students in Computer Science and Linguistics, Informatics,
 - Possibly also Mathematics, Statistics, Management, and Engineering
- **Related courses**
 - More introductory than Collins on Coursera (2013)
 - More focused on linguistics and resources than Jurafsky and Manning on Coursera (2012)

What is Natural Language Processing

- Natural Language Processing (NLP) is the study of the computational treatment of natural (human) language.
- In other words, teaching computers how to understand (and generate) human language.

Quiz

- Where is this quote from?

Dave Bowman: Open the pod bay doors, HAL.

HAL: I'm sorry Dave. I'm afraid I can't do that.

Quiz answer

- “2001: A Space Odyssey”
 - 1968 film by Stanley Kubrick
 - based on a joint screenplay with Arthur C. Clarke.

Modern Applications

- Search engines (Google, Yahoo!, Bing, Baidu)
- Question answering (IBM's Watson)
- Natural language assistants (Apple's Siri)
- Translation systems (Google Translate)
- News digest (Yahoo!)
- Automatic earthquake reports (LA Times)

Notes

- Computers are confused by (human) language
 - Specific techniques are needed
 - NLP draws on research in Linguistics, Theoretical Computer Science, Mathematics, Statistics, Artificial Intelligence, Psychology, Databases, etc.
- Goals of this class
 - Understand that language processing is hard (and why)
 - Understand the key problems in NLP
 - Learn about the methods used to address these problems
 - Understand the limitations of these methods

Language and Communication

- Speaker
 - Intention (goals, shared knowledge and beliefs)
 - Generation (tactical)
 - Synthesis (text or speech)
- Listener
 - Perception
 - Interpretation (syntactic, semantic, pragmatic)
 - Incorporation (internalization, understanding)
- Both
 - Context (grounding)

Basic NLP Pipeline

- (U)nderstanding and (G)eneration



NLP

NLP

Introduction to NLP

Examples of Text

Understanding a News Story

Brazil crowds attend funeral of late candidate Campos

More than 100,000 people in Brazil have paid their last respects to the late presidential candidate, Eduardo Campos, who died in a plane crash on Wednesday.

They attended a funeral Mass and filled the streets of the city of Recife to follow the passage of his coffin.

Later this week, Mr Campos's Socialist Party is expected to appoint former Environment Minister Marina Silva as a replacement candidate.

Mr Campos's jet crashed in bad weather in Santos, near Sao Paulo.

Investigators are still trying to establish the exact causes of the crash, which killed six other people.

Mr Campos's private plane - a Cessna 560XL - was travelling from Rio de Janeiro to the sea-side resort of Guaruja, near the city of Santos.

President Dilma Rousseff, who's running for re-election in October, was among many prominent politicians who travelled to Recife for the funeral.

Understanding a News Story

Brazil crowds attend funeral of late candidate Campos

More than 100,000 people in Brazil have paid their last respects to the late presidential candidate, Eduardo Campos, who died in a plane crash on Wednesday.

They attended a funeral Mass and filled the streets of the city of Recife to follow the passage of his coffin. Later this week, Mr Campos's Socialist Party **is expected** to appoint former Environment Minister Marina Silva as a replacement candidate.

Mr Campos's jet crashed in bad weather in Santos, near Sao Paulo.

Investigators are still trying to establish the exact causes of **the crash**, which killed six other people.

Mr Campos's private plane - a Cessna 560XL - was travelling from Rio de Janeiro to the sea-side resort of Guaruja, near the city of Santos.

President Dilma Rousseff, who's running for re-election in October, was among many prominent politicians who travelled to Recife for the funeral.

Why did I highlight some of the phrases above?

Answer to the quiz

- Brazil crowds attend funeral of late candidate Campos
 - Current event
- Mr Campos's jet crashed in bad weather in Santos
 - Background event
- Mr Campos's Socialist Party is expected to appoint...
 - Speculation
- President Dilma Rousseff
 - Property
- They attended a funeral Mass
 - Pronominal reference to an entity in a previous sentence

Genres of Text

- Blogs, emails, press releases, chats, debates, etc.
- Each presents different challenges to NLP

Firefox - Do you think that schools should provide... - www.createdebate.com/debate/show/Do_you_think_that_schools_should_provide_more_help_to_developmentally_challenged

First time here? Show Helpful Hints

CreateDebate

New Debate Browse Petitions About

Show All Replies Share This Debate Sort By Time

MAGNISES Experience New York City with Magnises "NYC's most exclusive card" - NY Post Join Today!

Sign Up | Login Debates | People Search

Debate Info

1 vs they should 2 what they are doing is fine

Debate Score: 3 Arguments: 3 Total Votes: 3 More Stats

Argument Ratio

what they should (1) what they are doing is fine (2)

Debate Creator 11wolff (454)

What links here Recent changes Upload file Permanent link Page information Redirects to this page Cite this page Print/export Create a book Download as PDF Printable version Languages Afrikaans Asturianu Azərbaycanca ବାନ୍ଦାମାଳ ବାନ୍ଦାମାଳ ବାନ୍ଦାମାଳ ବାନ୍ଦାମାଳ ବାନ୍ଦାମାଳ ବାନ୍ଦାମାଳ ବାନ୍ଦାମାଳ ବାନ୍ଦାମାଳ

Alfredo1356 (2) FromWibin (1) Alextherend (1)

1 point 1 point 1 point

Public schools should teach behavioral skills and give the mentally challenged listening exercises. Every student deserves to be given education, and these kinds of students should be able to communicate clearly with others. To start off, this is what they need. Mentally challenged students have trouble forming words, so they will need to learn them more accurately as others and have trouble forming words. If you focus on their speech, they will learn what they have learned. I noticed that the students in my class have trouble forming words, so I could make this stuff up because it is so difficult for them to pronounce the words they are saying. There were assistance guides around the school and giving her full sentences on what she wanted to do. She could not understand them. She was having trouble with her speech, so I had to help her with her speech. Her teacher struggled, but she didn't know what they were feeling. The school needs to provide more support for the students who have trouble forming words. They need to have more support and the teacher needs language. This was what she needed because she could never understand. TRY to understand the teachers.

This debate grows in English, the teachers will need to give them some sort of daily lesson. Every student learns math, science, English, and social studies. How about provide the mentally challenged with those subjects. Do not try and give them English. That should be provided after they have listened to whatever audio mathematics lesson or English lesson. They should listen every day as a replacement for homework. The teacher needs to give them work. They will be accustomed to the language and now need to be taught to read and write. This time, though, it will not be hard. They will know most of the words and will learn to say them aloud as they have learned from the audio lessons.

Firefox - W:Tony Blair - Wikipedia, the free encyclopedia - en.wikipedia.org/w/index.php?title=Tony_Blair&oldid=50000000

Article Talk

WIKIPEDIA The Free Encyclopedia

Tony Blair

From Wikipedia, the free encyclopedia

Main page Contents Featured events Current events Recent changes Donate to Wikipedia Wikipedia Shop

Interaction Help About Wikipedia Community portal Recent changes Contact page Tools What links here Recent changes Upload file Special pages Permanent link Page information Redirects to this page Cite this page Print/export Create a book Download as PDF Printable version

In office 2 May 1997 – 27 June 2007 Monarch Elizabeth II Deputy John Prescott Preceded by John Major Succeeded by Gordon Brown Leader of the Opposition In office 21 July 1994 – 24 May 1997 Monarch Elizabeth II Prime Minister John Major Preceded by Margaret Beckett Succeeded by John Major Leader of the Labour Party

Anthony Charles Lynton Blair (born 6 May 1953)¹⁰ is a British Labour Party politician who served as the Prime Minister of the United Kingdom from 1997 to 2007. He was the Member of Parliament (MP) for Sedgefield from 1983 to 2007 and Leader of the Labour Party from 1994 to 2007. Blair led a landslide victory in the 1997 general election, winning 418 seats, the most the party has ever held. The party went on to win two more elections under his leadership, in 2001, in which it won another landslide victory, and in 2005, with a reduced majority.

Blair was elected Labour Party leader in the leadership election of July 1994, following the sudden death of his predecessor, John Smith. Under his leadership, the party used the phrase "New Labour" to distance it from previous Labour policies. Blair declared opposition to the traditional conception of socialism, and declared support for a new conception that he referred to as "social-tens", involving policies that recognised individuals as socially interdependent, and advocated social justice, cohesion, equal worth of each citizen, and equal opportunity.¹⁰ Critics of Blair denounced him for having the Labour Party abandon genuine socialism and accepting capitalism.¹⁰ Supporters, including the party's public opinion polster Philip Gould, stated that after four consecutive General Election defeats, Labour had to demonstrate that it had made a decisive break from its left-wing past, in order to win again.¹⁰

In May 1997, at 43 years old, Blair won a landslide General Election victory, the largest in Labour Party history, to become the youngest Prime Minister since Lord Liverpool in 1812. He attained early personal popularity, receiving a 95% public approval rating in September 1997, after the death of Diana, Princess of Wales.¹⁰ In the first years of the New Labour government, Blair's government introduced the National Minimum Wage Act, Human Rights Act and Freedom of Information Act, and carried out devolution, establishing the Scottish Parliament, the National Assembly for Wales, and the Northern Ireland Assembly, fulfilling four of the promises in its 1997 manifesto.

In Northern Ireland, Blair was involved in the 1998 Good Friday Agreement. From the start of the War on Terror in 2001, he strongly supported much of the foreign policy of US President George W. Bush, and ensured that British armed forces participated in the 2001 invasion of Afghanistan and, more controversially, the 2003 invasion of Iraq. Blair is the Labour Party's longest-serving Prime Minister, the only person to have led the Labour Party to more than two consecutive general election victories, and the only Labour Prime Minister to serve consecutive terms more than one of which was at least four years long.

Blair was succeeded as Leader of the Labour Party on 24 June 2007 and as Prime Minister on 27 June 2007 by Gordon Brown.¹⁰ On the day he resigned as Prime Minister, he was appointed the official Envoy of the Quartet on the Middle East. In May 2008, Blair launched his Tony Blair Foundation.¹⁰ This was followed in July 2009 by the launching of the Faith and Globalisation Initiative with Yale University in the US, Durham University in the UK and the National University of Singapore in Asia to deliver a postgraduate programme in partnership with the Foundation.¹⁰

The Right Honourable Tony Blair



Induction of Influenza-Specific Mucosal Immunity by an Attenuated Recombinant Sendai Virus

Thuc-vy L. Le¹, Elena Mironova², Dominique Garcin², Richard W. Compans^{1*}

¹ Department of Microbiology and Immunology and Emory Vaccine Center, Emory University School of Medicine, Atlanta, Georgia, United States of America,

² Department of Microbiology and Molecular Medicine, University of Geneva School of Medicine, Geneva, Switzerland

Recent advances in molecular genetics have permitted the development of novel virus-based vectors for the delivery of genes and expression of gene products [6,7,8]. These live vectors have the advantage of promoting robust immune responses due to their ability to replicate, and induce expression of genes at high efficiency. Sendai virus is a member of the *Paramyxoviridae* family, belongs in the genus respirovirus and shares 60–80% sequence homology to human parainfluenza virus type 1 (HPIV-1) [9,10]. The viral genome consists of a negative sense, non-segmented RNA. Although Sendai virus was originally isolated from humans during an outbreak of pneumonitis [11] subsequent human exposures to Sendai virus have not resulted in observed pathology [12]. The virus is commonly isolated from mouse colonies and Sendai virus infection in mice leads to bronchopneumonia, causing severe pathology and inflammation in the respiratory tract. The sequence homology and similarities in respiratory pathology have made Sendai virus a mouse model for HPIV-1. Immunization with Sendai virus promotes an immune response in non-human primates that is protective against HPIV-1 [13,14] and clinical trials are underway to determine the efficacy of this virus for protection against HPIV-1 in humans [15]. Sendai virus naturally infects the respiratory tract of mice and recombinant viruses have been reported to efficiently transduce luciferase, lac Z and green fluorescent protein (GFP) genes in the airways of mice or ferrets as well as primary human nasal epithelial cells [16]. These data support the hypothesis that intranasal (i.n.) immunization with a recombinant Sendai virus will mediate heterologous gene expression in mucosal tissues and induce antibodies that are specific to a recombinant protein. A major advantage of a recombinant Sendai virus based vaccine is the observation that recurrence of parainfluenza virus infections is common in humans [12,17] suggesting that anti-vector responses are limited, making repeated administration of such a vaccine possible.

We have previously described the recovery of a mutant virus GP42-SeV containing a replacement of nucleotides 1–42 of the 3' leader sequence encoding the virus genomic promoter (P^L), with the corresponding sequence of the 5' trailer sequence encoding the anti-genomic promoter (P^R) [6]. This mutant exhibits enhanced chronic infection and reduced virus-induced programmed cell death *in vitro* [18]. In the present study, we introduced the hemagglutinin (HA) gene from influenza virus A/Puerto Rico/8/1934 (PR/8) into an attenuated sequence of GP42-SeV. This vector, designated GP42-H1, was evaluated for induction of systemic and mucosal antibodies as well as effectiveness of protection from lethal challenge with influenza virus.

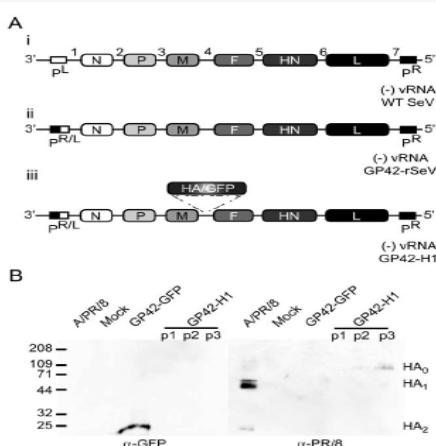


Figure 1. Generation of the GP42-H1 vector. A) Representation of wild-type Sendai virus genome (i). It illustrates the major viral genes N, P, M, F, HN and L. The left promoter (P^L) and right promoter (P^R) function as the genomic promoter and anti-genomic promoter respectively. Seven gene boundaries that encode the conserved regulatory transcription start and transcription stop sequence are represented numerically. The mutant GP42-SeV (ii) genomic RNA is identical to WT Sendai virus with the exception of 3' P^L in which 42 nucleotides of the P^L were replaced with the corresponding sequence from P^R . Additionally, a transcription start, stop poly-A signal sequences and a unique Mlu I restriction site was introduced into the intergenic region between the Sendai M and F genes. Using this unique Mlu I restriction site, the GFP or HA gene was inserted (respecting the rule of six) generating the recombinant Sendai GP42-GFP (19) or GP42-H1 vectors (iii). B) Recombinant GP42-H1 virus was cultured in BSR-T2 cells for three passages. Cell free supernatant containing virus suspensions were collected and used to infect CV-1 cells (refer to Materials and Methods). Proteins from GP42-GFP or GP42-H1 infected cell extracts were resolved on SDS-PAGE and screened for GFP (left) or HA (right) expression by western analysis. Mock infected cells and alantico fluid from PR/8 infected eggs are also shown.

Plos ONE

DOI: 10.1371/journal.pone.0018780

doi:10.1371/journal.pone.0018780.g001

Recent advances in molecular genetics have permitted the development of novel virus-based vectors for the delivery of genes and expression of gene products [6,7,8]. These live vectors have the advantage of promoting robust immune responses due to their ability to replicate, and induce expression of genes at high efficiency. **Sendai virus** is a member of the Paramyxoviridae family, belongs in the genus respirovirus and shares 60–80% sequence homology to **human parainfluenza virus type 1 (HPIV-1)** [9,10].

The viral genome consists of a negative sense, non-segmented RNA. Although **Sendai virus** was originally isolated from humans during an outbreak of pneumonitis [11] subsequent human exposures to **Sendai virus** have not resulted in observed pathology [12]. The virus is commonly isolated from mouse colonies and Sendai virus infection in mice leads to bronchopneumonia, causing severe pathology and inflammation in the respiratory tract. The sequence homology and similarities in respiratory pathology have made Sendai virus a mouse model for HPIV-1.

Immunization with Sendai virus promotes an immune response in non-human primates that is protective against **HPIV-1** [13,14] and clinical trials are underway to determine the efficacy of this virus for protection against HPIV-1 in humans [15]. Sendai virus naturally infects the respiratory tract of mice and **recombinant viruses** have been reported to efficiently transduce **luciferase, lac Z and green fluorescent protein (GFP) genes in the airways of mice or ferrets as well as primary human nasal epithelial cells** [16].

These data support the hypothesis that intranasal (i.n.) immunization with a recombinant Sendai virus will mediate heterologous gene expression in mucosal tissues and induce antibodies that are specific to a recombinant protein. A major advantage of a recombinant Sendai virus based vaccine is the observation that recurrence of **parainfluenza virus** infections is common in humans [12,17] suggesting that anti-vector responses are limited, making repeated administration of such a vaccine possible.

Named entities + variants (**human parainfluenza virus type, HPIV-1**)

Speculation (**reported, suggesting**)

Species (**human**)

Cell types (**nasal epithelial cells**)

Facts

References

Medical Records

TITLE: PC ACUTE CARE VISIT
DATE OF NOTE: FEB 04, 2000@11:18 ENTRY DATE: FEB 04, 2000@11:20
AUTHOR: EXP COSIGNER:
URGENCY: STATUS: COMPLETED

Chief Complaint: Patient notes 1 month history of blurred vision and frequent urination

HISTORY OF PRESENT ILLNESS:

DEMO,FATHER is a 44 year-old MALE who presents complaining of blurred vision for the past 1 month. He finds it is difficult for him to read clearly and is even effecting his driving. He also notes that he has been getting up to the bathroom frequently, esp. at night. He now routinely goes to urinate 3-4 times a night. He is not aware of any particular weight loss, but does feel thirsty much of the time.

PAST MEDICAL HISTORY:

Illnesses: Hypertension

Surgeries: None

Allergies: PENICILLINS

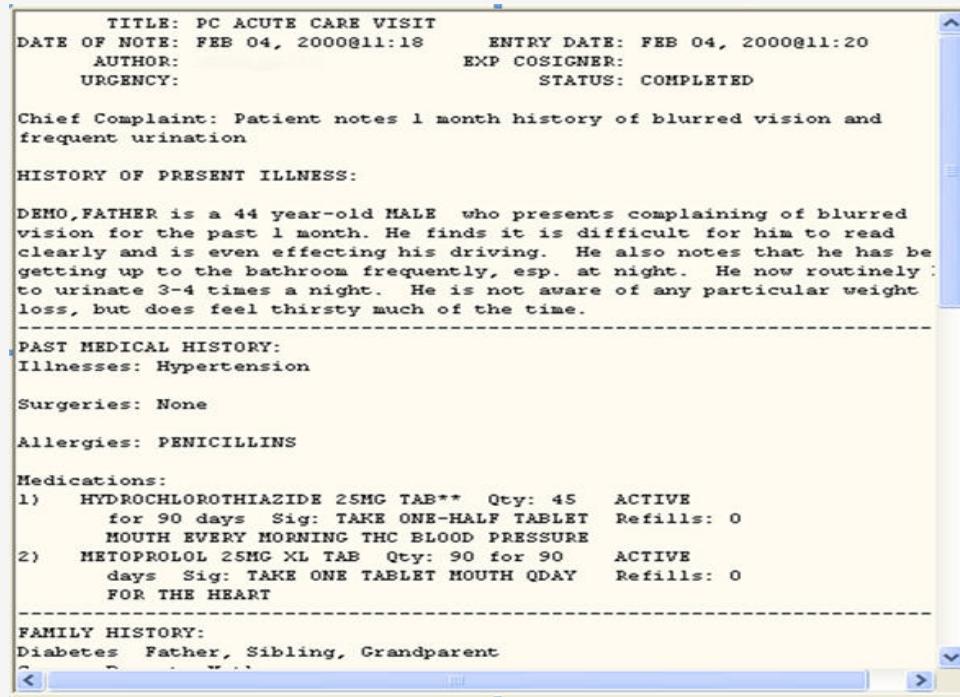
Medications:

1) HYDROCHLORTIAZIDE 25MG TAB** Qty: 45 ACTIVE
for 90 days Sig: TAKE ONE-HALF TABLET Refills: 0
MOUTH EVERY MORNING THC BLOOD PRESSURE

2) METOPROLOL 25MG XL TAB Qty: 90 for 90 ACTIVE
days Sig: TAKE ONE TABLET MOUTH QDAY Refills: 0
FOR THE HEART

FAMILY HISTORY:

Diabetes Father, Sibling, Grandparent



Medical Records

TITLE: PC ACUTE CARE VISIT
DATE OF NOTE: FEB 04, 2000@11:18 ENTRY DATE: FEB 04, 2000@11:20
AUTHOR: EXP COSIGNER:
URGENCY: STATUS: COMPLETED

Chief Complaint: Patient notes 1 month history of blurred vision and frequent urination

HISTORY OF PRESENT ILLNESS:

DEMO,FATHER is a 44 year-old MALE who presents complaining of blurred vision for the past 1 month. He finds it is difficult for him to read clearly and is even effecting his driving. He also notes that he has been getting up to the bathroom frequently, esp. at night. He now routinely goes to urinate 3-4 times a night. He is not aware of any particular weight loss, but does feel thirsty much of the time.

PAST MEDICAL HISTORY:

Illnesses: Hypertension

Surgeries: None

Allergies: PENICILLINS

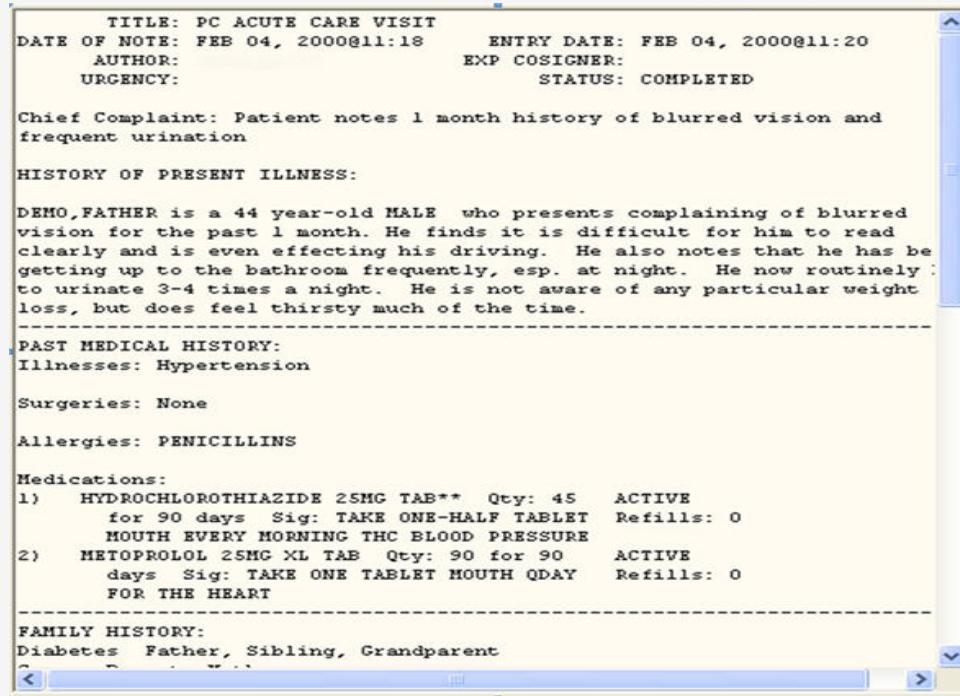
Medications:

1) HYDROCHLORTIAZIDE 25MG TAB** Qty: 45 ACTIVE
for 90 days Sig: TAKE ONE-HALF TABLET Refills: 0
MOUTH EVERY MORNING THC BLOOD PRESSURE

2) METOPROLOL 25MG XL TAB Qty: 90 for 90 ACTIVE
days Sig: TAKE ONE TABLET MOUTH QDAY Refills: 0
FOR THE HEART

FAMILY HISTORY:

Diabetes Father, Sibling, Grandparent



Literary Texts

- Project Gutenberg (<http://www.gutenberg.org/browse/scores/top>)
- A team of horses passed from Finglas with toiling plodding tread, dragging through the funereal silence a creaking waggon on which lay a granite block. The waggoner marching at their head saluted.
 - Ulysses – <http://www.gutenberg.org/files/4300/4300-h/4300-h.htm>
- There was no possibility of taking a walk that day. We had been wandering, indeed, in the leafless shrubbery an hour in the morning; but since dinner (Mrs. Reed, when there was no company, dined early) the cold winter wind had brought with it clouds so sombre, and a rain so penetrating, that further out-door exercise was now out of the question.
 - Jane Eyre – <http://www.gutenberg.org/files/1260/1260-h/1260-h.htm>
- Dorothy lived in the midst of the great Kansas prairies, with Uncle Henry, who was a farmer, and Aunt Em, who was the farmer's wife. Their house was small, for the lumber to build it had to be carried by wagon many miles. There were four walls, a floor and a roof, which made one room; and this room contained a rusty looking cookstove, a cupboard for the dishes, a table, three or four chairs, and the beds. Uncle Henry and Aunt Em had a big bed in one corner, and Dorothy a little bed in another corner. There was no garret at all, and no cellar--except a small hole dug in the ground, called a cyclone cellar, where the family could go in case one of those great whirlwinds arose, mighty enough to crush any building in its path. It was reached by a trap door in the middle of the floor, from which a ladder led down into the small, dark hole.
 - The Wizard of Oz – <http://www.gutenberg.org/files/55/55-h/55-h.htm>

A Really Long Literary Sentence

- Try parsing this
 - “Bloat is one of the co-tenants of the place, a maisonette erected last century, not far from the Chelsea Embankment, by Corydon Throsp, an acquaintance of the Rossettis' who wore hair smocks and liked to cultivate pharmaceutical plants up on the roof (a tradition young Osbie Feel has lately revived), a few of them hardy enough to survive fogs and frosts, but most returning, as fragments of peculiar alkaloids, to rooftop earth, along with manure from a trio of prize Wessex Saddleback sows quartered there by Throsp's successor, and dead leaves off many decorative trees transplanted to the roof by later tenants, and the odd unstomachable meal thrown or vomited there by this or that sensitive epicurean—all got scumbled together, eventually, by the knives of the seasons, to an impasto, feet thick, of unbelievable black topsoil in which anything could grow, not the least being bananas.”
- Do you know the source?

Quiz Answer

- “Gravity’s Rainbow” (by Thomas Pynchon), known for its use of very arcane words and complicated sentence (and plot) structure.
- Another such work is “Finnegans Wake” by James Joyce.
- Poetry is even more difficult.

NLP

NLP

Introduction to NLP

Funny Sentences

Silly Sentences

- Children make delicious snacks
- Stolen painting found by tree
- I saw the Rockies flying to San Francisco
- Court to try shooting defendant
- Ban on nude dancing on Governor's desk
- Red tape holds up new bridges
- Government head seeks arms
- Blair wins on budget, more lies ahead
- Local high school dropouts cut in half
- Hospitals are sued by seven foot doctors
- Dead expected To rise
- Miners refuse to work after death
- Patient at death's door – doctors pull him through
- In America a woman has a baby every 15 minutes. How does she do that?

More Ambiguous Headlines

- Vancouver Police Shoot Man Holding Box Cutter – PPSC Forums
 - <http://www.theppsc.org/forums/showthread.php?t=2657>
- Police: Man Armed With Box Cutters Shot By Officers – CBS New York
 - <http://newyork.cbslocal.com/2011/08/10/police-man-armed-with-box-cutters-shot-by-officer-in-queens/>
- Police Shoot Man With Box Cutters in Astoria
 - <http://nyc.pointslocal.com/news/2011/08/10/nyc/366208/police-shoot-man-with-box-cutters-in-astoria>
- Police shoot man yielding box cutters – Astoria, NY
 - <http://www.astorians.com/community/index.php?topic=21457.0>
- Police Shoot Man With Box Cutters
 - <http://nyc.rawsignal.com/news/Story.aspx?mode=0&id=27330161&ald=5804622&d=110811>

More Real Headlines

- New worm protests Sri Lankan violence
 - CNN – 05/17/2001
- Cartoon protesters stone embassy
 - <http://www.cnn.com/2006/WORLD/asiapcf/02/06/cartoon.protests/index.html>
- Child killer sentenced to die
 - <http://edition.cnn.com/2006/LAW/03/15/taped.abduction/index.html>
- Probe drills dentist

More Ambiguous Headlines

- Bulgaria Doubles Money for Parties
 - http://www.novinite.com/view_news.php?id=68597
- Motorola To Hire 300 Android Developers
 - http://rss.slashdot.org/~r/Slashdot/slashdot/~3/ir7e66GuL_A/article.pl
- Massachusetts exhales as Bill passes, heads to Canada
 - <http://www.cnn.com/2009/US/weather/08/23/bill.east.coast/index.html>
- U.S. eyes return to the moon
 - <http://www.cnn.com/2003/TECH/space/12/04/us.moon/index.html>
- Flesh-eating bug survivor goes home
 - http://www.cnn.com/2012/08/22/health/georgia-flesh-eating-bacteria/index.html?hpt=hp_t1



Ambiguous Recommendations



Ambiguous Recommendations

Ambiguous Recommendations

Lexical ambiguity

For a chronically absent employee	A man like him is hard to find.
For a dishonest employee	He's an unbelievable worker.
For a lazy employee	You would indeed be fortunate to get this person to work for you.
For the office drunk	Every hour with him was a happy hour.

Structural ambiguity

For a chronically absent employee	It seemed her career was just taking off.
For a dishonest employee	Her true ability was deceiving.
For a stupid employee	I most enthusiastically recommend this candidate with no qualifications whatsoever.
For the office drunk	He generally found him loaded with work to do.

Scope ambiguity

For an employee who is not worth further consideration as a job candidate	All in all, I cannot say enough good things about this candidate or recommend him too highly.
For an employee who is so unproductive that the job is better left unfilled	I can assure you that no person would be better for the job.

Other

For a lazy employee	He could not care less about the number of hours he has to put in.
For an employee who is not worth further consideration as a job candidate	I would urge you to waste no time in making this candidate an offer of employment.
For a stupid employee	There is nothing you can teach a man like him.

From Beatrice Santorini's collection

NLP

NLP

Introduction to NLP

Administrative

Structure of the Course

- Four major parts:
 - Linguistic, mathematical, and computational background
 - Computational models of morphology, syntax, semantics, discourse, pragmatics
 - Core NLP technology: parsing, part of speech tagging, text generation, etc.
 - Applications: text classification, machine translation, information extraction, etc.
- Three major goals:
 - Learn the basic principles and theoretical issues underlying natural language processing
 - Learn techniques and tools used to develop practical, robust systems that can understand text and communicate with users in one or more languages
 - Gain insight into some open research problems in natural language

Available Books

- **Speech and Language Processing**
 - Daniel Jurafsky and James Martin
 - <http://www.cs.colorado.edu/~martin/slp.html>
- **Foundations of Statistical Natural Language Processing**
 - Chris Manning and Hinrich Schütze
 - <http://nlp.stanford.edu/fsnlp/>
- **Natural Language Understanding**
 - James Allen

Courses at Other Places

- Brick-and-Mortar
 - Johns Hopkins University (Jason Eisner)
 - Cornell University (Lillian Lee)
 - Stanford University (Chris Manning)
 - U. Maryland (Hal Daumé)
 - Berkeley (Dan Klein)
 - U. Texas (Ray Mooney)
- Coursera
 - Manning/Jurafsky (2012, survey)
 - Michael Collins (2013, more advanced)

The Association for Computational Linguistics (ACL)

The screenshot shows a Mozilla Firefox browser window displaying the ACL website (<http://www.acweb.org/>). The page features a navigation menu on the left, a main content area with several sections, and a poll on the right.

Navigation Menu:

- Home
- News
- Conferences
- ACL Member Portal
- Publications
- CL Journal
- TACL
- Resources
- Affiliations
- SIGs
- About the ACL
- Contact Us
- ACL Policies

Main Content Area:

New ACL journal – Transactions of the Association for Computational Linguistics (TACL)

Following the decision of the ACL Executive Committee, which was announced at the ACL Business Meeting this year in Portland, we are now happy to launch the TACL Journal - Transactions of the Association for Computational Linguistics.

TACL targets papers of quite similar scope to our current (long) conference papers. Papers published at TACL will also receive a presentation slot at the annual ACL conference. TACL is thus intended to become an additional submission venue for presentation at the conference, coupled with fast turnaround journal-style reviewing and journal publication. In the future, additional conferences may join as presentation venues for TACL papers. While coupling with a conference presentation is a major feature of TACL, authors may choose to avoid such presentation.

The first two co-editors of TACL are Michael Collins and Dekang Lin, who will lead the journal setup and operation. The journal is expected to open for submissions in spring 2012, with papers accepted by a pre-set deadline receiving a presentation slot at ACL-2013. Additional details will be provided in the near future on the journal website and in a call for papers.

You are all welcome to plan your submissions to TACL!

Ido Dagan
Past President of the ACL
On behalf of the ACL Executive Committee

ACL Fellows Program

Greetings,

I am happy to announce that the ACL Executive Committee has approved the creation of an ACL Fellows program, which recognizes ACL members whose contributions to the field have been most extraordinary. To be named a Fellow, a candidate must have been a member of the ACL for the past three consecutive years and be nominated by a current ACL member.

Seventeen ACL members are among the founding group of Fellows. Each of these 2011 Fellows has been a major force in computational linguistics, and each has been a member of ACL for the last three years. Please see the list at http://acweb.org/acwiki/index.php?title=ACL_Fellows and congratulate them!

A small group of new Fellows will be announced each year. If you would like to nominate a candidate, please make your nomination at <http://www.acweb.org/portal/nominations>. If you are unsure about a candidate's eligibility, please send a query to ac@acweb.org. A nominator must provide a comprehensive case for the candidate, and so-called "letter of support". The nominator should direct the recommenders to fill out recommendation forms (ACL does not contact individual recommenders or solicit letters). All forms submitted by October 1 of a given year will be considered, and submitted forms will be kept confidential.

Regards,

POLLS
What do you think of the new site?

Love it.
 Hate it.
 This is a new site?

Monday, 09 January 2012

Find: children Next Previous Highlight all Match case Phrase not found Done

The Alphabet Soup

- NLP (Natural Language Processing)
- CL (Computational Linguistics)
- IR (Information Retrieval)
- SP (Speech Processing)
- HLT (Human Language Technology)
- NLE (Natural Language Engineering)
- ML (Machine Learning)

Research in NLP

- Conferences:
 - ACL/NAACL, EMNLP, SIGIR, AAAI/IJCAI, Coling, HLT, EACL/NAACL, AMTA/MT Summit, ICSLP/Eurospeech
- Journals:
 - Computational Linguistics, TACL, Natural Language Engineering, Information Retrieval, Information Processing and Management, ACM Transactions on Information Systems, ACM TALIP, ACM TSLP
- University centers:
 - Berkeley, Columbia, Stanford, CMU, JHU, Brown, UMass, MIT, UPenn, USC/ISI, Illinois, Michigan, UW, Maryland, etc.
 - Toronto, Edinburgh, Cambridge, Sheffield, Saarland, Trento, Prague, QCRI, NUS, and many others
- Industrial research sites:
 - Google, MSR, Yahoo!, IBM, SRI, BBN, MITRE, AT&T Labs
- The ACL Anthology
 - <http://www.aclweb.org/anthology>
- The ACL Anthology Network (AAN)
 - <http://clair.eecs.umich.edu/aan/index.php>

NLP

NLP

Introduction to NLP

Why is NLP hard?

Example

Time flies like an arrow.

- How many different interpretations does the above sentence have?
- How many of them are reasonable/grammatical?

Quiz Answer

- The most obvious meaning is
 - time flies very fast; as fast as an arrow.
- This is a metaphorical interpretation.
 - Computers are not really good at metaphors.
- Other interpretations:
 - Flies like honey -> flies like an arrow -> fruit flies like an arrow
 - Take a stopwatch and time the race -> time the flies

More Classic Examples

- Beverly Hills
- Beverly Sills
- The box is in the pen
- The pen is in the box
- Mary and Sue are mothers
- Mary and Sue are sisters
- Every American has a mother
- Every American has a president
- We gave the monkeys the bananas because they were hungry
- We gave the monkeys the bananas because they were over-ripe
 - <http://specgram.com/CLIII.4/08.phlogiston.cartoon.zhe.html>

Syntax vs. Semantics

* *Little a has Mary lamb.*

? *Colorless green ideas sleep furiously.*

[Chomsky 1957]

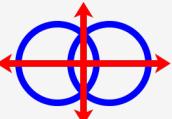
Ambiguous Words

- ball, board, plant
 - meaning
- fly, rent, tape
 - part of speech
- address, resent, entrance, number, unionized
 - pronunciation – give it a try

Answer To The Quiz

- **address**
 - The stress can be on either syllable. Compare with transport, effect, outline
- **resent**
 - As a verb infinitive or as “re-sent” a letter
- **entrance**
 - As a noun or as a verb meaning to put someone in a trance
- **number**
 - As a noun but also as the comparative of the adjective “numb”

Ambiguity

- Not in computer languages (by design)!
- Or Lojban 
- Noun-noun phrases: $(XY)Z$ vs. $X(YZ)$
 - *science fiction writer*
 - *customer service representative*
 - *state chess tournament*

NACLO Problems

- One Two Tree, by Noah Smith, Kevin Gimpel, and Jason Eisner
 - <http://www.naclo.cs.cmu.edu/problems2012/N2012-R.pdf>
- Fakepapershelfmaker, by Willie Costello
 - <http://www.naclo.cs.cmu.edu/problems2008/N2008-F.pdf>

NACLO Problem Solutions

- One Two Tree
 - <http://www.naclo.cs.cmu.edu/problems2012/N2012-RS.pdf>
- Fakepapershelfmaker
 - <http://www.naclo.cs.cmu.edu/problems2008/N2008-FS.pdf>

Types of Ambiguity 1/2

- **Morphological:**
 - Joe is quite impossible. Joe is quite important.
- **Phonetic:**
 - Joe's finger got number.
- **Part of speech:**
 - Joe won the first round.
- **Syntactic:**
 - Call Joe a taxi.
- **Pp attachment:**
 - Joe ate pizza with a fork / with meatballs / with Samantha / with pleasure.
- **Sense:**
 - Joe took the bar exam.
- **Modality:**
 - Joe may win the lottery.

Types of Ambiguity 2/2

- **Subjectivity:**
 - Joe believes that stocks will rise.
- **Cc attachment:**
 - Joe likes ripe apples and pears.
- **Negation:**
 - Joe likes his pizza with no cheese and tomatoes.
- **Referential:**
 - Joe yelled at Mike. He had broken the bike.
 - Joe yelled at Mike. He was angry at him.
- **Reflexive:**
 - John bought him a present.
 - John bought himself a present.
- **Ellipsis and parallelism:**
 - Joe gave Mike a beer and Jeremy a glass of wine.
- **Metonymy:**
 - Boston called and left a message for Joe.

Other Sources of Difficulties

- Non-standard, slang, and novel words and usages
 - A360, 7342.67, +1-646-555-2223
 - “spam” or “friend” as verbs
 - yolo, selfie, chillax – recently recognized as dictionary words
 - www.urbandictionary.com – (Parental Warning!)
- Inconsistencies
 - junior college, college junior
 - pet spray, pet llama
- Typoses and grammatical errors ☺
 - Receipt, John Hopkins, should of
- Parsing problems
 - Cup holder
 - Federal Reserve Board Chairman

Other Sources of Difficulties

- Complex sentences
- Counterfactual sentences
- Humor and sarcasm
- Implicature/inference/world knowledge:
 - I was late because my car broke down.
 - Implies I have a car, I use the car to get to places, the car has wheels, etc.
 - What is not explicitly mentioned, what is world knowledge?
- Semantics vs. pragmatics
 - Do you know the time?
- Language is hard even for humans (both L1 and L2)

Synonyms and Paraphrases

The S&P 500 climbed 6.93, or 0.56 percent, to 1,243.72, its best close since June 12, 2001.

The Nasdaq gained 12.22, or 0.56 percent, to 2,198.44 for its best showing since June 8, 2001.

The DJIA rose 68.46, or 0.64 percent, to 10,705.55, its highest level since March 15.

Synonyms and Paraphrases

The S&P 500 climbed 6.93, or 0.56 percent, to 1,243.72,

The Nasdaq gained 12.22, or 0.56 percent, to 2,198.44

The DJIA rose 68.46, or 0.64 percent, to 10,705.55,

since June 12, 2001.

its best close since June 8, 2001.

for its best showing since March 15.

NLP

NLP

Introduction to NLP

Background

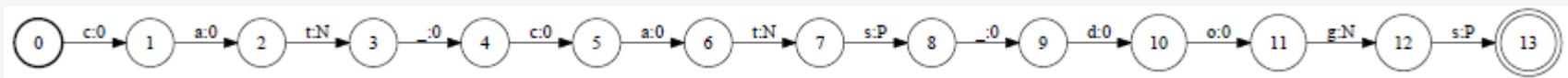
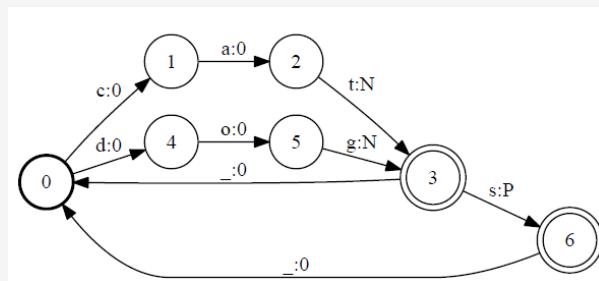
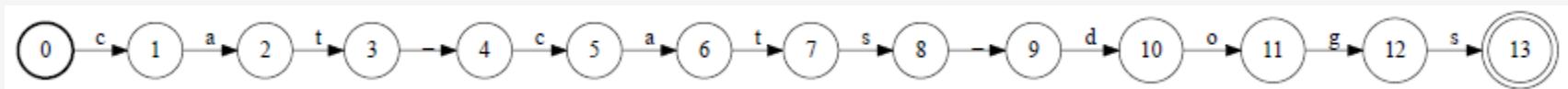
Linguistic Knowledge

- Constituents:
 - Children eat pizza.
 - They eat pizza.
 - My cousin's neighbor's children eat pizza.
 - Eat pizza!
- Collocations:
 - Strong beer but *powerful beer
 - Big sister but *large sister
 - Stocks rise but ?stocks ascend
 - in the past: 225,000 hits vs. 47 hits on Google, now 550,000 vs 57,000
- How to get this knowledge in the system:
 - Manual rules
 - Automatically acquired from large text collections (corpora)

Linguistic Knowledge

- Knowledge about language:
 - Phonetics and phonology – the study of sounds
 - Morphology – the study of word components
 - Syntax – the study of sentence and phrase structure
 - Lexical semantics – the study of the meanings of words
 - Compositional semantics – how to combine words
 - Pragmatics – how to accomplish goals
 - Discourse conventions – how to deal with units larger than utterances
- Separate lecture

Finite-state Automata



Theoretical Computer Science

- Automata
 - Deterministic and non-deterministic finite-state automata
 - Push-down automata
- Grammars
 - Regular grammars
 - Context-free grammars
 - Context-sensitive grammars
- Complexity
- Algorithms
 - Dynamic programming

Mathematics and Statistics

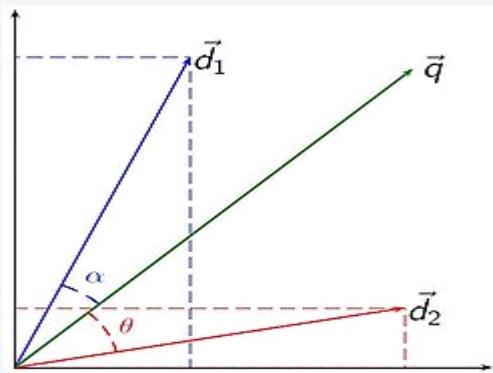
- Probabilities
- Statistical models
- Hypothesis testing
- Linear algebra
- Optimization
- Numerical methods

Mathematical and Computational Tools

- Language models
- Estimation methods
- Context-free grammars (CFG)
 - for trees
- Hidden Markov Models (HMM)
 - for sequences
- Conditional Random Fields (CRF)
- Generative/discriminative models
- Maximum entropy models

Statistical Techniques

- Vector space representation for WSD
- Noisy channel models for MT
- Graph-based Random walk methods for sentiment analysis



$$\begin{aligned}\hat{E} &= \operatorname{argmax}_{E \in English} P(E | F) \\ &= \operatorname{argmax}_{E \in English} \frac{P(F | E)P(E)}{P(F)} \\ &= \operatorname{argmax}_{E \in English} P(F | E)P(E)\end{aligned}$$

Artificial Intelligence

- Logic
 - First-order logic
 - Predicate calculus
- Agents
 - Speech acts
- Planning
- Constraint satisfaction
- Machine learning

NLP

NLP

Introduction to NLP

Linguistics

IPA Chart (consonants)

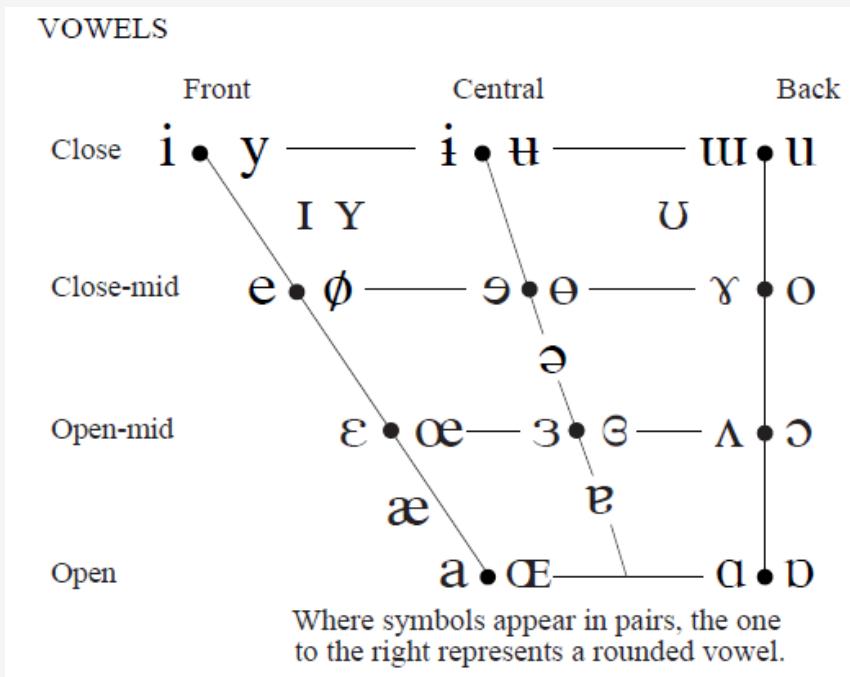
CONSONANTS (PULMONIC)

© 2005 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		t̪ d̪	c ɟ	k ɡ	q ɢ		?
Nasal	m	n̪j		n		n̪	jn̪	ŋ	N		
Trill	B			r					R		
Tap or Flap		v̪		f		t̪					
Fricative	ɸ β	f v̪	θ ð	s z	ʃ ʒ	s̪ z̪	ç ɟ	x ɣ	χ ʁ	h ʕ	h̪ ɦ
Lateral fricative				ɬ ɭ							
Approximant		v̪		ɹ		ɺ	ɟ	m̪			
Lateral approximant				l̪		ɻ	ʎ	L			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

IPA Chart (vowels)

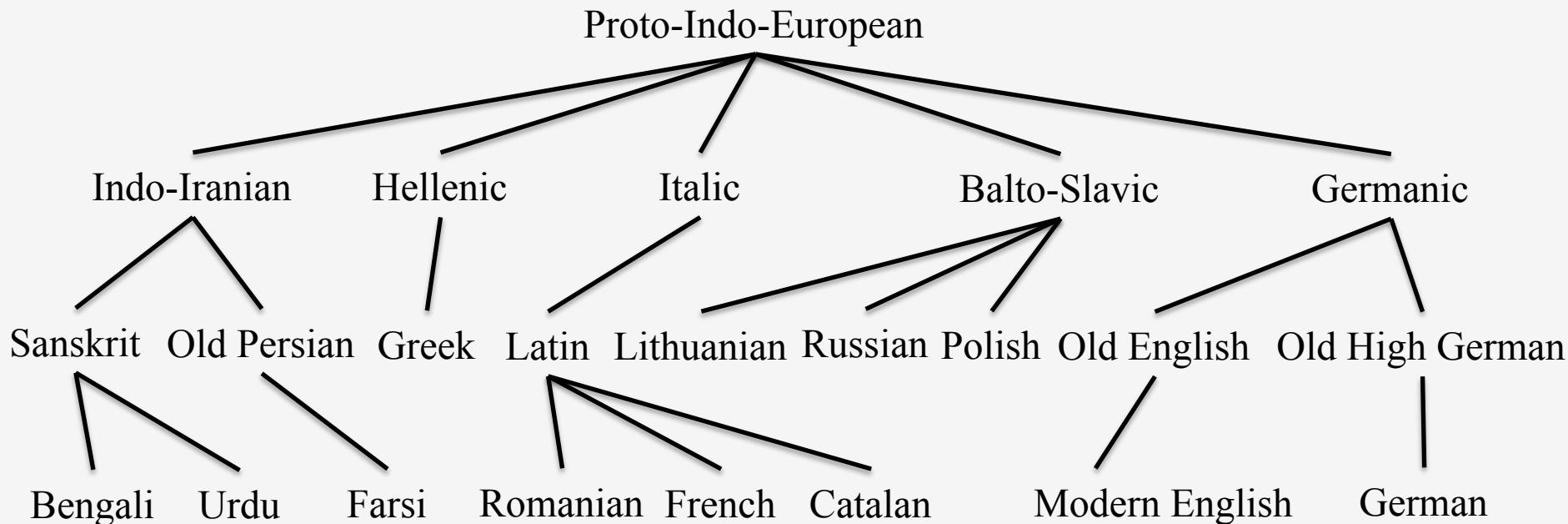


(Many) Languages are Related

- Cognates
 - night (English), nuit (French), Nacht (German), nacht (Dutch), nag (Afrikaans), nicht (Scots), natt (Swedish, Norwegian), nat (Danish), nátt (Faroese), nótt (Icelandic), noc (Czech, Slovak, Polish), ночь, noch (Russian), ноќ, ноќ (Macedonian), ноќ, nosht (Bulgarian), ніч, nich (Ukrainian), ноќ, noch/noć (Belarusian), noč (Slovene), noć (Serbo-Croatian), νύξ, nyx (Ancient Greek, νύχτα/nychta in Modern Greek), nox/nocte (Latin), nakt- (Sanskrit), natë (Albanian), noche (Spanish), nos (Welsh), nueche (Asturian), noite (Portuguese and Galician), notte (Italian), nit (Catalan), nuèch/nuèit (Occitan), noapte (Romanian), nakts (Latvian) and naktis (Lithuanian), all meaning "night" and derived from the Proto-Indo-European (PIE) *nókʷts, "night".

From wikipedia

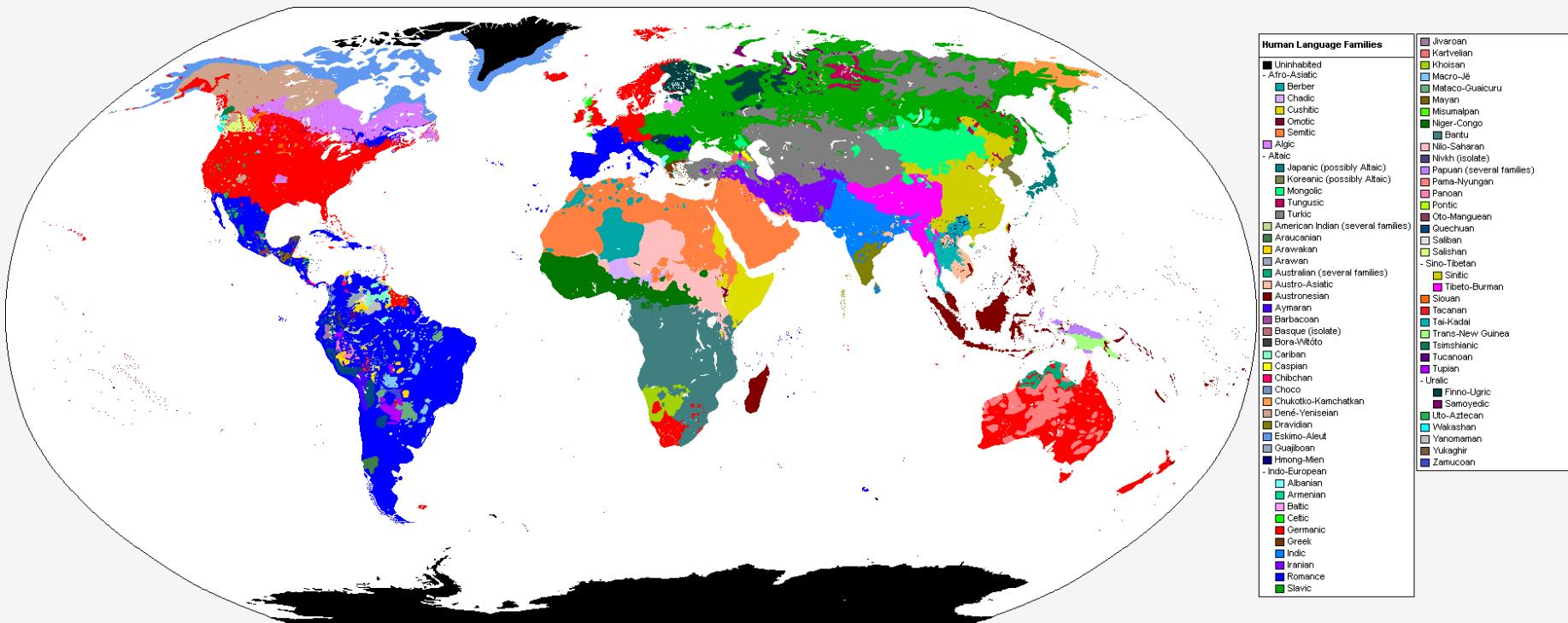
Some Indo-European languages



Some non-Indo-European Languages

- Altaic
 - Turkish
- Uralic (Finno-Ugric)
 - Finnish
 - Hungarian
- Semitic
 - Arabic
 - Hebrew
- Uto-Aztecán

Language Families



By Industrius at English Wikipedia. Later version(s) were uploaded by Mtll at English Wikipedia. (Image:BlankMap-World.png by User:Vardion) [GFDL (www.gnu.org/copyleft/fdl.html)], via Wikimedia Commons

Language Diversity

[Afro-Asiatic](#) (374)
[Alacalufan](#) (2)
[Algic](#) (44)
[Altaic](#) (66)
[Amto-Musan](#) (2)
[Andamanese](#) (13)
[Arafundi](#) (3)
[Arai-Kwomtari](#) (10)
[Arauan](#) (5)
[Araucanian](#) (2)
[Arawakan](#) (59)
[Arutani-Sape](#) (2)
[Australian](#) (264)
[Austro-Asiatic](#) (169)
[Austronesian](#) (1257)
[Aymaran](#) (3)
[Barbacoan](#) (7)
[Basque](#) (1)
[Bayono-Awbono](#) (2)
[Border](#) (15)
[Caddoan](#) (5)
[Cahuapanan](#) (2)

[Carib](#) (31)
[Central Solomons](#) (4)
[Chapacura-Wanham](#) (5)
[Chibchan](#) (21)
[Chimakuan](#) (1)
[Choco](#) (12)
[Chon](#) (2)
[Chukotko-Kamchatkan](#) (5)
[Chumash](#) (7)
[Coahuiltecan](#) (1)
[Constructed language](#) (1)
[Creole](#) (82)
[Deaf sign language](#) (130)
[Dravidian](#) (85)
[East Bird's Head-Sentani](#) (8)
[East Geelvink Bay](#) (11)
[East New Britain](#) (7)
[Eastern Trans-Fly](#) (4)
[Eskimo-Aleut](#) (11)
[Guahiban](#) (5)
[Gulf](#) (4)

[Harakmbet](#) (2)
[Hibito-Cholon](#) (2)
[Hmong-Mien](#) (38)
[Hokan](#) (23)
[Huavean](#) (4)
[Indo-European](#) (439)
[Iroquoian](#) (9)
[Japonic](#) (12)
[Jivaroan](#) (4)
[Kartvelian](#) (5)
[Katukinan](#) (3)
[Kaure](#) (4)
[Keres](#) (2)
[Khoisan](#) (27)
[Kiowa Tanoan](#) (6)
[Lakes Plain](#) (20)
[Language isolate](#) (50)
[Left May](#) (2)
[Lower Mamberamo](#) (2)
[Lule-Vilela](#) (1)
[Macro-Ge](#) (32)
[Mairasi](#) (3)

[Maku](#) (6)
[Mascoian](#) (5)
[Mataco-Guaicuru](#) (12)
[Mayan](#) (69)
[Maybrat](#) (2)
[Misumalpan](#) (4)
[Mixed language](#) (23)
[Mixe-Zoque](#) (17)
[Mongol-Langam](#) (3)
[Mura](#) (1)
[Muskocean](#) (6)
[Na-Dene](#) (46)
[Nambiquaran](#) (7)
[Niger-Congo](#) (1532)
[Nilo-Saharan](#) (205)
[Nimboran](#) (5)
[North Bougainville](#) (4)
[North Brazil](#) (1)
[North Caucasian](#) (34)
[Oto-Manguean](#) (177)
[Panoan](#) (28)

[Pauwasi](#) (5)
[Peba-Yaguan](#) (2)
[Penutian](#) (33)
[Piawi](#) (2)
[Pidgin](#) (17)
[Quechuan](#) (46)
[Ramu-Lower Sepik](#) (32)
[Salishan](#) (26)
[Salivan](#) (3)
[Senagi](#) (2)
[Sepik](#) (56)
[Sino-Tibetan](#) (449)
[Siouan](#) (17)
[Sko](#) (7)
[Somahai](#) (2)
[South Bougainville](#) (9)
[South-Central Papuan](#) (22)
[Tacanan](#) (6)
[Tai-Kadai](#) (92)
[Tarascan](#) (2)
[Tequistlatecan](#) (2)
[Tor-Kwerba](#) (24)

[Torricelli](#) (56)
[Totonacan](#) (12)
[Trans-New Guinea](#) (477)
[Tucanoan](#) (25)
[Tupi](#) (76)
[Unclassified](#) (73)
[Uralic](#) (37)
[Uru-Chipaya](#) (2)
[Uto-Aztecian](#) (61)
[Wakashan](#) (5)
[West Papuan](#) (23)
[Witotoan](#) (6)
[Yanomam](#) (4)
[Yele-West New Britain](#) (3)
[Yeniseian](#) (2)
[Yuat](#) (6)
[Yukaghir](#) (2)
[Yuki](#) (2)
[Zamucoan](#) (2)
[Zaparoan](#) (7)

Language Changes

- Grimm's Law
 - Voiceless stops turn into voiceless fricatives
 - Voiced stops become voiceless stops
 - Voiced aspirated stops change to voiced stops or fricatives
- Example 1
 - Ancient Greek: πούς, Latin: *pēs*, Sanskrit: *pāda*
 - English: *foot*, German: *Fuß*, Swedish: *fot*
- Example 2
 - Ancient Greek: κύων, Latin: *canis*, Welsh: *ci*
 - English: *hound*, Dutch: *hond*, German: *Hund*

NACLO Problem

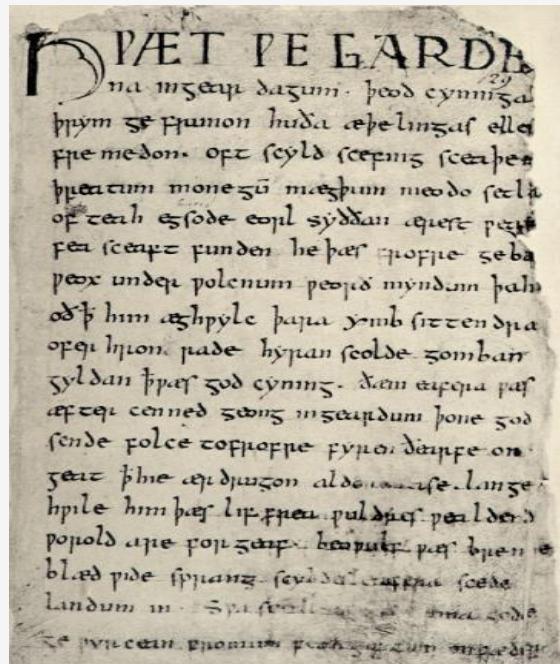
- All in the Family
 - <http://www.naclo.cs.cmu.edu/problems2012/N2012-D.pdf>

NACLO Solution

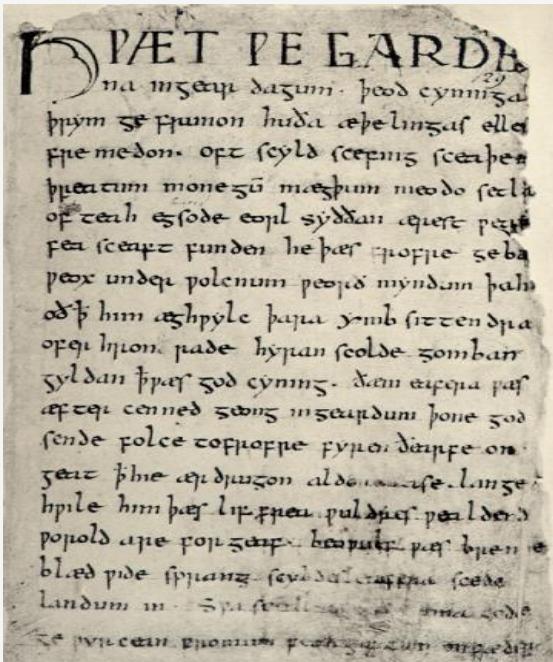
- All in the Family
 - <http://www.naclo.cs.cmu.edu/problems2012/N2012-DS.pdf>

Question

- Can you guess the source, language, and period of this text?



Answer



- Beowulf
- Epic poem
- 8th-11th Century
- Old English

Beowulf

Hwæt! We Gardena in geardagum,
þeodcyniga, þrym gefrunon,
hu ða æþelingas ellen fremedon.
Oft Scyld Scefing sceapena þreatum,
monegum mægþum, meodosetla
ofteah,
egsode eorlas. Syððan **aerest** wearð
feasceaft funden, he þæs frofre gebad,
weox under wolcnum, weorðmyndum
bah,
oðþæt him æghwylc þara ymbsittendra

erst (as in *erstwhile*) = first

Lo! the Spear-Danes' glory through splendid
achievements
The folk-kings' former fame we have heard of,
How princes displayed then their prowess-in-battle.
Oft Scyld the Scefing from scathers in numbers
From many a people their mead-benches tore.
Since **first** he found him friendless and wretched,
The earl had had terror: comfort he got for it,
Waxed 'neath the welkin, world-honor gained,
Till all his neighbors o'er sea were compelled to ...

<http://lit.genius.com/> <http://www8.georgetown.edu/departments/medieval/labyrinth/library/oe/texts/a4.1.html>

<http://www.gutenberg.org/files/16328/16328-h/16328-h.htm>

<http://www.nvcc.edu/home/vpoulakis/Translation/beowulf1.htm>

<http://en.wikipedia.org/wiki/File:Beowulf.firstpage.jpeg>

ATHELING.—Prince, nobleman.
BAIRN.—Son, child.
BARROW.—Mound, rounded hill, funeral-mound.
BATTLE-SARK.—Armor.
BEAKER.—Cup, drinking-vessel.
BEGEAR.—Prepare.
BIGHT.—Bay, sea.
BILL.—Sword.
BOSS.—Ornamental projection.
BRACATEATE.—A round ornament on a necklace.
BRAND.—Sword.
BURN.—Stream.
BURNIE.—Armor.
CARLE.—Man, hero.
EARL.—Nobleman, any brave man.
EKE.—Also.
EMPRISE.—Enterprise, undertaking.
ERST.—Formerly.
ERST-WORTHY.—Worthy for a long time past.
FAIN.—Glad.
FERRY.—Bear, carry.
FEY.—Fated, doomed.
FLOAT.—Vessel, ship.
FOIN.—To lunge (Shaks.).
GLORY OF KINGS.—God.
GREWSOME.—Cruel, fierce.
HEFT.—Handle, hilt; used by synecdoche for ‘sword.’
HELM.—Helmet, protector.
HENCHMAN.—Retainer, vassal.
HIGHT.—Am (was) named.
HOLM.—Ocean, curved surface of the sea.
HIMSEEMED.—(It) seemed to him.

LIEF.—Dear, valued.
MERE.—Sea; in compounds, ‘mere-ways,’ ‘mere-currents,’ etc.
MICKLE.—Much.
NATHLESS.—Nevertheless.
NAZE.—Edge (nose).
NESS.—Edge.
NICKER.—Sea-beast.
QUIT, QUITE.—Requite.
RATHE.—Quickly.
REAVE.—Bereave, deprive.
SAIL-ROAD.—Sea.
SETTLE.—Seat, bench.
SKINKER.—One who pours.
SOOTHLY.—Truly.
SWINGE.—Stroke, blow.
TARGE, TARGET.—Shield.
THROUGHLY.—Thoroughly.
TOLD.—Counted.
UNCANNY.—Ill-featured, grizzly.
UNNETHE.—Difficult.
WAR-SPEED.—Success in war.
WEB.—Tapestry (that which is ‘woven’).
WEDED.—Clad (cf. widow’s weeds).
WEEN.—Suppose, imagine.
WEIRD.—Fate, Providence.
WHILOM.—At times, formerly, often.
WIELDER.—Ruler. Often used of God;
WIGHT.—Creature.
WOLD.—Plane, extended surface.
WOT.—Knows.
YOUNKER.—Youth.

Diversity of Languages

- Articles
- Cases (e.g., in Latin)
 - Puer puellam vexat
- Sound systems
 - Glottal stop (the middle sound in “uh-oh”) – pro
 - Velar fricatives – articulated with the back of the tongue at the soft palate
 - Voiceless /χ/ – used e.g., in Russian
 - Voiced /ɣ/ – used e.g., in Modern Greek
- Social status (e.g., in Japanese)
 - otousan, お父さん = someone else’s father
 - chichi, 父 = one’s own father
- Kinship systems (e.g., in Warlpiri) – see next slide

NACLO Problem

- Warlpiri Kinship – by Alan Chang
 - <http://www.naclo.cs.cmu.edu/pdf-split/N2013-O.pdf>

NACLO Solution

- Warlpiri Kinship
 - <http://www.naclo.cs.cmu.edu/pdf-split/N2013-OS.pdf>

Language Universals

- Two types
 - unconditional
 - conditional
- Examples
 - All languages have verbs and nouns
 - All spoken languages have consonants and vowels
 - [Greenberg 1] “In declarative sentences with nominal subject and object, the dominant order is almost always one in which the subject precedes the object.”
 - [Greenberg 29] “If a language has inflection, it always has derivation.”

WALS: the World Atlas of Language Structures

- <http://wals.info>
- Feature 83A: Order of Object and Verb
 - by Matthew S. Dryer
 - OV (713 languages), VO (705), no dominant order (101)
 - <http://wals.info/feature/83A#2/18.0/152.9>
- Other features:
 - 18A Absence of common consonants (by Ian Maddieson):
no bilabials (5 languages), no fricatives (49), no nasals (12)
 - 67A Inflectional future tense (by Östen Dahl, Viveka Velupillai):
yes (110), no (112)

Links about World Languages

- Ethnologue
 - <http://www.ethnologue.com/>
- Number words in many languages
 - <http://www.zompist.com/numbers.shtml>
- Endangered languages
 - <http://www.endangeredlanguages.com/>
- Google fights to save 3,054 dying languages
 - <http://www.cnn.com/2012/06/21/tech/web/google-fights-save-language-mashable/index.html>

NLP

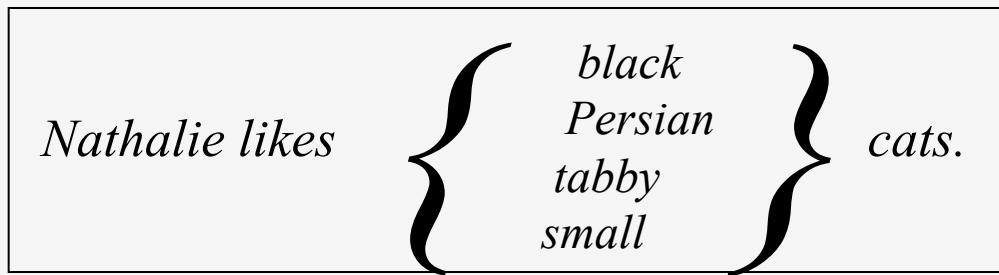
NLP

Introduction to NLP

Parts of speech

Syntactic Categories

- Substitution test:



- Open (lexical) and closed (functional) categories:

No-fly-zone
twerk

the
in

Example

The dog chased the yellow bird.

- Parts of speech
 - eight (or so) general types
 - nouns, verbs, adjectives...

Nouns

- Examples
 - *dog, tree, computer, idea*
- Nouns vary in
 - number (singular, plural)
 - gender (masculine, feminine, neuter)
 - case (nominative, genitive, accusative, dative)
- Case example in Latin
 - *Singular: puer (nominative), puerum (accusative), puerī (genitive)*
 - *Plural: puerī (nominative), puerōs (accusative), puerōrum (genitive)*
- Gender example in German
 - *Mädchen* (neuter gender)

Jabberwocky (Lewis Carroll)

'Twas **brillig**, and the **slithy toves**
Did gyre and gimble in the **wabe**:
All **mimsy** were the borogoves,
And the **mome raths** **outgrabe**.

- What are the parts of speech for the words in bold?

Answers

'Twas **brillig**, and the **slithy toves**
Did gyre and gimble in the **wabe**:
All **mimsy** were the borogoves,
And the **mome raths outgrabe**.

- **Wabe, borogoves**
 - Nouns (after “the”)
- **brillig**
 - adjective?
 - noun? (“noon”)
- **mimsy**
 - adjective
- **slightly toves**
 - adjective+noun?
 - noun+verb? (“the bell tolls”)
- **mome raths outgrabe**
 - Adjective+noun+verb?
 - Noun+verb+adverb? (“birds fly outside”)

Why is this an Important Example?

- Computers see text that they don't really understand.
- They have to use some prior knowledge.
- They reason probabilistically.
- They use context.
- They can be wrong.

Pronouns

- Examples
 - *she, ourselves, mine*
- Pronouns vary in
 - person
 - gender
 - number
 - case (in English: nominative, accusative, possessive, 2nd possessive)
- Reflexive and anaphoric forms
 - *herself, each other*

Samantha gave her a haircut.

Samantha gave herself a haircut.

Determiners and Adjectives

- Determiners
 - Articles
 - *the, a*
 - Demonstratives
 - *this, that*
- Adjectives
 - describe properties
 - attributive and predicative adjectives
 - agreement
 - in gender, number
 - comparative and superlative forms
 - derivative and periphrastic
 - positive form

Verbs

- **Describe**
 - actions, activities, and states (*throw, walk, have*)
- **English**
 - four verb forms
- **tense**
 - present, past, future
- **other inflection**
 - number, person
- **gerunds and infinitive**
- **aspect**
 - progressive, perfective
- **Voice**
 - active, passive

Verbs

- **Participles, auxiliaries**
- **Arguments:**
 - The dog sleeps (intransitive)
 - The dog chased the cat (transitive)
 - Mary gave the dog a bone (ditransitive)
- **Irregular verbs**
- **Richer inflections**
 - E.g., French, Latin, and Finnish

Verb Conjugation in French

Present

je vais I go
tu vas you go
il va he goes
nous allons we go
vous allez you go
ils vont they go

Conditional

j'irais I would go
tu irais you would go
il irait he would go
nous irions we would go
vous iriez you would go
ils iraient they would go

Past

je suis allé(e) I went
tu es allé(e) you went
il est allé(e) he went
nous sommes allé(e)s we went
vous êtes allé(e)s you went
ils sont allé(e)s they went

Future

j'irai I will go
tu iras you will go
il ira he will go
nous irons we will go
vous irez you will go
ils iront they will go

Imperfect

j'allais I used to go
tu allais you used to go
il allait he used to go
nous allions we used to go
vous alliez you used to go
ils allaient they used to go

Subjunctive

que j'aille that I go
que tu ailles that you go
qu'il aille that he go
que nous allions that we go
que vous alliez that you go
qu'ils aillent that they go

Other Parts of Speech

- Adverbs
 - happily, here, never
- Prepositions
 - of, through, in
- Particles
 - Phrasal verbs
 - the plane took off, take it off
- Particles vs. prepositions
 - *She ran up a bill/hill*

Other Parts of Speech

- Coordinating conjunctions
 - *and, or, but*
- Subordinating conjunctions
 - *if, because, that, although*
- Interjections
 - *Ouch!*

Part of Speech Tags

```
NN    /* singular noun */
IN    /* preposition */
AT    /* article */
NP    /* proper noun */
JJ    /* adjective */
,     /* comma */
NNS   /* plural noun */
CC    /* conjunction */
RB    /* adverb */
VB    /* un-inflected verb */
VBN   /* verb +en (taken, looked (passive,perfect)) */
VBD   /* verb +ed (took, looked (past tense)) */
CS    /* subordinating conjunction */
```

NLP

NLP

Introduction to NLP

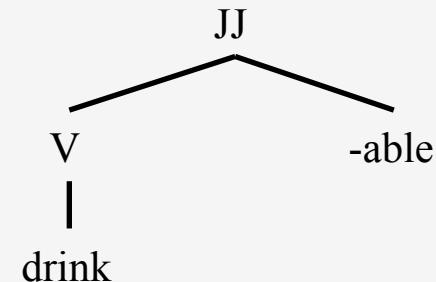
Morphology and the Lexicon

Mental Lexicon

- What is the meaning of cat? Its pronunciation? Part of speech?
- What is the meaning of wug?
- What is the meaning of cluvious?
- Compare trافتful and trافتless?
- Morphology of these words
- Intuition and productivity
- “Runs”
- Allomorphs – “cats/oxen”, “played/swung”
- Affixes

Derivational Morphology

- Er (many examples)
- What do these morphemes mean?
 - Ness, able, ing, re, un, er (adj)
 - JJ → V + “-able”
- Recursion:
 - unconcernednesses
- Ambiguous – undoable
- Not ambiguous – unbelievable – why?



Answer to the Quiz

- **Undoable**
 - unable to be done
 - able to be undone
- **Unbelievable**
 - unable to be believed
 - ? able to be unbelieved

Morphological Examples

- Reduplication
 - amigo = friend, amimígo = friends (in Pangasinan) [Rubino 2001]
 - savali = he travels, savavali = they travel (in Samoan)
- Circumfixes
 - spielen – gespielt (in German)
- Pig Latin
 - appyhay
- Verlan
 - “céfran”, “ripou” (from “l’envers”, “Français”, “pourri”)
- Massa-*freakin’-chusetts*
 - where can you insert “freakin” in “education”?

Answer to the Quiz

- The “freakin” infix is inserted
- ... to the left of the syllable that bears the main stress
 - edu-*freakin’-cation*
 - * educa-*freakin’-tion*
 - * e-*freakin’-ducation*
- though there can be exceptions

Morphemes

- Stems, affixes
- Concatenative morphology
- Templatistic morphology (e.g., Semitic languages):
 - lmd (learn), lamad (he studied), limed (he taught), lumad (he was taught)

Inflectional Morphology

- Tense, number, person, mood, aspect
- Five verb forms in English
- 40+ forms in French
- Six cases in Russian:
<http://www.departments.bucknell.edu/russian/language/case.html>
- Up to 40,000 forms in Turkish (you cause X to cause Y to ... do Z)

Morphological Analysis

- *sleeps* = *sleep* + V + 3P + SG
- *done* = *do* + V + PP

Turkish Vowel Harmony

	Front		Back	
	Unrounded	Rounded	Unrounded	Rounded
High	i	ü	ı	u
Low	e	ö	a	o

- **Back vowels**
 - in the room → oda~~da~~
 - at the door → kapı~~da~~
- **Front vowels**
 - at home → ev~~de~~
 - at the lake → göl~~de~~
 - on the bridge → köprü~~de~~

NACLO Problem

- Turkish
 - www.naclo.cs.cmu.edu/problems2010/F.pdf

NACLO Solution

- Turkish
 - www.naclo.cs.cmu.edu/problems2010/FS.pdf

Agglutinative Languages

- How does English become Turkish?

if we will be able to make ... become strong

if we will be able to make ... become strong

... strong become to make be able will if we

... sağlam +laş +tır +abil +ecek +se +k

↓
... sağlamlaştıra**bilece**ksek

アメフト	amefuto	Ame(rican) Foot(ball)
アイスクリーム	aisu kurīmu	ice cream
アイドル	aidoru	idol
アパート	apāto	apartment
バイク	baiku	bike
バリアフリー	bariafuri	barrier free
コンピューター	konpyūtā	computer
デスク	desuku	desk (at a news agency)
ラマ	dorama	drama (on TV)
エレベーター	erebētā	elevator
エスカレーター	esukarētā	escalator
フライドポテト	furaidopoteto	fried potato (French fries)
グラス	gurasu	glass (for drinking)
ハッピーエンド	happīendo	happy end(ing)
ホットケーキ	hottokēki	hotcake (pancake)
カシューなツツ	kashū nattsu	cashew nut
コーヒー	kōhī	coffee
クラブ	kurabu	club
キーボード	kībōdo	keyboard
キャンペーン	kyanpēn	campaign
キャップ	kyappu	cap
パソコン	pāsokon	perso(nal) com(puter)
パソコン用コンピューター	pāsonaru konpyūtā	personal computer
レジュメ	rejume	resume
レストラン	resutoran	restaurant
リモコン	rimokon	remo(te) con(trol)
サラダ	sarada	salad
タバコ	tabako	tobacco
テレビゲーム	terebigēmu	television game
ゼミナー	zemināru	seminar

Introduction to NLP

Other Levels of Linguistic Analysis

Semantics

- Semantics
 - Lexical semantics and compositional semantics
- Lexical Semantics
 - Hyponyms, hyponyms, antonyms, meronyms and holonyms (part-whole relationship, tire is a meronym of car), synonyms, homonyms
 - Senses of words, polysemous words
 - Collocations
 - *white hair, white wine*
 - Idioms
 - *to kick the bucket*
- Compositional Semantics
 - How to understand the meaning of a sentence based on the meaning of its components.

Pragmatics

- The study of how knowledge about the world and language conventions interact with literal meaning.
- Speech acts
- Resolution of anaphoric relations
- Modeling of speech acts in dialogue

Other Areas

- **Sociolinguistics**
 - interactions of social organization and language.
- **Historical linguistics**
 - change over time.
- **Linguistic typology**
- **Language acquisition**
 - L1 and L2
- **Psycholinguistics**
 - real-time production and perception of language

NLP

NLP

Text similarity

Introduction

Text Similarity

- People can express the same concept (or related concepts) in many different ways. For example, “the plane leaves at 12pm” vs “the flight departs at noon”
- Text similarity is a key component of Natural Language Processing
- If the user is looking for information about cats, we may want the NLP system to return documents that mention kittens even if the word “cat” is not in them.
- If the user is looking for information about “fruit dessert”, we want the NLP system to return documents about “peach tart” or “apple cobbler”.
- A speech recognition system should be able to tell the difference between similar sounding words like the “Dulles” and “Dallas” airports.
- This set of lectures will teach you how text similarity can be modeled computationally.

Human Judgments of Similarity

tiger	cat	7.35
tiger	tiger	10.00
book	paper	7.46
computer	keyboard	7.62
computer	internet	7.58
plane	car	5.77
train	car	6.31
telephone	communication	7.50
television	radio	6.77
media	radio	7.42
drug	abuse	6.85
bread	butter	6.19
cucumber	potato	5.92

[Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin,
"Placing Search in Context: The Concept Revisited", ACM Transactions on Information Systems, 20(1):116–131,
January 2002]

Human Judgments of Similarity

delightful	wonderful	A	8.65
modest	flexible	A	0.98
clarify	explain	V	8.33
remind	forget	V	0.87
get	remain	V	1.6
realize	discover	V	7.47
argue	persuade	V	6.23
pursue	persuade	V	3.17
plane	airport	N	3.65
uncle	aunt	N	5.5
horse	mare	N	8.33

[SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation. 2014. Felix Hill, Roi Reichart and Anna Korhonen. Preprint published on arXiv. arXiv:1408.3456]

Automatic Similarity Computation

spain	0.679
belgium	0.666
netherlands	0.652
italy	0.633
switzerland	0.622
luxembourg	0.610
portugal	0.577
russia	0.572
germany	0.563
catalonia	0.534

- Words most similar to “France”
- Computed using “word2vec”
- [Mikolov et al. 2013]

Types Of Text Similarity

- Many types of text similarity exist:
 - Morphological similarity (e.g., respect–respectful)
 - Spelling similarity (e.g., theater–theatre)
 - Synonymy (e.g., talkative–chatty)
 - Homophony (e.g., raise–raze–rays)
 - Semantic similarity (e.g., cat–tabby)
 - Sentence similarity (e.g., paraphrases)
 - Document similarity (e.g., two news stories on the same event)
 - Cross-lingual similarity (e.g., Japan–Nihon)

NLP

NLP

Text Similarity

*Morphological Similarity:
Stemming*

Morphological Similarity

- Words with the same root:
 - scan (base form)
 - scans, scanned, scanning (inflected forms)
 - scanner (derived forms, suffixes)
 - rescan (derived forms, prefixes)
 - rescanned (combinations)

Stemming

- To stem a word is to reduce it to a base form, called the *stem*, after removing various suffixes and endings and, sometimes, performing some additional transformations
- Examples
 - *scanned* → *scan*
 - *indication* → *indicate*
- In practice, prefixes are sometimes preserved, so *rescan* will not be stemmed to *scan*

Porter's Stemming Method

- Porter's stemming method is a rule-based algorithm introduced by Martin Porter in 1980
- The paper ("An algorithm for suffix stripping") has been cited more than 7,000 times according to Google Scholar
- The input is an individual word. The word is then transformed in a series of steps to its stem
- The method is not always accurate

Porter's Algorithm

- Example 1:
 - Input = *computational*
 - Output = *comput*
- Example 2:
 - Input = *computer*
 - Output = *comput*
- The two input words end up stemmed the same way

Porter's Algorithm

- The *measure* of a word is an indication of the number of syllables in it
 - Each sequence of consonants is denoted by C
 - Each sequence of vowels is denoted as V
 - The initial C and the final V are optional
 - So, each word is represented as [C]VCVC ... [V], or [C](VC){k}[V], where *k* is its *measure*

Examples of Measures

- $k=0$: I, AAA, CNN, TO, GLEE
- $k=1$: OR, EAST, BRICK, STREET, DOGMA
- $k=2$: OPAL, EASTERN, DOGMAS
- $k=3$: EASTERNMOST, DOGMATIC

Porter's Algorithm

- The initial word is then checked against a sequence of transformation patterns, in order.
- An example pattern is:
 - $(m > 0)$ ATION \rightarrow ATE
 - medicate \rightarrow medication
- Note that this pattern matches *medication* and *dedication*, but not *nation*.
- Whenever a pattern matches, the word is transformed and the algorithm restarts from the beginning of the list of patterns with the transformed word.
- If no pattern matches, the algorithm stops and outputs the most recently transformed version of the word.

Example Rules

- Step 1a

SSES → SS

presses → press

IES → I

lies → li

SS → SS

press → press

S → Ø

lots → lot

- Step 1b

(m>0) EED → EE

refereed → referee

(doesn't apply to bleed since m('BL')=0)

Example Rules

- Step 2

(m>0) ATIONAL	->	ATE	inflational	->	inflate
(m>0) TIONAL	->	TION	notional	->	notion
(m>0) IZER	->	IZE	nebulizer	->	nebulize
(m>0) ENTLI	->	ENT	intelligentli	->	intelligent
(m>0) OUSLI	->	OUS	analogousli	->	analogous
(m>0) IZATION	->	IZE	realization	->	realize
(m>0) ATION	->	ATE	predication	->	predicate
(m>0) ATOR	->	ATE	indicator	->	indicate
(m>0) IVENESS	->	IVE	attentiveness	->	attentive
(m>0) ALITI	->	AL	realiti	->	real
(m>0) BILITI	->	BLE	abiliti	->	able

Example Rules

- Step 3

(m>0) ICATE -> IC
(m>0) ATIVE -> ø
(m>0) ALIZE -> AL
(m>0) ICAL -> IC
(m>0) FUL -> ø
(m>0) NESS ->

replicate -> replic
informative -> inform
realize -> real
electrical -> electric
blissful -> bliss
tightness -> tight

- Step 4

(m>1) AL -> ø
(m>1) ANCE -> ø
(m>1) ER -> ø
(m>1) IC -> ø
(m>1) ABLE -> ø
(m>1) IBLE -> ø
(m>1) EMENT -> ø
(m>1) MENT -> ø
(m>1) ENT -> ø

appraisal -> apprais
conductance -> conduct
container -> contain
electric -> electr
countable -> count
irresistible -> irresist
displacement -> displac
investment -> invest
respondent -> respond

Examples

- Example 1:
 - Input = *computational*
 - Step 2: replace *ational* with *ate*: *compute*
 - Step 4: replace *ate* with \emptyset : *comput*
 - Output = *comput*
- Example 2:
 - Input = *computer*
 - Step 4: replace *er* with \emptyset : *comput*
 - Output = *comput*
- The two input words end up stemmed the same way

External Pointers

- Online demo
 - <http://text-processing.com/demo/stem/>
- Martin Porter's official site
 - <http://tartarus.org/martin/PorterStemmer/>

Quiz

- How will the Porter stemmer stem these words?

construction ?

increasing ?

unexplained ?

differentiable ?

- Check the Porter paper (or the code for the stemmer) in order to answer these questions.
- Is the output what you expected? If not, explain why.

Answers to the Quiz

construction ?

increasing ?

unexplained ?

differentiable ?

construction **construct**

increasing **increas**

unexplained **unexplai**

differentiable **differenti**

NACLO Problem

- Thorny Stems, NACLO 2008 problem by Eric Breck
 - [http://www.naclo.cs.cmu.edu/assets/problems/
NACLO08h.pdf](http://www.naclo.cs.cmu.edu/assets/problems/NACLO08h.pdf)

Solution to the NACLO Problem

- Thorny Stems
 - <http://www.naclo.cs.cmu.edu/problems2008/N2008-HS.pdf>

NLP

NLP

Text Similarity

*Spelling Similarity:
Edit Distance*

Spelling Similarity

- **Typos:**
 - Brittany Spears → Britney Spears
 - Catherine Hepburn → Katharine Hepburn
 - Reciept → receipt
- **Variants in spelling:**
 - Theater → theatre

Who Is This?

مُعْمَر القذافي

Hints

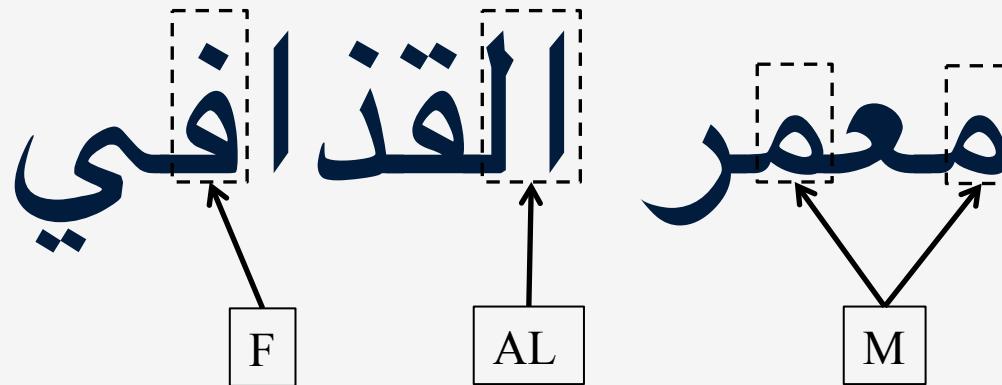
القذافي
معلم

M

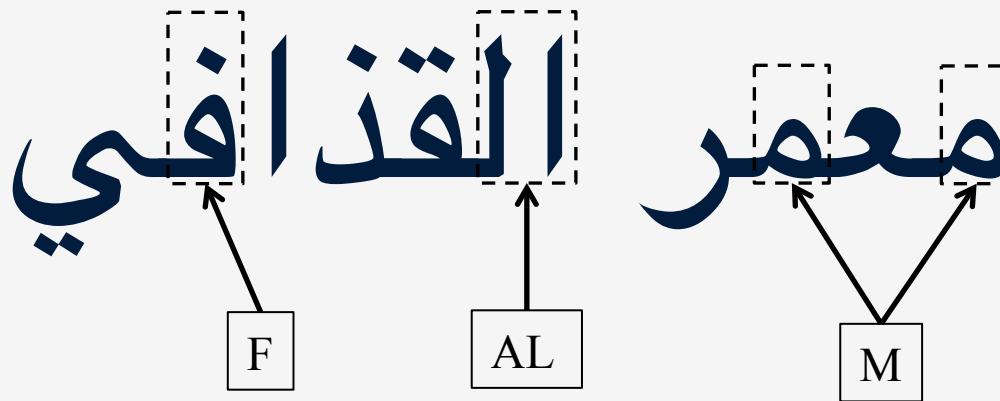
Hints



Hints



Hints



Muammar (al-)Gaddafi, or Moamar Khadafi, or ...

Quiz

How many different transliterations can there be?

m
u o
a
m mm
a e
r

el al El Al ø

Q G Gh K Kh
a e u
d dh ddh dh dh th
zz
a
f ff
i y

A Lot!

m
u o
a
m mm
a e
r

el al El Al ø

Q G Gh K Kh
a e u
d dh ddh dhdh th
zz
a
f ff
i y

8

x

5

x

360

=

14,400

Edit Operations

- behaviour – behavior (insertion/deletion) (“al”)
- string – spring (substitution) (“k”–“q”)
- sleep – slept (multiple edits)

Levenshtein Method

- Based on dynamic programming
- Insertions, deletions, and substitutions usually all have a cost of 1.

Example

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1								
r	2								
e	3								
n	4								
d	5								

Recurrence Relation

- Definitions
 - $s_1(i)$ – i^{th} character in string s_1
 - $s_2(j)$ – j^{th} character in string s_2
 - $D(i, j)$ – edit distance between a prefix of s_1 of length i and a prefix of s_2 of length j
 - $t(i, j)$ – cost of aligning the i^{th} character in string s_1 with the j^{th} character in string s_2
- Recursive dependencies
$$\begin{aligned} D(i, 0) &= i \\ D(0, j) &= j \\ D(i, j) &= \min [\\ &\quad D(i-1, j) + 1 \\ &\quad D(i, j-1) + 1 \\ &\quad D(i-1, j-1) + t(i, j) \\] \end{aligned}$$
- Simple edit distance:
$$\begin{aligned} t(i, j) &= 0 \text{ } iff \text{ } s_1(i) = s_2(j) \\ t(i, j) &= 1, \text{ otherwise} \end{aligned}$$

Example

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1	1							
r	2								
e	3								
n	4								
d	5								

Example

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1	1	1						
r	2								
e	3								
n	4								
d	5								

Example

Example

Example

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1	1	1	2	3	4	5	6	7
r	2	2	2	1	2	3	4	5	6
e	3	3	3	2	1	2	3	4	5
n	4	4	4	3	2	1	2	3	4
d	5	5	5	4	3	2	2	3	4

Edit Transcript

		s	t	r	e	n	g	t	h	
		0	1	2	3	4	5	6	7	8
t	1	1	1	2	3	4	5	6	7	
r	2	2	2	1	2	3	4	5	6	
e	3	3	3	2	1	2	3	4	5	
n	4	4	4	3	2	1	2	3	4	
d	5	5	5	4	3	2	2	3	4	

Other Costs

- Damerau modification
 - Swaps of two adjacent characters also have a cost of 1
 - E.g., $\text{Lev}(\text{"cats"}, \text{"cast"}) = 2$,
 $\text{Dam}(\text{"cats"}, \text{"cast"}) = 1$

Quiz

- Some distance functions can be more specialized.
- Why do you think that the edit distances for these pairs are as follows?
 - $\text{Dist}(\text{"sit clown"}, \text{"sit down"}) = 1$
 - $\text{Dist}(\text{"qeather"}, \text{"weather"}) = 1$, but $\text{Dist}(\text{"leather"}, \text{"weather"}) = 2$

Quiz Answers

- $\text{Dist}(\text{"sit down"}, \text{"sit clown"})$ is lower in this example because we want to model the type of errors common with optical character recognition (OCR)
- $\text{Dist}(\text{"qeather"}, \text{"weather"}) < \text{Dist}(\text{"leather"}, \text{"weather"})$ because we want to model spelling errors introduced by “fat fingers” (clicking on an adjacent key on the keyboard)



Quiz: Guess the Language

AACCTGCGGAAGGATCATTACCGAGTGC GGTCCTTGGGCCAACCTCCCATCCGTGTCTATTGTACCC
TGTTGCTTCGGCGGGCCCGCCGCTGTGC GGCCGCCGGGGGGCGCCTCTGCCCCCCGGGCCGTGCCCGC
CGGAGACCCCAACACGAACACTGTCTGAAAGCGTGCAGTCTGAGTTGATTGAATGCAATCAGTTAAA
ACTTTCAACAATGGATCTCTGGTTCCGGC

Quiz Answer

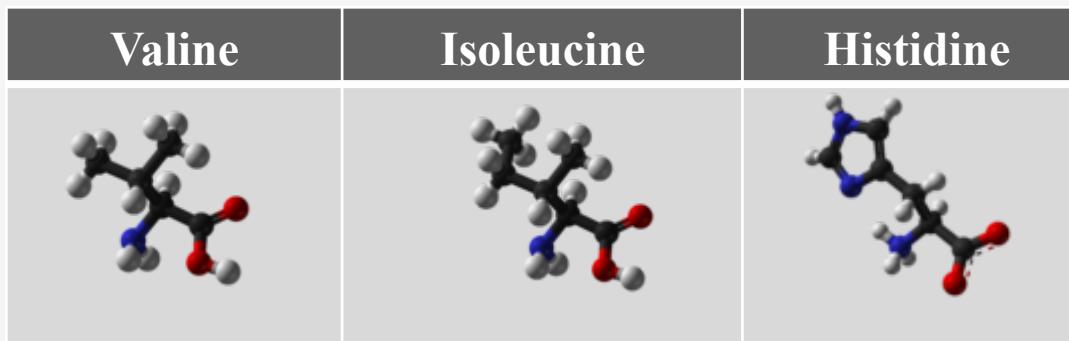
- This is a genetic sequence (nucleotides AGCT)

>U03518 **Aspergillus awamori internal transcribed spacer 1 (ITS1)**

AACCTGCGGAAGGATCATTACCGAGTGCAGGTCTTGCCGGCAACCTCCATCCGTGTCTATTGTACCC
TGTTGCTTCGGCGGGCCCGCTGTGGCCGCCGGGGGGCGCCTCTGCCCGGGCCGTGCCCGC
CGGAGACCCAACACGAACACTGTCTGAAAGCGTGCAGTCTGAGTTGATTGAATGCAATCAGTTAAA
TTCAACAATGGATCTCTGGTTCCGGC

Other Uses of Edit Distance

- In biology, similar methods are used for aligning non-textual sequences
 - Nucleotide sequences, e.g., GTTCGTGATGGAGCG, where A=adenine, C=cytosine, G=guanine, T=thymine, U=uracil, “-”=gap of any length, N=either one of ACGTU, etc.
 - Amino acid sequences, e.g., FMELSEDGIEMAGSTGVI, where A=alanine, C=cystine, D=aspartate, E=glutamate, F=phenylalanine, Q=glutamine, Z=either glutamate or glutamine, X=“any”, etc. The costs of alignment are determined empirically and reflect evolutionary divergence between protein sequences. For example, aligning V (valine) and I (isoleucine) is lower-cost than aligning V and H (histidine).



External URLs

- Levenshtein demo
 - <http://www.let.rug.nl/~kleiweg/lev/>
- Biological sequence alignment
 - http://www.bioinformatics.org/sms2/pairwise_align_dna.html
 - <http://www.sequence-alignment.com/sequence-alignment-software.html>
 - <http://www.ebi.ac.uk/Tools/msa/clustalw2/>
 - <http://www.animalgenome.org/bioinfo/resources/manuals/seqformats>

NACLO Problem

- “Nok-Nok”, NACLO 2009 problem by Eugene Fink:
 - <http://www.naclo.cs.cmu.edu/problems2009/N2009-B.pdf>

Solution to the NACLO Problem

- “Nok–Nok”
 - <http://www.naclo.cs.cmu.edu/problems2009/N2009-BS.pdf>

NACLO Problem

- “The Lost Tram”, NACLO 2007 problem by Boris Iomdin:
 - <http://www.naclo.cs.cmu.edu/problems2007/N2007-F.pdf>

Solution to the NACLO problem

- “The Lost Tram”
 - <http://www.naclo.cs.cmu.edu/problems2007/N2007-FS.pdf>

NLP

NLP

Introduction to NLP

NACLO

NACLO

- Competition in Linguistics (including Computational Linguistics)
 - Since 2007
 - <http://www.naclo.cs.cmu.edu>
- Best individual US performers so far:
 - Adam Hesterberg (2007)
 - Hanzhi Zhu (2008)
 - Rebecca Jacobs (2007–2009) – 3 team golds + 2 individual medals
 - Ben Sklaroff (2010)
 - Morris Alper (2011)
 - Alex Wade (2012, 2013) – 2 team golds + 2 individual golds + 1 individual silver
 - Darryl Wu (2012, 2014)
- Other strong countries:
 - Russia, UK, Netherlands, Poland, Bulgaria, South Korea, Canada, China
- IOL – the International contest
 - Since 2003
 - IOL 2013 in Manchester, IOL 2014 in Beijing, IOL 2015 in Bulgaria
 - <http://www.ioling.org>
- Other high school competitions, e.g., IMO, IOI, IPhO, IChO, IBO, IOAA, etc.

Consider these phrases in Ancient Greek (in a Roman-based transcription) and their unordered English translations:

- (A) *ho tōn hyiōn dulos*
- (B) *hoi tōn dulōn cyrioi*
- (C) *hoi tu emporu adelphoi*
- (D) *hoi tōn onōn emporoi*
- (E) *ho tu cyriu onos*
- (F) *ho tu oicu cyrios*
- (G) *ho tōn adelphōn oicos*
- (H) *hoi tōn cyriōn hyioi*

- (1) the donkey of the master
- (2) the brothers of the merchant
- (3) the merchants of the donkeys
- (4) the sons of the masters
- (5) the slave of the sons
- (6) the masters of the slaves
- (7) the house of the brothers
- (8) the master of the house

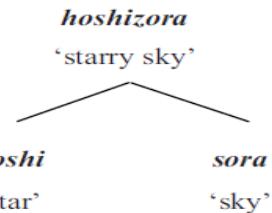
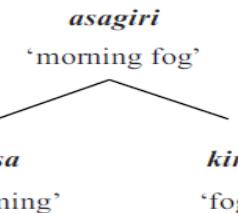
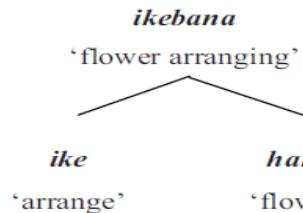
C1. Place the number of the correct English translation in the space following each Greek sentence. Explain your answers!

C2. Translate into Ancient Greek:

the houses of the merchants;
the donkeys of the slave

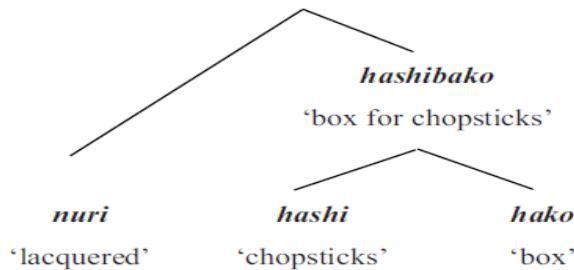
Explain your answers!

In English, we can combine two nouns to get a compound noun, such as in ‘mailbox’ or ‘sandcastle’. We can do this in Japanese as well, but just sticking the two words together isn’t enough. Instead, the words themselves undergo predictable changes:

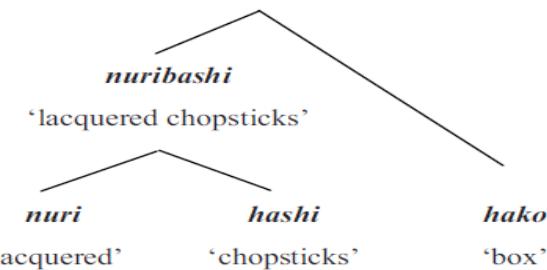


Compound words can then be compounded again, creating compounds with three or more members. Study the diagrams below carefully. You’ll notice that the order in which the compound is built affects both the meaning and the final form of the word.

nurihashibako
‘lacquered box for chopsticks’



nuribashibako
‘box for lacquered chopsticks’



An excerpt from a well known text is shown below. It is in two languages (X and Y) that are closely linguistically related to each other and also to English. However the two versions are not perfect translations of one another.

Text in language X

- X1. Rödluvan: Men mormor, varför har du så stora ögon?
- X2. "Mormor": Det är bara för att jag skall se dig bättre, mitt barn.
- X3. Rödluvan: Men mormor, varför har du så stora öron?
- X4. "Mormor": Det är bara för att jag skall höra dig bättre, mitt barn.
- X5. Rödluvan: Men mormor, varför har du så stora tänder?
- X6. "Mormor": Det är bara för att jag skall kunna äta upp dig!

(almost) the same text in language Y

- Y1. - Så store ører du har, bestemor, sa Rødhette.
- Y2. - Det er fordi jeg skal kunne høre deg bedre, svarte ulven.
- Y3. - Så store øyne du har, bestemor, sa Rødhette.
- Y4. - Det er fordi jeg skal kunne se deg bedre, svarte ulven.
- Y5. - Så store hender du har, bestemor, sa Rødhette.
- Y6. - Det er fordi jeg skal kunne klemme deg bedre, svarte ulven.
- Y7. - Så stor munn du har, bestemor, sa Rødhette.
- Y8. - Det er fordi jeg skal kunne ete deg bedre, svarte ulven.

To the right is a Japanese word written in the *tenji* ("dot characters") writing system. The large dots represent the raised bumps; the tiny dots represent empty positions.

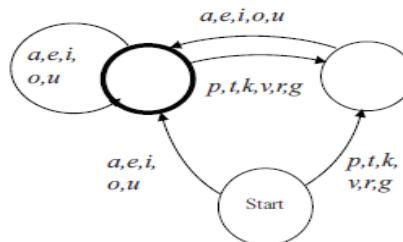
karaoke :: :: :: ::

A1. The following *tenji* words represent *atari*, *haiku*, *katana*, *kimono*, *koi*, and *sake*. Which is which? You don't need to know either Japanese or Braille to figure it out; you'll find that the system is highly logical.

a. _____		b. _____	
c. _____		d. _____	
e. _____		f. _____	

Finite-state automata (FSA) are a type of abstract “machine” with many possible uses. One possible use is to guess what language a document (such as a webpage) is in. If we make an automaton that can distinguish between possible English words and impossible ones, and then give it a webpage with a bunch of words that are impossible in English (like “*aioaepa*” or “*ragaiiare*”), we can be pretty sure that the webpage isn’t written in English. (Or, at least, isn’t *entirely* written in English.)

Here is a finite state automaton that can distinguish between possible and impossible words in Rotokas, a language spoken on the island of Bougainville off the coast of New Guinea. Rotokas has a very simple system of sounds and allows us to create a very small FSA.



An FSA works like a board game. Choose a word, and place your pencil on the space marked “Start”. Going through the letters of the word one at a time, move your pencil along the path marked with that letter. If the word ends and you’re at a space marked with a thicker circle, the word succeeds: it’s a possible Rotokas word! If the word ends and you’re not at a thicker circle, or you’re midway through the word and there’s no path corresponding to the next letter, the word fails: it’s *not* a possible Rotokas word!

Try it out with these possible and impossible words; the automaton should accept all the possible words and reject the impossible ones.

Possible Rotokas words	
<i>tauo</i>	<i>kareveiepa</i>
<i>puraveva</i>	<i>ovokirovua</i>
<i>avaopa</i>	<i>ouragaveva</i>

Impossible Rotokas words	
<i>grio</i>	<i>ouag</i>
<i>ovgi</i>	<i>vonoka</i>
<i>gataap</i>	<i>oappa</i>

11. Now, using the automaton above, put a check mark next to each possible Rotokas word:

- | | | |
|--------------------------------------|--|---|
| <input type="checkbox"/> <i>iu</i> | <input type="checkbox"/> <i>uente</i> | <input type="checkbox"/> <i>voav</i> |
| <input type="checkbox"/> <i>idau</i> | <input type="checkbox"/> <i>urioo</i> | <input type="checkbox"/> <i>uaia</i> |
| <input type="checkbox"/> <i>oire</i> | <input type="checkbox"/> <i>raorao</i> | <input type="checkbox"/> <i>oratreopaveiepa</i> |

On her visit to Armenia, Millie has gotten lost in Yerevan, the nation's capital. She is now at the Metropoliten (subway) station named Shengavit, but her friends are waiting for her at the station named Barekamutyun. Can you help Millie meet up with her friends?

1. Assuming Millie takes a train in the right direction, which will be the first stop after Shengavit?

Note that all names of stations listed below appear on the map.

- a. Gortsaranayin
 - b. Zoravar Andranik
 - c. Charbakh
 - d. Garegin Njdehi Hraparak
 - e. none of the above

2. After boarding at Shengavit, how many stops will it take Millie to get to Barekamutyun (don't include Shengavit itself in the number of stops)?



NACLO: Computational Problems

- <http://clair.si.umich.edu/naclo/resources/resources.html>
- List of computational problems:
 - <http://www.naclo.cs.cmu.edu/problems2014/N2014-O.pdf>
 - <http://www.naclo.cs.cmu.edu/problems2014/N2014-C.pdf>
 - <http://www.naclo.cs.cmu.edu/problems2014/N2014-I.pdf>
 - <http://www.naclo.cs.cmu.edu/problems2014/N2014-L.pdf>
 - <http://www.naclo.cs.cmu.edu/problems2013/N2013-C.pdf>
 - <http://www.naclo.cs.cmu.edu/problems2013/N2013-F.pdf>
 - <http://www.naclo.cs.cmu.edu/problems2013/N2013-H.pdf>
 - <http://www.naclo.cs.cmu.edu/problems2013/N2013-L.pdf>
 - <http://www.naclo.cs.cmu.edu/problems2013/N2013-N.pdf>
 - <http://www.naclo.cs.cmu.edu/problems2013/N2013-O.pdf>
 - <http://www.naclo.cs.cmu.edu/problems2012/N2012-C.pdf>
 - <http://www.naclo.cs.cmu.edu/problems2012/N2012-K.pdf>
 - <http://www.naclo.cs.cmu.edu/problems2012/N2012-O.pdf>
 - <http://www.naclo.cs.cmu.edu/problems2012/N2012-R.pdf>
 - <http://www.naclo.cs.cmu.edu/problems2011/F.pdf>
 - <http://www.naclo.cs.cmu.edu/problems2011/M.pdf>
 - <http://www.naclo.cs.cmu.edu/problems2010/D.pdf>
 - <http://www.naclo.cs.cmu.edu/problems2010/E.pdf>
 - <http://www.naclo.cs.cmu.edu/problems2010/I.pdf>
 - <http://www.naclo.cs.cmu.edu/problems2010/K.pdf>
 - <http://www.naclo.cs.cmu.edu/problems2009/N2009-E.pdf>
 - <http://www.naclo.cs.cmu.edu/problems2009/N2009-G.pdf>
 - <http://www.naclo.cs.cmu.edu/problems2009/N2009-I.pdf>
 - <http://www.naclo.cs.cmu.edu/problems2009/N2009-M.pdf>
 - <http://www.naclo.cs.cmu.edu/problems2008/N2008-F.pdf>
 - <http://www.naclo.cs.cmu.edu/problems2008/N2008-H.pdf>
 - <http://www.naclo.cs.cmu.edu/problems2008/N2008-I.pdf>
 - <http://www.naclo.cs.cmu.edu/problems2008/N2008-L.pdf>
 - <http://www.naclo.cs.cmu.edu/problems2007/N2007-A.pdf>
 - <http://www.naclo.cs.cmu.edu/problems2007/N2007-H.pdf>

NLP

NLP

Introduction to NLP

Preprocessing

Text Preprocessing

- Removing non-text (e.g., ads, javascript)
- Dealing with text encoding (e.g., Unicode)
- Sentence segmentation
- Normalization
 - labeled/labelled, extra-terrestrial/extraterrestrial, extra terrestrial
- Stemming
 - computer/computation
- Morphological analysis
 - car/cars
- Capitalization
 - Now/NOW, led/LED
- Named entity extraction
 - USA/usa

Text Preprocessing

- **Types vs. Tokens**
 - To be or not to be
- **Tokenization:**
 - ALS vs. A.L.S.
 - Paul's, Willow Dr., Dr. Willow, New York, ad hoc, can't
 - “The New York–Los Angeles flight” vs. “Minneapolis–St.Paul”
 - Numbers, e.g., (888) 555–1313, 1–888–555–1313
 - Dates, e.g., Jan–13–2012, 20120113, 13 January 2012, 01/13/12
 - URLs

Word Segmentation

- 金属製品製造の日立金属は19日、世界最大手の鉄鋳物メーカー「ワウパカ ファウンドリー ホールディングス」(米国・デラウェア州)を米投資ファンドから買収し、完全子会社にすると発表した。買収額は13億ドル(約1330億円)で、10月中にも手続きを終える。

Word Segmentation

- Arabic:

كتاب

- Japanese:

この本は重い。

(kono hon ha omoi)

- German:

Finanzdienstleistung = financial services

- Chinese:

电视 (television)

电 (diàn = electric) 视 (shì = to look at)

Text Preprocessing

ニューヨーク (New York) は、アメリカ合衆国 ニューヨーク州にある都市

- Kanji, Katakana, Hiragana, Rōmaji, (numbers)
- Nyūyōku wa, Amerikagasshūkoku nyūyōku-shū ni aru toshi

Sentence Boundary Recognition

- Decision trees
- Features
 - punctuation
 - formatting
 - fonts
 - spacing
 - capitalization
 - case
 - use of abbreviations, e.g., Dr., a.m.
- Example
 - If there is no space after a period, don't assume that there is a sentence boundary

NLP

NLP

Text Similarity

*Semantic Similarity:
Synonymy and other Semantic
Relations*

Synonyms and Paraphrases

- Example: post-close market announcements

The S&P 500 climbed 6.93, or 0.56 percent, to 1,243.72,

its best close

since June 12, 2001.

The Nasdaq gained 12.22, or 0.56 percent, to 2,198.44

for its best showing

since June 8, 2001.

The DJIA rose 68.46, or 0.64 percent, to 10,705.55,

its highest level

since March 15.

Synonyms

- Different words (and also word compounds) can have similar meanings.
 - For example, the adjectives *tepid* and *lukewarm* have very similar meanings and can be substituted for one another (*tepid water* vs. *lukewarm water*).
- True synonyms are actually relatively rare.
 - For example, even though *big* and *large* are often thought of as synonyms, consider the difference between *Big Leagues* and *Large Leagues*. ☺
- The verbs *sweat* and *perspire* are also near synonyms.
 - However, they differ in their frequency of use and the type of text in which they are likely to appear.

Polysemy

- Polysemy is the property of words to have multiple senses.
- For example, the noun *book* can refer to the following:
 - A literary work (e.g., “Anna Karenina”)
 - A stack of pages (e.g., a notebook)
 - A record of business transactions (think “bookkeeper”)
 - A record of bets (think “bookmaker”)
 - A list of buy and sell orders in a financial market

Polysemy

- The same word can also have multiple parts of speech, each with its own set of senses. For example, the word *book*, as a verb can mean “make a reservation for” or “occupy”.
- The different senses of the same word don’t have to be equally frequent.
- Some of the senses may overlap (e.g., the first two senses of *book* on the previous slide). That’s partially why different dictionaries list different sets of word senses for the same word.
 - “My favorite books are Anna Karenina and my father’s checkbook” ☺
- Some words can be highly polysemous (e.g., the verb “get” has at least 35 different meanings, according to Wordnet).

Other Semantic Relations

- Antonymy (near opposites)
 - *raise-lower*
- Hyponymy
 - a *deer* is a hyponym for *elk*
- Hyponymy (the inverse of hyponymy)
- Membership Meronymy:
 - a *flock* includes *sheep* (or *birds*)
- Part Meronymy:
 - a *table* has *legs*

Synsets

- Semantic relations hold between word senses, not between words.
- Examples:
 - the antonym of *hot* can be either *mild* or *cold* (or *unattractive*) depending on the specific sense of *hot*.
 - the immediate hypernym of *bar* can be one of the following, among others: *room*, *musical notation*, *obstruction*, *profession*, depending on the sense of *bar*.
- The term *synset* is used to group together all synonyms of the same word. If a word is polysemous, it may be associated with multiple synsets.

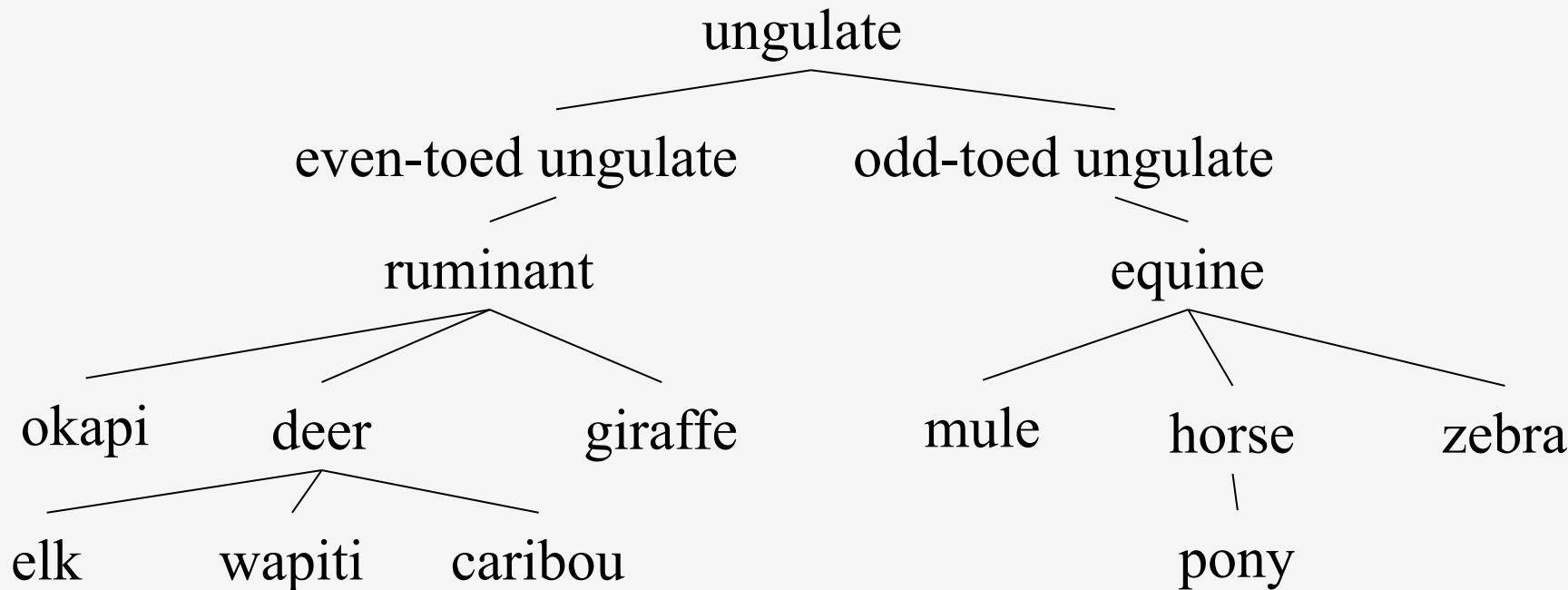
Text Similarity

Wordnet

Wordnet

- Wordnet is a project run by George Miller (1920–2012) and Christiane Fellbaum at Princeton University.
- It includes a database of words (mainly nouns and verbs but also adjectives and adverbs) and semantic relations between them.
- The main relation is hypernymy, so the overall structure of the database is more tree-like (see next slide).
- References:
 - George A. Miller (1995). WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39–41.
 - Christiane Fellbaum (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Tree-like Structure of Wordnet



Wordnet Example (1/6)

The noun bar has 11 senses

1. barroom, bar, saloon, ginmill, taproom -- (a room where alcoholic drinks are served over a counter)
2. bar -- (a counter where you can purchase food or drink)
3. bar -- (a rigid piece of metal)
4. measure, bar -- (notation for a repeating pattern of musical beats; written followed by a vertical bar)
5. bar -- (usually metal placed in windows to prevent escape)
6. prevention, bar -- (the act of preventing)
7. bar -- (a unit of pressure equal to a million dynes per square centimeter)
8. bar -- (a submerged (or partly submerged) ridge in a river or along a shore)
9. legal profession, bar, legal community -- (the body of individuals qualified to practice law)
10. cake, bar -- (a block of soap or wax)
11. bar -- ((law) a railing that encloses the part of the courtroom where the judges and lawyers sit and the case is tried)

The verb bar has 4 senses

1. bar, debar, exclude -- (prevent from entering; keep out; "He was barred from membership in the club")
2. barricade, block, blockade, block off, block up, bar -- (render unsuitable for passage; "block the way"; "barricade the streets")
3. banish, relegate, bar -- (expel, as if by official decree; "he was banished from his own country")
4. bar -- (secure with, or as if with, bars; "He barred the door")

Wordnet Example (2/6)

Sense 1

barroom, bar, saloon, ginmill, taproom

=> room

=> area

=> structure, construction

=> artifact, artefact

=> object, physical object

=> entity, something

Sense 2

bar

=> counter

=> table

=> furniture, piece of furniture, article of furniture

=> furnishings

=> instrumentality, instrumentation

=> artifact, artefact

=> object, physical object

=> entity, something

Wordnet Example (3/6)

Sense 3

bar

- => implement
- => instrumentality, instrumentation
- => artifact, artefact
- => object, physical object
- => entity, something

Sense 4

measure, bar

- => musical notation
- => notation, notational system
- => writing, symbolic representation
- => written communication, written language
- => communication
- => social relation
- => relation
- => abstraction

Wordnet Example (4/6)

Sense 5

bar

- => obstruction, impediment, impedimenta
- => structure, construction
- => artifact, artefact
- => object, physical object
- => entity, something

Sense 6

prevention, bar

- => hindrance, interference, interfering
- => act, human action, human activity

Sense 7

bar

- => pressure unit
- => unit of measurement, unit
- => definite quantity
- => measure, quantity, amount, quantum
- => abstraction

Wordnet Example (5/6)

Sense 8

bar

=> ridge

=> natural elevation, elevation

=> geological formation, geology, formation

=> natural object

=> object, physical object

=> entity, something

=> barrier

=> mechanism

=> natural object

=> object, physical object

=> entity, something

Wordnet Example (6/6)

Sense 9

legal profession, bar, legal community

=> profession, community

=> occupation, vocation, occupational group

=> body

=> gathering, assemblage

=> social group

=> group, grouping

Sense 10

cake, bar

=> block

=> artifact, artefact

=> object, physical object

=> entity, something

Familiarity and Polysemy

board used as a noun is familiar (polysemy count = 9)

bird used as a noun is common (polysemy count = 5)

cat used as a noun is common (polysemy count = 7)

house used as a noun is familiar (polysemy count = 11)

information used as a noun is common (polysemy count = 5)

retrieval used as a noun is uncommon (polysemy count = 3)

serendipity used as a noun is very rare (polysemy count = 1)

Text Similarity

Other Lexical Networks

External Links

- EuroWordNet
 - <http://www illc uva nl/EuroWordNet/>
- Open Thesaurus
 - <http://www openthesaurus de/>
- Freebase
 - <http://www freebase com>
- DBPedia
 - <http://www dbpedia org>
- BabelNet
 - <http://babelnet org>
- Various thesauri

BabelNet Example

MeSH

Medical Subject Headings

The screenshot shows a web browser window titled "MeSH Browser Record". The URL in the address bar is www.nlm.nih.gov/cgi/mesh/2014/MB.cgi. The page displays a hierarchical list of Medical Subject Headings (MeSH) terms under the category "Nervous System Diseases [C10]". The terms are listed in a tree structure with plus signs indicating expandable categories. Some terms have arrows pointing to their definitions or detailed descriptions.

- [Nervous System Diseases \[C10\]](#)
 - [Central Nervous System Diseases \[C10.228\]](#)
 - [Brain Diseases \[C10.228.140\]](#)
 - [Akinetic Mutism \[C10.228.140.042\]](#)
 - [Amblyopia \[C10.228.140.055\]](#)
 - [Amnesia, Transient Global \[C10.228.140.060\]](#)
 - [Auditory Diseases, Central \[C10.228.140.068\]](#) +
 - [Basal Ganglia Diseases \[C10.228.140.079\]](#) +
 - [Brain Abscess \[C10.228.140.116\]](#) +
 - [Brain Damage, Chronic \[C10.228.140.140\]](#) +
 - [Brain Death \[C10.228.140.151\]](#)
 - [Brain Diseases, Metabolic \[C10.228.140.163\]](#) +
 - [Brain Edema \[C10.228.140.187\]](#)
 - [Brain Injuries \[C10.228.140.199\]](#) +
 - [Brain Neoplasms \[C10.228.140.211\]](#) +
 - [Cerebellar Diseases \[C10.228.140.252\]](#) +
 - [Cerebrovascular Disorders \[C10.228.140.300\]](#) +
 - [Dementia \[C10.228.140.380\]](#) +
 - [Diffuse Cerebral Sclerosis of Schilder \[C10.228.140.400\]](#)
 - [Encephalitis \[C10.228.140.430\]](#)
 - [Anti-N-Methyl-D-Aspartate Receptor Encephalitis \[C10.228.140.430.124\]](#)
 - [Cerebral Ventriculitis \[C10.228.140.430.249\]](#)
 - [Encephalomyelitis \[C10.228.140.430.500\]](#)
 - [Limbic Encephalitis \[C10.228.140.430.525\]](#)
 - [Meningoencephalitis \[C10.228.140.430.550\]](#) +
 - [Encephalomalacia \[C10.228.140.461\]](#) +
 - [Epilepsy \[C10.228.140.490\]](#) +
 - [Headache Disorders \[C10.228.140.546\]](#) +
 - [Hydrocephalus \[C10.228.140.602\]](#) +
 - [Hypothalamic Disease \[C10.228.140.617\]](#) +

- <http://www.nlm.nih.gov/mesh/MBrowser.html>

NLP

NLP

Text Similarity

*Thesaurus-based Word Similarity
Methods*

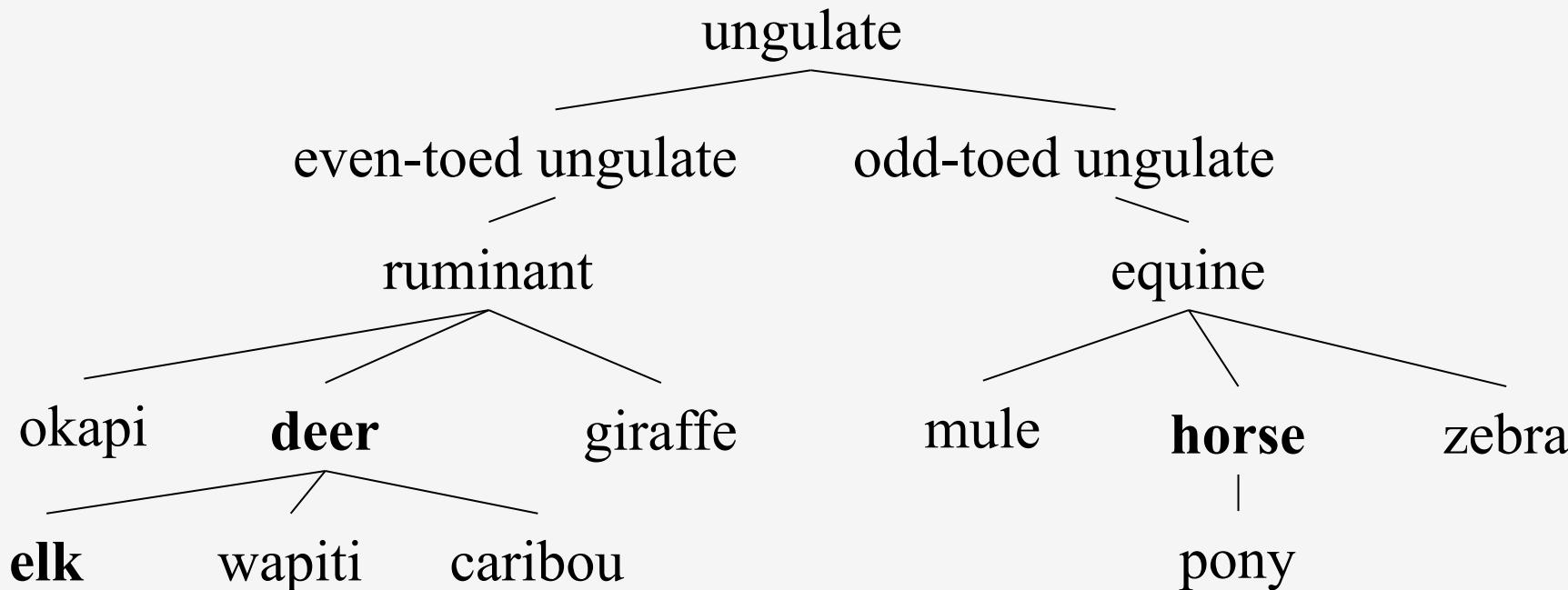
Quiz

- Which pair of words exhibits the greatest similarity?
 - 1. Deer-elk
 - 2. Deer-horse
 - 3. Deer-mouse
 - 4. Deer-roof

Quiz Answer

- Which pair of words exhibits the greatest similarity?
 - 1. Deer-elk
 - 2. Deer-horse
 - 3. Deer-mouse
 - 4. Deer-roof
- Why?
- Remember the Wordnet tree:

Remember Wordnet



Path Similarity

- Version 1
 - $\text{Sim}(v,w) = -\text{pathlength}(v,w)$
- Version 2
 - $\text{Sim}(v,w) = -\log \text{pathlength}(v,w)$

Problems With This Approach

- There may be no tree for the specific domain or language
- A specific word (e.g., a term or a proper noun) may not be in any tree
- IS-A (hypernym) edges are not all equally apart in similarity space

Path Similarity Between Two Words

- Version 3 (Philip Resnik)

$$\text{Sim}(v, w) = -\log P(\text{LCS}(v, w))$$

where LCS = lowest common subsumer,
e.g.

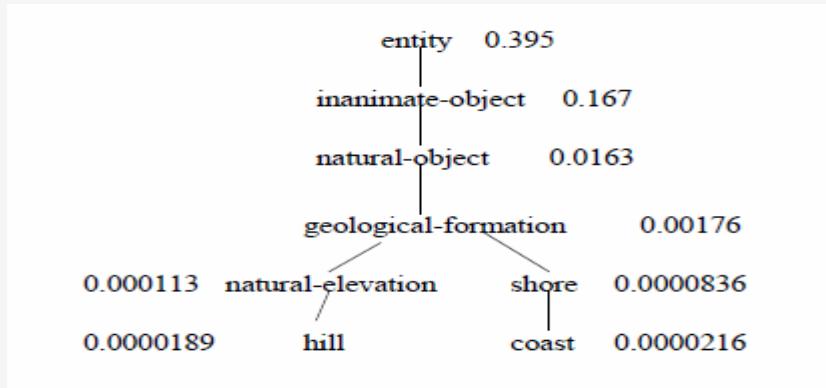
ungulate for deer and horse

deer for deer and elk

Information Content

- Version 4 (Dekang Lin)

- Wordnet augmented with probabilities (Lin 1998)
- $IC(c) = -\log P(c)$
- $Sim(v,w) = \frac{2 \times \log P(LCS(v,w))}{\log P(v) + \log P(w)}$



$$\text{sim}(\text{Hill}, \text{Coast}) = \frac{2 \times \log P(\text{Geological-Formation})}{\log P(\text{Hill}) + \log P(\text{Coast})}$$

$$= 0.59$$

Wordnet Similarity Software

- WordNet::Similarity (Perl)
 - <http://www.d.umn.edu/~tpederse/similarity.html>
- NLTK (Python)
 - <http://www.nltk.org>

```
>>> dog.lin_similarity(cat, brown_ic)
0.879
>>> dog.lin_similarity(elephant, brown_ic)
0.531
>>> dog.lin_similarity(elk, brown_ic)
0.475
```

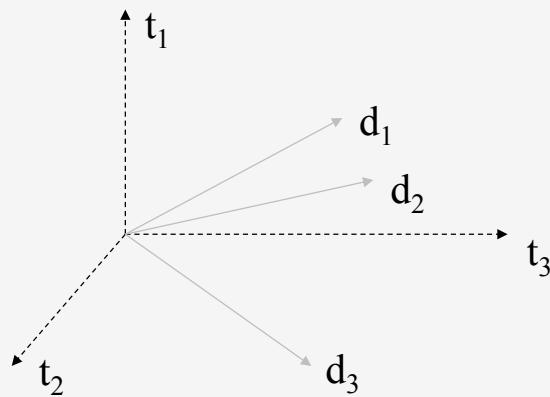
NLP

NLP

Text Similarity

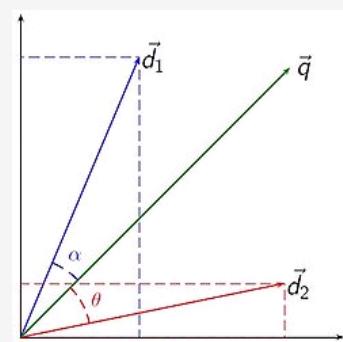
The Vector Space Model

The Vector Space Model



Document Similarity

- Used in information retrieval to determine which document (d_1 or d_2) is more similar to a given query q .
- Note that documents and queries are represented in the same space.
- Often, the angle between two vectors (or, rather, the cosine of that angle) is used as a proxy for the similarity of the underlying documents.



Cosine Similarity

- The Cosine measure is computed as the normalized dot product of two vectors:

$$\sigma(D, Q) = \frac{|D \cap Q|}{\sqrt{|D||Q|}} = \frac{\sum(d_i q_i)}{\sqrt{\sum(d_i)^2} \sqrt{\sum(q_i)^2}}$$

- A variant of Cosine is the Jaccard coefficient:

$$\sigma(D, Q) = \frac{|D \cap Q|}{|D \cup Q|}$$

Example

- What is the cosine similarity between:
 - D= “cat,dog,dog” = <1,2,0>
 - Q= “cat,dog,mouse,mouse” = <1,1,2>
- Answer:

$$\sigma(D, Q) = \frac{1 \times 1 + 2 \times 1 + 0 \times 2}{\sqrt{1^2 + 2^2 + 0^2} \sqrt{1^2 + 1^2 + 2^2}} = \frac{3}{\sqrt{5} \sqrt{6}} \approx 0.55$$

- In comparison:

$$\sigma(D, D) = \frac{1 \times 1 + 2 \times 2 + 0 \times 0}{\sqrt{1^2 + 2^2 + 0^2} \sqrt{1^2 + 2^2 + 0^2}} = \frac{5}{\sqrt{5} \sqrt{5}} = 1$$

Quiz

- Given the three documents
 - $D_1 = \langle 1, 3 \rangle$
 - $D_2 = \langle 10, 30 \rangle$
 - $D_3 = \langle 3, 1 \rangle$
- Compute the cosine scores
 - $\sigma(D_1, D_2)$
 - $\sigma(D_1, D_3)$
- What do the numbers tell you?

Answers to the Quiz

$$\sigma(D_1, D_2) = 1$$

one of the two documents is a scaled version of the other

$$\sigma(D_1, D_3) = 0.6$$

swapping the two dimensions results in a lower similarity

Quiz

- What is the range of values that the cosine score can take?

Answer to the Quiz

- In general, the cosine function has a range of $[-1,1]$
- However, when the two vectors are both in the first quadrant (since all word counts are non-negative), the range is $[0,1]$.

Text Similarity

The Vector Space Model Applied to Word Similarity

Distributional Similarity

- Two words that appear in similar contexts are likely to be semantically related, e.g.,
 - schedule a test **drive** and investigate **Honda**'s financing options
 - **Volkswagen** debuted a new version of its front-wheel-**drive** Golf
 - the **Jeep** reminded me of a recent **drive**
 - Our test **drive** took place at the wheel of loaded **Ford** EL model
- “You will know a word by the company that it keeps.” (J.R. Firth 1957)

Distributional Similarity

- The context can be any of the following:
 - The word before the target word
 - The word after the target word
 - Any word within n words of the target word
 - Any word within a specific syntactic relationship with the target word (e.g., the head of the dependency or the subject of the sentence)
 - Any word within the same sentence
 - Any word within the same document

Association Strength

- Frequency matters: we want to ignore spurious word pairings.
- However, frequency alone is not sufficient.
- A common technique is to use pointwise mutual information (PMI).
- Here w is a word and c is a feature from the context $\text{PMI}(w,c) = \log P(w,c)/P(w)P(c)$

NLP

NLP

Text Similarity

Dimensionality Reduction

Problems with the Simple Vector Approaches to Similarity

- Polysemy ($\text{sim} < \cos$)
 - bar, bank, jaguar, hot
- Synonymy ($\text{sim} > \cos$)
 - building/edifice, large/big, spicy/hot
- Relatedness (people are really good at figuring this)
 - doctor/patient/nurse/treatment
- Sparse matrix
- Needed
 - dimensionality reduction

TOEFL Synonyms and SAT Analogies

- Word similarity vs. analogies

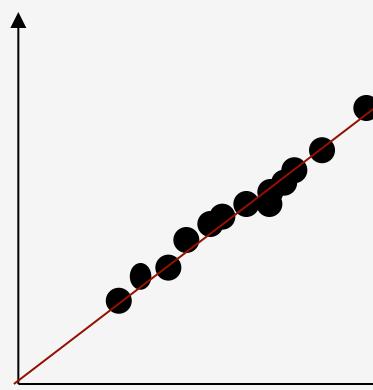
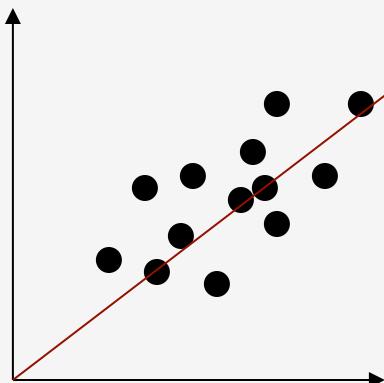
Stem:	levied
Choices:	(a) imposed (b) believed (c) requested (d) correlated
Solution:	(a) imposed

Stem:	mason:stone
Choices:	(a) teacher:chalk (b) carpenter:wood (c) soldier:gun (d) photograph:camera (e) book:word
Solution:	(b) carpenter:wood

Example from Peter Turney

Dimensionality Reduction

- Looking for hidden similarities in data
- Based on matrix decomposition
- Height/weight example



Vectors and Matrices

- A matrix is an $m \times n$ table of objects (in our case, numbers)
- Each row (or column) is a vector.
- Matrices of compatible dimensions can be multiplied together.
- What is the result of the multiplication below?

$$\begin{bmatrix} 1 & 2 & 4 \\ 2 & 5 & 7 \\ 4 & 9 & 14 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \\ -1 \end{bmatrix} = \begin{bmatrix} \quad \end{bmatrix}$$

Answer to the Quiz

$$\begin{bmatrix} 1 & 2 & 4 \\ 2 & 5 & 7 \\ 4 & 9 & 14 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \times 2 + 2 \times 1 + 4 \times (-1) \\ 2 \times 2 + 5 \times 1 + 7 \times (-1) \\ 4 \times 2 + 9 \times 1 + 14 \times (-1) \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \\ 3 \end{bmatrix}$$

Eigenvectors and Eigenvalues

- An eigenvector is an implicit “direction” for a matrix $A\vec{v} = \lambda\vec{v}$
- v (the eigenvector) is non-zero, though λ (the eigenvalue) can be any complex number in principle.
- Computing eigenvalues: $\det(A - \lambda I) = 0$

Eigenvectors and Eigenvalues

- Example:

$$A = \begin{pmatrix} -1 & 3 \\ 2 & 0 \end{pmatrix} \quad A - \lambda I = \begin{pmatrix} -1 - \lambda & 3 \\ 2 & -\lambda \end{pmatrix}$$

- $\det(A - \lambda I) = (-1 - \lambda)(-\lambda) - 3 \cdot 2 = 0$
- Then: $\lambda + \lambda^2 - 6 = 0$; $\lambda_1 = 2$; $\lambda_2 = -3$
- For $\lambda_1 = 2$:

$$\begin{pmatrix} -3 & 3 \\ 2 & -2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0$$

- Solutions: $x_1 = x_2$

Matrix Decomposition

- If Σ is a square matrix, it can be decomposed into $U\Lambda U^{-1}$, where

U = matrix of eigenvectors

Λ = diagonal matrix of eigenvalues

$$\Sigma U = U \Lambda$$

$$U^{-1} \Sigma U = \Lambda$$

$$\Sigma = U \Lambda U^{-1}$$

Example

$$S = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \lambda_1 = 1, \lambda_2 = 3$$

$$U = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$$

$$U^{-1} = \begin{pmatrix} 1/2 & -1/2 \\ 1/2 & 1/2 \end{pmatrix}$$

$$S = U\Lambda U^{-1} = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix} \begin{pmatrix} 1/2 & -1/2 \\ 1/2 & 1/2 \end{pmatrix}$$

SVD: Singular Value Decomposition

- $A = U\Sigma V^T$
 - U is the matrix of orthogonal eigenvectors of AA^T
 - V is the matrix of orthogonal eigenvectors of A^TA
 - The components of Σ are the eigenvalues of A^TA
- This decomposition exists for all matrices, dense or sparse
- If A has 5 columns and 3 rows, then U will be 5×5 and V will be 3×3
- In Matlab, use $[U, S, V] = svd(A)$

Example

D1: T6, T9

D2: T1, T2

D3: T2, T5, T8

D4: T1, T4, T6, T8, T9

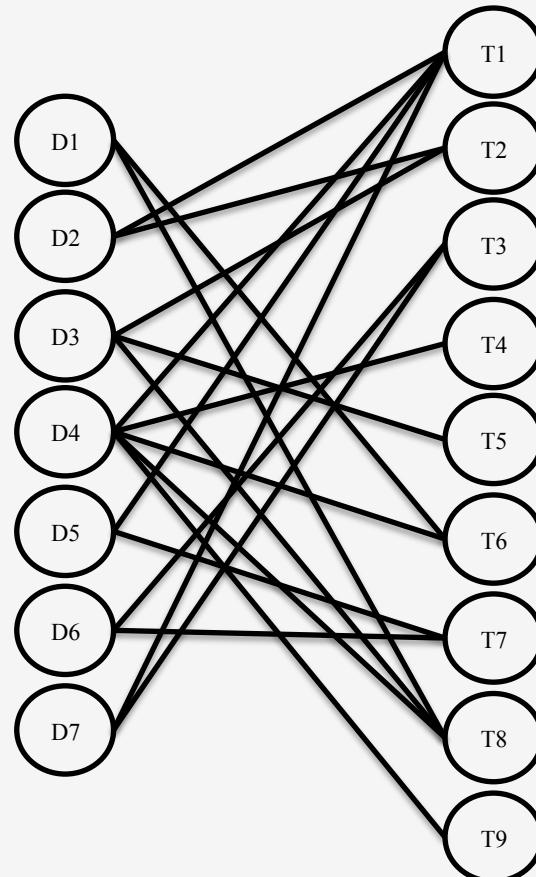
D5: T1, T7

D6: T3, T7

D7: T1, T3

Example

D1: T6, T9
D2: T1, T2
D3: T2, T5, T8
D4: T1, T4, T6, T8, T9
D5: T1, T7
D6: T3, T7
D7: T1, T3



Document-Term Matrix

$$A = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

raw

$$A^{(n)} = \begin{bmatrix} 0 & 0.58 & 0 & 0.45 & 0.71 & 0 & 0.71 \\ 0 & 0.58 & 0.58 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.71 & 0.71 \\ 0 & 0 & 0 & 0.45 & 0 & 0 & 0 \\ 0 & 0.58 & 0.58 & 0 & 0 & 0 & 0 \\ 0.71 & 0 & 0 & 0.45 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.71 & 0.71 & 0 \\ 0 & 0 & 0.58 & 0.45 & 0 & 0 & 0 \\ 0.71 & 0 & 0 & 0.45 & 0 & 0 & 0 \end{bmatrix}$$

normalized

Decomposition

u =

-0.6976	-0.0945	0.0174	-0.6950	0.0000	0.0153	0.1442	-0.0000	0
-0.2622	0.2946	0.4693	0.1968	-0.0000	-0.2467	-0.1571	-0.6356	0.3098
-0.3519	-0.4495	-0.1026	0.4014	0.7071	-0.0065	-0.0493	-0.0000	0.0000
-0.1127	0.1416	-0.1478	-0.0734	0.0000	0.4842	-0.8400	0.0000	-0.0000
-0.2622	0.2946	0.4693	0.1968	0.0000	-0.2467	-0.1571	0.6356	-0.3098
-0.1883	0.3756	-0.5035	0.1273	-0.0000	-0.2293	0.0339	-0.3098	-0.6356
-0.3519	-0.4495	-0.1026	0.4014	-0.7071	-0.0065	-0.0493	0.0000	-0.0000
-0.2112	0.3334	0.0962	0.2819	-0.0000	0.7338	0.4659	-0.0000	0.0000
-0.1883	0.3756	-0.5035	0.1273	-0.0000	-0.2293	0.0339	0.3098	0.6356

v =

-0.1687	0.4192	-0.5986	0.2261	0	-0.5720	0.2433
-0.4472	0.2255	0.4641	-0.2187	0.0000	-0.4871	-0.4987
-0.2692	0.4206	0.5024	0.4900	-0.0000	0.2450	0.4451
-0.3970	0.4003	-0.3923	-0.1305	0	0.6124	-0.3690
-0.4702	-0.3037	-0.0507	-0.2607	-0.7071	0.0110	0.3407
-0.3153	-0.5018	-0.1220	0.7128	-0.0000	-0.0162	-0.3544
-0.4702	-0.3037	-0.0507	-0.2607	0.7071	0.0110	0.3407

Decomposition

Spread on the v1 axis

$$s = \boxed{1.5849}$$

	0	0	0	0	0	0	0
0	1.2721	0	0	0	0	0	0
0	0	1.1946	0	0	0	0	0
0	0	0	0.7996	0	0	0	0
0	0	0	0	0.7100	0	0	0
0	0	0	0	0	0.5692	0	0
0	0	0	0	0	0	0.1977	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

Related to Dimensionality Reduction

- Singular value decomposition:
 - $A = U\Sigma V^T$
- Dimensionality reduction
 - $A^* = U\Sigma^* V^T$
 - Where Σ^* keeps only the largest eigenvalues

Rank-4 Approximation

s4 =

Rank-4 Approximation

u^*s4^*v'

-0.0019	0.5985	-0.0148	0.4552	0.7002	0.0102	0.7002
-0.0728	0.4961	0.6282	0.0745	0.0121	-0.0133	0.0121
0.0003	-0.0067	0.0052	-0.0013	0.3584	0.7065	0.3584
0.1980	0.0514	0.0064	0.2199	0.0535	-0.0544	0.0535
-0.0728	0.4961	0.6282	0.0745	0.0121	-0.0133	0.0121
0.6337	-0.0602	0.0290	0.5324	-0.0008	0.0003	-0.0008
0.0003	-0.0067	0.0052	-0.0013	0.3584	0.7065	0.3584
0.2165	0.2494	0.4367	0.2282	-0.0360	0.0394	-0.0360
0.6337	-0.0602	0.0290	0.5324	-0.0008	0.0003	-0.0008

Rank-4 Approximation

u^*s4

-1.1056	-0.1203	0.0207	-0.5558	0	0	0
-0.4155	0.3748	0.5606	0.1573	0	0	0
-0.5576	-0.5719	-0.1226	0.3210	0	0	0
-0.1786	0.1801	-0.1765	-0.0587	0	0	0
-0.4155	0.3748	0.5606	0.1573	0	0	0
-0.2984	0.4778	-0.6015	0.1018	0	0	0
-0.5576	-0.5719	-0.1226	0.3210	0	0	0
-0.3348	0.4241	0.1149	0.2255	0	0	0
-0.2984	0.4778	-0.6015	0.1018	0	0	0

Rank-4 Approximation

 $s4 * v'$

-0.2674	-0.7087	-0.4266	-0.6292	-0.7451	-0.4996	-0.7451
0.5333	0.2869	0.5351	0.5092	-0.3863	-0.6384	-0.3863
-0.7150	0.5544	0.6001	-0.4686	-0.0605	-0.1457	-0.0605
0.1808	-0.1749	0.3918	-0.1043	-0.2085	0.5700	-0.2085
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0

Rank-2 Approximation

s2 =

Rank-2 Approximation

u^*s^2v'

0.1361	0.4673	0.2470	0.3908	0.5563	0.4089	0.5563
0.2272	0.2703	0.2695	0.3150	0.0815	-0.0571	0.0815
-0.1457	0.1204	-0.0904	-0.0075	0.4358	0.4628	0.4358
0.1057	0.1205	0.1239	0.1430	0.0293	-0.0341	0.0293
0.2272	0.2703	0.2695	0.3150	0.0815	-0.0571	0.0815
0.2507	0.2412	0.2813	0.3097	-0.0048	-0.1457	-0.0048
-0.1457	0.1204	-0.0904	-0.0075	0.4358	0.4628	0.4358
0.2343	0.2454	0.2685	0.3027	0.0286	-0.1073	0.0286
0.2507	0.2412	0.2813	0.3097	-0.0048	-0.1457	-0.0048

Rank-2 Approximation

u^*s_2 - word vector representation in concept space

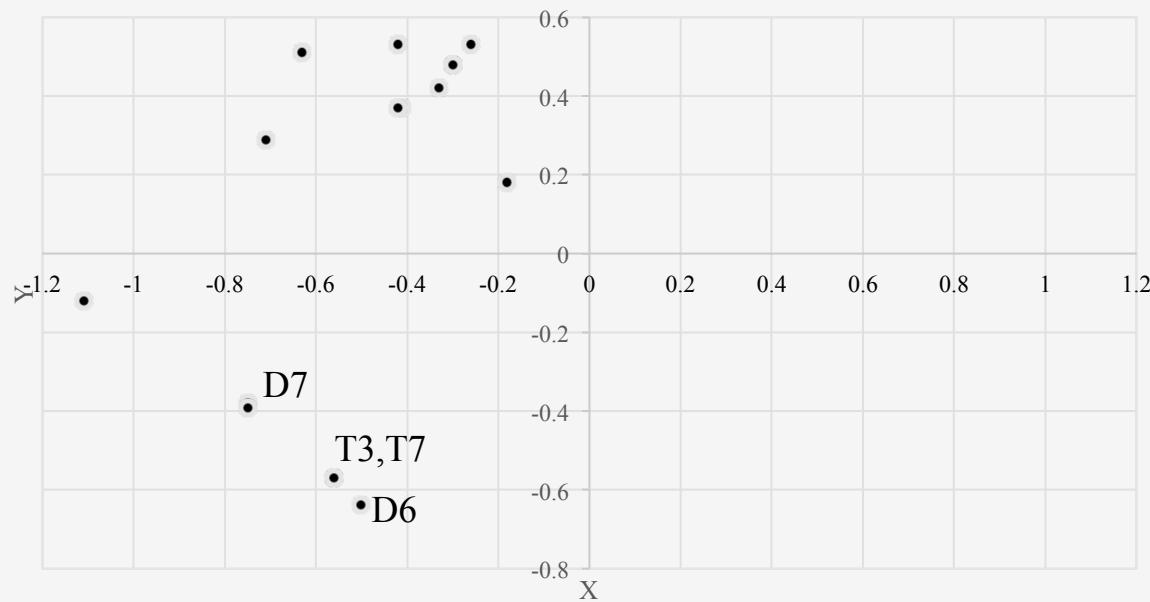
-1.1056	-0.1203	0	0	0	0	0
-0.4155	0.3748	0	0	0	0	0
-0.5576	-0.5719	0	0	0	0	0
-0.1786	0.1801	0	0	0	0	0
-0.4155	0.3748	0	0	0	0	0
-0.2984	0.4778	0	0	0	0	0
-0.5576	-0.5719	0	0	0	0	0
-0.3348	0.4241	0	0	0	0	0
-0.2984	0.4778	0	0	0	0	0

Rank-2 Approximation

s2*v' - new concept representation of the documents

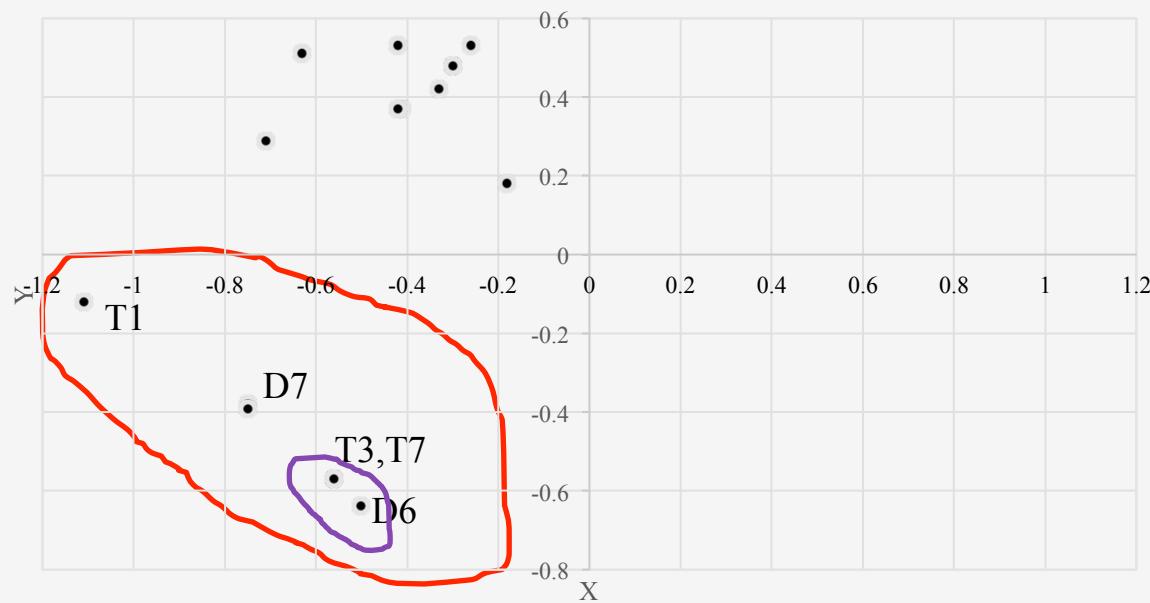
D1	D2	D3	D4	D5	D6	D7
-0.26	-0.71	-0.42	-0.63	-0.75	-0.5	-0.75
0.53	0.29	0.53	0.51	-0.38	-0.64	-0.39

T1	T2	T3	T4	T5	T6	T7	T8	T9
-1.11	-0.41	-0.56	-0.18	-0.42	-0.3	-0.56	-0.33	-0.3
-0.12	0.37	-0.57	0.18	0.37	0.48	-0.57	0.42	0.48



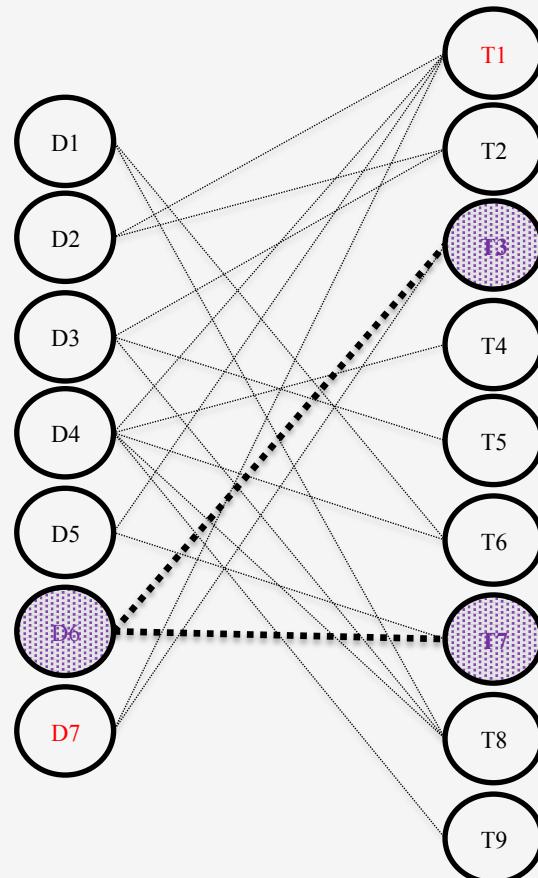
D1	D2	D3	D4	D5	D6	D7
-0.26	-0.71	-0.42	-0.63	-0.75	-0.5	-0.75
0.53	0.29	0.53	0.51	-0.38	-0.64	-0.39

T1	T2	T3	T4	T5	T6	T7	T8	T9
-1.11	-0.41	-0.56	-0.18	-0.42	-0.3	-0.56	-0.33	-0.3
-0.12	0.37	-0.57	0.18	0.37	0.48	-0.57	0.42	0.48



Example

D1: T6, T9
D2: T1, T2
D3: T2, T5, T8
D4: T1, T4, T6, T8, T9
D5: T1, T7
D6: T3, T7
D7: T1, T3



Documents to Concepts and Terms to Concepts

```
>> A(:,1)'*u*s
-0.4238    0.6784   -0.8541    0.1446   -0.0000   -0.1853    0.0095
```

```
>> A(:,1)'*u*s4
-0.4238    0.6784   -0.8541    0.1446      0      0      0
```

```
>> A(:,1)'*u*s2
-0.4238    0.6784      0      0      0      0      0
```

```
>> A(:,2)'*u*s2
-1.1233    0.3650      0      0      0      0      0
```

```
>> A(:,3)'*u*s2
-0.6762    0.6807      0      0      0      0      0
```

Documents to Concepts and Terms to Concepts

```
>> A(:, 4)' * u * s2
-0.9972    0.6478      0      0      0      0      0
>> A(:, 5)' * u * s2
-1.1809   -0.4914      0      0      0      0      0
>> A(:, 6)' * u * s2
-0.7918   -0.8121      0      0      0      0      0
>> A(:, 7)' * u * s2
-1.1809   -0.4914      0      0      0      0      0
```

Cont'd

```
>> (s2*v'*A(1,:))'  
-1.7523 -0.1530 0 0 0 0 0 0 0  
  
>> (s2*v'*A(2,:))'  
-0.6585 0.4768 0 0 0 0 0 0 0  
  
>> (s2*v'*A(3,:))'  
-0.8838 -0.7275 0 0 0 0 0 0 0  
  
>> (s2*v'*A(4,:))'  
-0.2831 0.2291 0 0 0 0 0 0 0  
  
>> (s2*v'*A(5,:))'  
-0.6585 0.4768 0 0 0 0 0 0 0
```

Cont'd

```
>> (s2*v'*A(6,:))'  
-0.4730    0.6078    0    0    0    0    0    0    0  
  
>> (s2*v'*A(7,:))'  
-0.8838   -0.7275    0    0    0    0    0    0    0  
  
>> (s2*v'*A(8,:))'  
-0.5306    0.5395    0    0    0    0    0    0    0  
  
>> (s2*v'*A(9,:))'  
-0.4730    0.6078    0    0    0    0    0    0    0
```

Properties

A is a document to term matrix. What is A^*A' ?

A^*A'

1.5471	0.3364	0.5041	0.2025	0.3364	0.2025	0.5041	0.2025	0.2025
0.3364	0.6728	0	0	0.6728	0	0	0.3364	0
0.5041	0	1.0082	0	0	0	0.5041	0	0
0.2025	0	0	0.2025	0	0.2025	0	0.2025	0.2025
0.3364	0.6728	0	0	0.6728	0	0	0.3364	0
0.2025	0	0	0.2025	0	0.7066	0	0.2025	0.7066
0.5041	0	0.5041	0	0	0	1.0082	0	0
0.2025	0.3364	0	0.2025	0.3364	0.2025	0	0.5389	0.2025
0.2025	0	0	0.2025	0	0.7066	0	0.2025	0.7066

Properties

What about A^*A ?

A^*A

1.0082	0	0	0.6390	0	0	0	0
0	1.0092	0.6728	0.2610	0.4118	0	0	0.4118
0	0.6728	1.0092	0.2610	0	0	0	0
0.6390	0.2610	0.2610	1.0125	0.3195	0	0	0.3195
0	0.4118	0	0.3195	1.0082	0.5041	0.5041	0.5041
0	0	0	0	0.5041	1.0082	0.5041	0.5041
0	0.4118	0	0.3195	0.5041	0.5041	0.5041	1.0082

Latent Semantic Indexing (LSI)

- Dimensionality reduction = identification of hidden (latent) concepts
- Query matching in latent space

External Pointers

- <http://lsa.colorado.edu>
- <http://www.cs.utk.edu/~lsi>

NLP

NLP

Introduction to NLP

NLP Tasks

Part of Speech Tagging

The swimmer is getting ready to run in the final race.

Part of Speech Tagging

The swimmer is getting ready to **run** in the final race.

- Run – verb or noun?
- Final – noun or adjective?
- Race – verb or noun?

Part of Speech Tagging

The candidate is preparing for his **run** for the presidency.
The swimmer is getting ready to **run** in the final race.

Parsing

- Myriam slept.
- Myriam wrote a novel.
- Myriam gave Sally flowers.
- Myriam ate pizza with olives.
- Myriam ate pizza with Sally.
- Myriam ate pizza with a fork.
- Myriam ate pizza with remorse.

Phrase-Structure Grammar

S → NP VP

NP → DET N

NP → NP PP

VP → VBD

VP → VBD NP

VP → VBD NP NP

VP → VP PP

PP → PRP NP

DET → *the*

DET → *that*

DET → *a*

N → *child*

N → *window*

N → *car*

VBD → *found*

VBD → *ate*

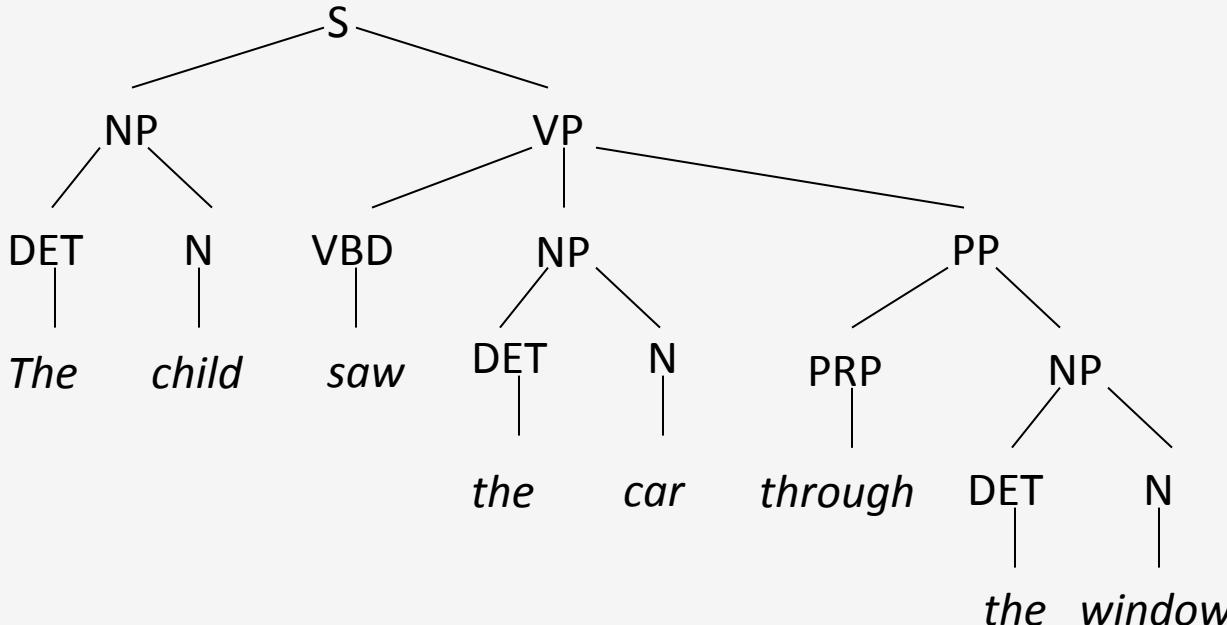
VBD → *saw*

PRP → *in*

PRP → *of*

PRP → *through*

Parse Trees



Stanford Parser

Stanford Parser

Please enter a sentence to be parsed:

```
Housing starts, the number of new homes being built, rose 7.2% in March to an annual rate of 549,000 units, up from a revised 512,000 in February, the Commerce Department said.
```

Language: English

Your query

Housing starts, the number of new homes being built, rose 7.2% in March to an annual rate of 549,000 units, up from a revised 512,000 in February, the Commerce Department said.

Tagging

```
Housing/NN starts/NNS ,/, the/DT number/NN of/IN new/JJ homes/NNS being/VBG built/VBN ,/, rose/VBD 7.2/CD %/NN in/IN March/NNP to/TO an/DT annual/JJ rate/NN of/IN 549,000/CD units/NNS ,/, up/RB from/IN a/DT revised/VBN 512,000/CD in/IN February/NNP ,/, the/DT Commerce/NNP Department/NNP said/VBD ./.
```

Parse

```
(ROOT
  (S
    (S
      (NP
        (NP (NN Housing) (NNS starts))
        (, ,)
        ...)
```

Done

Parser Output

```
(ROOT
  (S
    (S
      (NP
        (NP (NN Housing) (NNS starts))
        (, ,)
        (NP
          (NP (DT the) (NN number))
          (PP (IN of)
            (NP
              (NP (JJ new) (NNS homes))
              (VP (VBG being)
                (VP (VBN built)))))))
        (, ,)))
```

```
(VP (VBD rose)
  (NP (CD 7.2) (NN %))
  (PP (IN in)
    (NP (NNP March)))
  (PP (TO to)
    (NP
      (NP (DT an) (JJ annual) (NN rate))
      (PP (IN of)
        (NP (CD 549,000) (NNS units))))))
  (, ,)
  (ADVP (RB up)
    (PP (IN from)
      (NP
        (NP (DT a) (VBN revised) (CD 512,000))
        (PP (IN in)
          (NP (NNP February)))))))
  (, ,)
  (NP (DT the) (NNP Commerce) (NNP Department))
  (VP (VBD said))
  (. .)))
```

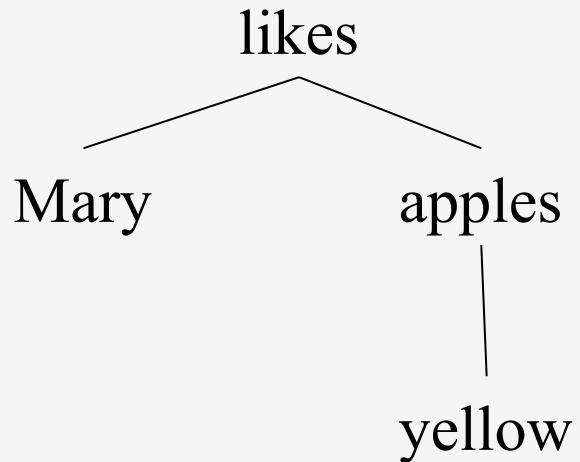
This Problem is Pretty // Easy

- Commercial for a phone company
- Garden path sentences
 - Don't bother coming
 - Don't bother coming early
 - Take the turkey out of the oven at five
 - Take the turkey out of the over at five to four
 - I got canned
 - I got canned peaches for dinner
 - All Americans need to buy a house
 - All Americans need to buy a house is a lot of money
- Can you think of more such examples?

Solution

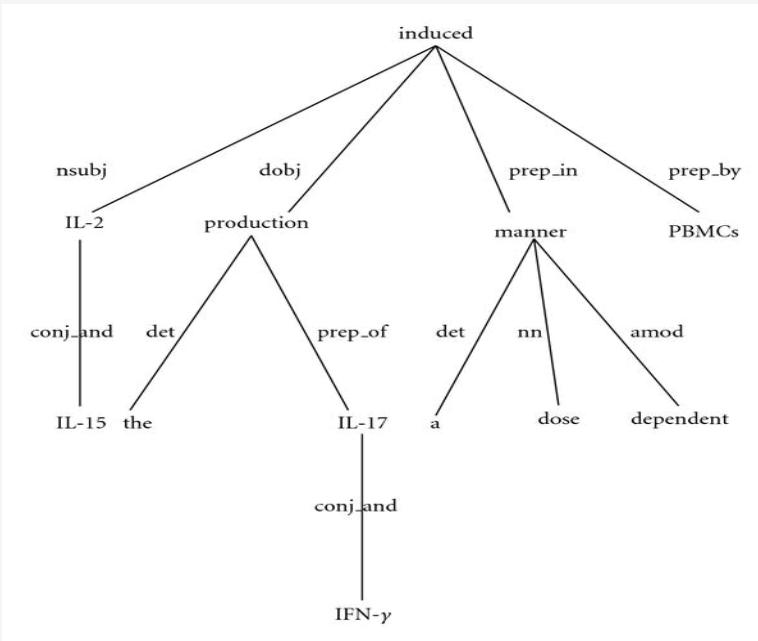
- This problem is pretty // easy
 - <http://www.naclo.cs.cmu.edu/problems2007/N2007-HS.pdf>
- Criteria
 - The part before // should be a complete sentence
 - The full sentence has a different meaning than the part before //
 - The part before // should not already be ambiguous

Dependency Parsing



Dependency Parsing

IL-2 and IL-15 induced the production of IL-17 and IFN- γ by PBMCS in a dose dependent manner.



Parser Output

nn(starts-2, Housing-1)
nsubj(rose-12, starts-2)
det(number-5, the-4)
appos(starts-2, number-5)
prep(number-5, of-6)
amod(homes-8, new-7)
pobj(of-6, homes-8)
auxpass(built-10, being-9)
partmod(homes-8, built-10)
ccomp(said-36, rose-12)
num(%-14, 7.2-13)
dobj(rose-12, %-14)
prep(rose-12, in-15)
pobj(in-15, March-16)
prep(rose-12, to-17)
det(rate-20, an-18)

amod(rate-20, annual-19)
pobj(to-17, rate-20)
prep(rate-20, of-21)
num(units-23, 549,000-22)
pobj(of-21, units-23)
advmod(rose-12, up-25)
dep(up-25, from-26)
det(512,000-29, a-27)
amod(512,000-29, revised-28)
pobj(from-26, 512,000-29)
prep(512,000-29, in-30)
pobj(in-30, February-31)
det(Department-35, the-33)
nn(Department-35, Commerce-34)
nsubj(said-36, Department-35)

NLP

NLP

Introduction to NLP

NLP Tasks (cont'd)

Information Extraction

- RESEARCH ALERT—Wells Fargo cuts PPD Inc to market perform
- China Southern Air Upgraded To Overweight From Neutral—HSBC
- CITIGROUP RAISES INGERSOLL RAND <IR.N> TO HOLD FROM SELL
- TCF Financial Corp Raised To Overweight From Neutral By JPMorgan
- BAIRD CUTS KIOR INC <KIOR.O> TO UNDERPERFORM RATING
- BRIEF—RESEARCH ALERT—Global Equities Research cuts LinkedIn to equal weight

Information Extraction

DATE/ TIME	TICKER	COMPANY	SOURCE	OLD	NEW	CHANGE
		PPD Inc	Wells Fargo		market perform	⬇️
		China Southern Air	HSBC	Neutral	Overweight	⬆️
	IR.N	INGERSOLL RAND	CITIGROUP	SELL	HOLD	⬆️
		TCF Financial Corp	JPMorgan	Neutral	Overweight	⬆️
	KIOR.O	KIOR INC	BAIRD		UNDERPERFORM	⬇️
		LinkedIn	Global Equities Research		equal weight	⬇️

False Positives

- Examples of false positives
 - BARCLAYS CUTS FLAGSTONE REINSURANCE <FSR.N> PRICE TARGET TO \$9 FROM \$11
 - Rimage To Buy Qumu For \$52M;; Raises Dividend;; Lowers EPS View
 - S&P rates Ameren Illinois commercial paper 'A-3'
 - BRIEF-Moody's changes otlk for Stirling Water Seafield Finance to positive
 - BRIEF-RESEARCH ALERT-HSBC cuts price targets on European telcos
 - Stifel cuts Philip Morris price target
 - Media General shares plummet on Moody's downgrade
- Explain why these are false positives.

Answers to the Quiz

- BARCLAYS CUTS FLAGSTONE REINSURANCE <FSR.N> PRICE TARGET TO \$9 FROM \$11
 - Didn't cut the ratings but the price target
- Rimage To Buy Qumu For \$52M;; Raises Dividend;; Lowers EPS View
 - Lowers eps view
- S&P rates Ameren Illinois commercial paper 'A-3'
 - Debt rating
- BRIEF–Moody's changes otk for Stirling Water Seafield Finance to positive
 - Changes outlook
- BRIEF–RESEARCH ALERT–HSBC cuts price targets on European telcos
 - Not a company but a group of companies
- Stifel cuts Philip Morris price target
 - Price target, not rating
- Media General shares plummet on Moody's downgrade
 - Event in the past

Semantics

- First order logic
- Inference
- Semantic analysis

$$\forall x,y: \text{Mother}(x,y) \Rightarrow \text{Parent}(x,y)$$

NACLO Problem

- “Bertrand and Russell”, 2014 problem by Ben King
 - <http://www.naclo.cs.cmu.edu/problems2014/N2014-H.pdf>

NACLO Solution

- Bertrand and Russell
 - <http://www.naclo.cs.cmu.edu/problems2014/N2014-HS.pdf>

Reading Comprehension

Mars Polar Lander - Where Are You?

(January 18, 2000) After more than a month of searching for a signal from NASA's Mars Polar Lander, mission controllers have lost hope of finding it. The Mars Polar Lander was on a mission to Mars to study its atmosphere and search for water, something that could help scientists determine whether life ever existed on Mars. Polar Lander was to have touched down December 3 for a 90-day mission. It was to land near Mars' south pole. The lander was last heard from minutes before beginning its descent. *The last effort to communicate with the three-legged lander ended with frustration at 8 a.m Monday.* "We didn't see anything," said Richard Cook, the spacecraft's project manager at NASA's Jet Propulsion Laboratory. The failed mission to the Red Planet cost the American government more than \$200 million dollars. Now, space agency scientists and engineers will try to find out what could have gone wrong. They do not want to make the same mistakes in the next mission.

(sources: CBC "For Kids" web page, Associated Press, CBC News Online, CBC Radio news, NASA)

1. When did the mission controllers lose hope of communicating with the lander?
Answer: *8AM, Monday Jan. 17, 2000*
2. Who is the Polar Lander's project manager?
3. Where on Mars was the spacecraft supposed to touch down?
4. What did the Mars Global Surveyor do?
5. What was the mission of the Mars Polar Lander?

Text Understanding

There are four bungalows in our cul-de-sac. They are made from these materials: straw, wood, brick and glass.

Mrs. Scott's bungalow is somewhere to the left of the wooden one and the third one along is brick. Mrs. Umbrella owns a straw bungalow and Mr. Tinsley does not live at either end, but lives somewhere to the right of the glass bungalow. Mr. Wilshaw lives in the fourth bungalow, whilst the first bungalow is not made from straw.

Who lives where, and what is their bungalow made from?

<http://www.brainbashers.com/showpuzzles.asp?puzzle=ZSOP>

Word Sense Disambiguation

- “The thieves took off with 100 gold **bars**”.
 - Did they steal 100 drinking establishments?
 - Or 100 measures of a song?

Word Sense Disambiguation

Bar=Noun

- S: (n) barroom, bar, saloon, ginmill, taproom (a room or establishment where alcoholic drinks are served over a counter) "he drowned his sorrows in whiskey at the bar"
- S: (n) bar (a counter where you can obtain food or drink) "he bought a hot dog and a coke at the bar"
- S: (n) bar (a rigid piece of metal or wood; usually used as a fastening or obstruction or weapon) "there were bars in the windows to prevent escape"
- S: (n) measure, bar (musical notation for a repeating pattern of musical beats) "the orchestra omitted the last twelve bars of the song"
- S: (n) bar (an obstruction (usually metal) placed at the top of a goal) "it was an excellent kick but the ball hit the bar"
- S: (n) prevention, bar (the act of preventing) "there was no bar against leaving"; "money was allocated to study the cause and prevention of influenza"
- S: (n) bar ((meteorology) a unit of pressure equal to a million dynes per square centimeter) "unfortunately some writers have used bar for one dyne per square centimeter"
- S: (n) bar (a submerged (or partly submerged) ridge in a river or along a shore) "the boat ran aground on a submerged bar in the river"
- S: (n) legal profession, bar, legal community (the body of individuals qualified to practice law in a particular jurisdiction) "he was admitted to the bar in New Jersey"
- S: (n) stripe, streak, bar (a narrow marking of a different color or texture from the background) "a green toad with small black stripes or bars"; "may the Stars and Stripes forever wave"
- S: (n) cake, bar (a block of solid substance (such as soap or wax)) "a bar of chocolate"
- S: (n) Browning automatic rifle, BAR (a portable .30 caliber automatic rifle operated by gas pressure and fed by cartridges from a magazine; used by United States troops in World War I and in World War II and in the Korean War)
- S: (n) bar (a horizontal rod that serves as a support for gymnasts as they perform exercises)
- S: (n) bar (a heating element in an electric fire) "an electric fire with three bars"
- S: (n) bar ((law) a railing that encloses the part of the courtroom where the judges and lawyers sit and the case is tried) "spectators were not allowed past the bar"

Bar=Verb

- S: (v) bar, debar, exclude (prevent from entering; keep out) "He was barred from membership in the club"
- S: (v) barricade, block, blockade, stop, block off, block up, bar (render unsuitable for passage) "block the way"; "barricade the streets"; "stop the busy road"
- S: (v) banish, relegate, bar (expel, as if by official decree) "he was banished from his own country"
- S: (v) bar (secure with, or as if with, bars) "He barred the door"

WSD is Important for Translation

- Paul plays soccer
 - Paul joue **au** football
- Paul plays the guitar
 - Paul joue **de la** guitare
- “wall” in German
 - die Chinesische **Mauer** (The Great Wall of China)
 - (otherwise Wand)
- “wall” in Spanish
 - pared, muro, muralla

Named Entity Recognition

Wolff, currently a journalist in Argentina, played with Del Bosque in the final years of the seventies in Real Madrid.

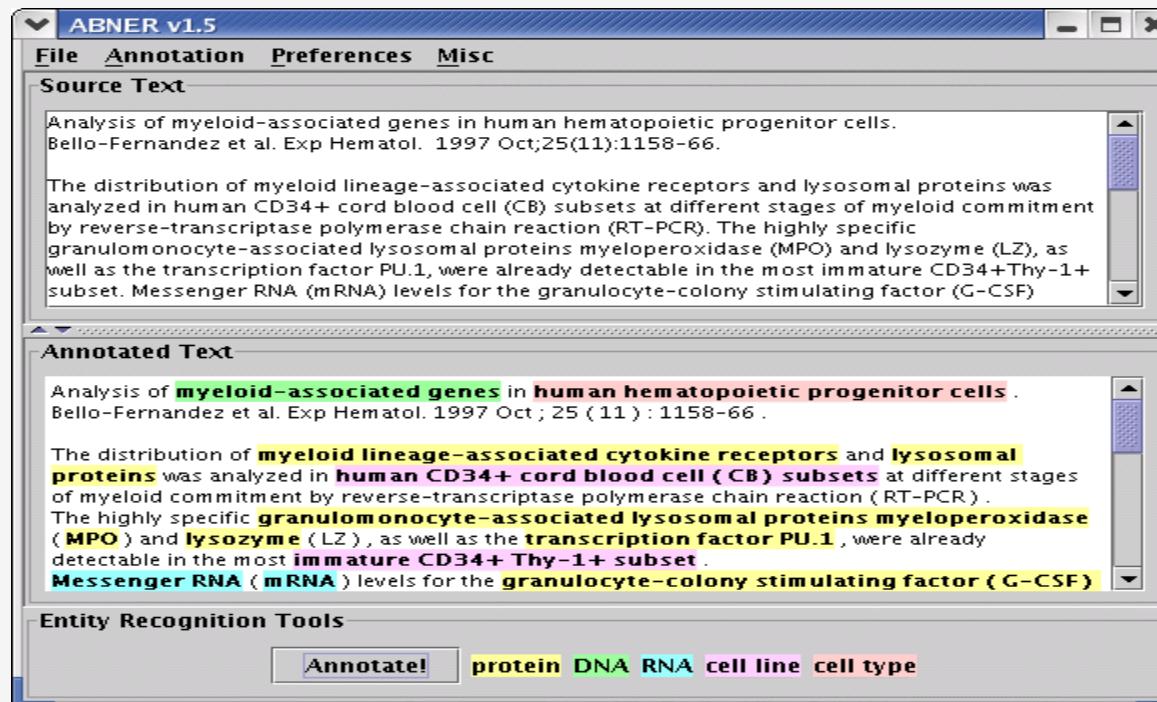
Wolff B-PER
,

currently O
a O
journalist O
in O
Argentina B-LOC
,

played O
with O
Del B-PER
Bosque I-PER
in O
the O
final O
years O
of O
the O
seventies O
in O
Real B-ORG
Madrid I-ORG
. O

- http://cogcomp.cs.illinois.edu/page/demo_view/NER
- <http://nlp.stanford.edu:8080/ner/>

Named Entity Recognition



<http://pages.cs.wisc.edu/~bsettles/abner>

Semantic Role Labeling

- [_{A0} He] [_{AM-MOD} would] [_{AM-NEG} n't] [_V accept] [_{A1} anything of value] from [_{A2} those he was writing about] .
- V: verb
A0: acceptor
A1: thing accepted
A2: accepted-from
A3: attribute
AM-MOD: modal
AM-NEG: negation
- http://cogcomp.cs.illinois.edu/page/demo_view/SRL

Coreference Resolution

- Barack Obama visited China. The US president met with his Chinese counterpart.
- Cynthia went to see her aunt at the hospital. She was scheduled for surgery on Monday.
- Because he was sick, Michael stayed home on Friday.

Ellipsis, Parallelism, and Underspecification

- Chen speaks Chinese. I don't.
- Santa gave Mary a book and Johnny a toy.

NLP

NLP

Introduction to NLP

NLP Tasks (cont'd)

Question Answering

- "The antagonist of Stevenson's Treasure Island." (Who is Long John Silver?)
- <http://blog.reddit.com/2011/02/ibm-watson-research-team-answers-your.html>
- "Watson is powered by 10 racks of IBM Power 750 servers running Linux, and uses 15 terabytes of RAM, 2,880 processor cores and is capable of operating at 80 teraflops. Watson was written in mostly Java but also significant chunks of code are written C++ and Prolog, all components are deployed and integrated using UIMA."

Jeopardy Questions

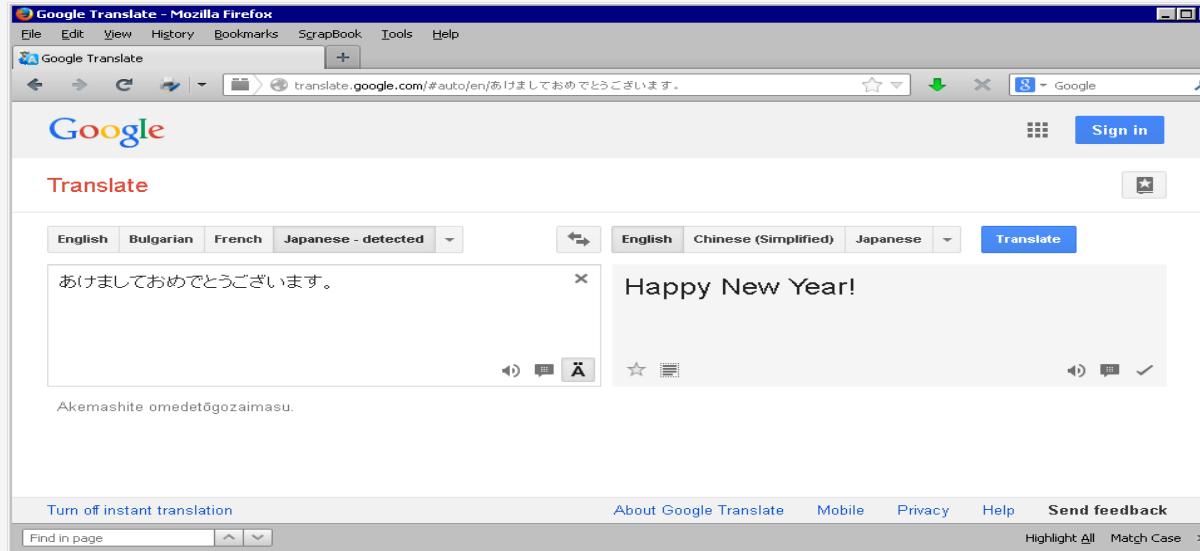
- From the competition between the IBM Watson system and two human champions (Ken Jennings and Brad Rutter)
- Sample questions:
 - On December 8, 2008 this national newspaper raised its newsstand price by 25 cents to \$1 : *USA Today*
 - In 2010 this former first lady published the memoir "Spoken From the Heart" : *Laura Bush**
 - This person is appointed by a testator to carry out the directions & requests in his will : *Executor**
 - Familiarity is said to breed this, from the Latin for "Despise" : *Contempt**
 - As of 2010, Croatia & Macedonia are candidates but this is the only former Yugoslav republic in the EU : *Slovenia*
 - The ancient "Lion of Nimrud" went missing from this city's national museum in 2003 (along with a lot of other stuff) : *Baghdad*
 - It's just a bloody nose! You don't have this hereditary disorder once endemic to European royalty : *Haemophilia*
 - It's Michelangelo's fresco on the wall of the Sistine Chapel, Depicting the saved and the damned : *The Last Judgement*
 - She "Died in the church and was buried along with her name. Nobody came" : *Eleanor Rigby*
 - It's a 4-letter term for a summit; the first 3 letters mean a type of simian : *Apex*
 - A camel is a horse designed by this : *Committee*
- Watson's answers: 66 correct and 9 incorrect (e.g., the one in the category "US Cities" about a city with two airports named after a World War II hero and a World War II battle)
- Watson's two day winning streak was \$77,147. Ken Jennings ended with \$24,000 and Brad Rutter with \$21,600. Watson donated \$500,000 to both World Vision and World Community Grid charities from the \$1,000,000 prize.
- <http://www.quora.com/What-questions-were-asked-in-the-Jeopardy-episode-involving-Watson>

Sentiment Analysis

- “I like the camera because I can edit images so easily, exactly as I do my iPad. I have found that its difficult to frame a picture when there isn't a zoom function as with the iPad. With this camera I can adjust my images by cropping as I did with my iPad but better yet, this camera has a built in zoom. A stretch or pinch of the fingers bring in the subject closer or back out again. With this iPhone I can also, as I dido with my iPad, enhance, crop, rotate, red eye reduce, and set a range of tints. I am also quite impressed with the quality of the images. Pretty darn good especially better than I expected for low light situations where I can use the built-in flash! Quite frankly I was quite surprised with these built in features. I also hope too experiment with and learn what HDR photography is. It's built into this iPhone and can be activated by a the touch of an icon. ”
- [http://www.epinions.com/review/
apple_iphone_5c_latest_model_16gb_graphite_unlocked_smartphone/
content_640679317124](http://www.epinions.com/review/apple_iphone_5c_latest_model_16gb_graphite_unlocked_smartphone/content_640679317124)

Machine Translation

- あけましておめでとうございます。
- Happy New Year!



Machine Translation

- Moses
- www.statmt.org

Moses Machine Translation Demo

Source:

Le gouvernement syrien a approuvé mardi un projet de loi permettant la levée de l'état d'urgence, en vigueur dans le pays depuis 1963, annonce l'agence officielle Sana.

Show Debug Output Show Alignment

Looking to translate a web page? Then click [here](#)

Translation:

The Syrian government has approved Tuesday a bill allowing the lifting of the state of emergency, in force in the country since 1963, the official Sana news agency.

Help to improve statistical machine translation!

The Syrian government has approved Tuesday a bill allowing the lifting of the state of emergency, in force in the country since 1963, the official Sana news agency.

Elephants are social animals. They live with their families, give hugs and call each other by using their trunks as trumpets. They also might know how to help each other.

In a recent elephant study by researchers from the United States and Thailand, pairs of giant animals learned to work together to get some ears of corn. Other animals, especially some primates, are already known to work together to complete tasks, but now elephants have joined the club. Perhaps the finding is not too surprising: Scientists suspect that elephants, with their big brains and survival savvy, may be among the smartest animals on the planet.

Joshua Plotnik, who worked on the study, told Science News that the animals didn't just learn a trick. Instead, the ways the elephants behaved show that they understand how working together brings benefits to everyone involved. Plotnik is a comparative psychologist now at the University of Cambridge in England. Psychology is the study of behaviors and mental processes, and comparative psychologists study how animals other than humans behave.

Les éléphants sont des animaux sociaux. Ils vivent avec leur famille, faire des câlins et appeler les uns les autres en utilisant leurs troncs trompettes. Ils pourraient également savoir comment aider les uns les autres.

Dans une étude récente d'éléphants par des chercheurs des États-Unis et la Thaïlande, des paires d'animaux géants ont appris à travailler ensemble pour obtenir des épis de maïs. D'autres animaux, en particulier des primates, sont déjà connus pour travailler ensemble pour accomplir des tâches, mais maintenant, les éléphants ont rejoint le club. Peut-être le résultat n'est pas trop surprenant: Les scientifiques soupçonnent que les éléphants, avec leurs gros cerveaux et de bon sens de survie, peut-être parmi les plus intelligents des animaux sur la planète.

Joshua Plotnick, qui a travaillé sur l'étude, dit Nouvelles de la Science que les animaux n'ont pas seulement appris un truc. Au lieu de cela, les moyens les éléphants se comportent montrent qu'ils comprennent comment travailler ensemble apporte des avantages à toutes les personnes impliquées. Plotnik est un psychologue comparatif maintenant à l'Université de Cambridge en Angleterre. **La psychologie est l'étude des comportements et des processus mentaux**, et étude comparative des psychologues comment les animaux autres que les humains se comportent.

Elephants are social animals. They live with their families, give hugs and call each other by using their trunks as trumpets. They also might know how to help each other.

In a recent elephant study by researchers from the United States and Thailand, pairs of giant animals learned to work together to get some ears of corn. Other animals, especially some primates, are already known to work together to complete tasks, but now elephants have joined the club. Perhaps the finding is not too surprising: Scientists suspect that elephants, with their big brains and survival savvy, may be among the smartest animals on the planet.

Joshua Plotnik, who worked on the study, told Science News that the animals didn't just learn a trick. Instead, the ways the elephants behaved show that they understand how working together brings benefits to everyone involved. Plotnik is a comparative psychologist now at the University of Cambridge in England. Psychology is the study of behaviors and mental processes, and comparative psychologists study how animals other than humans behave.

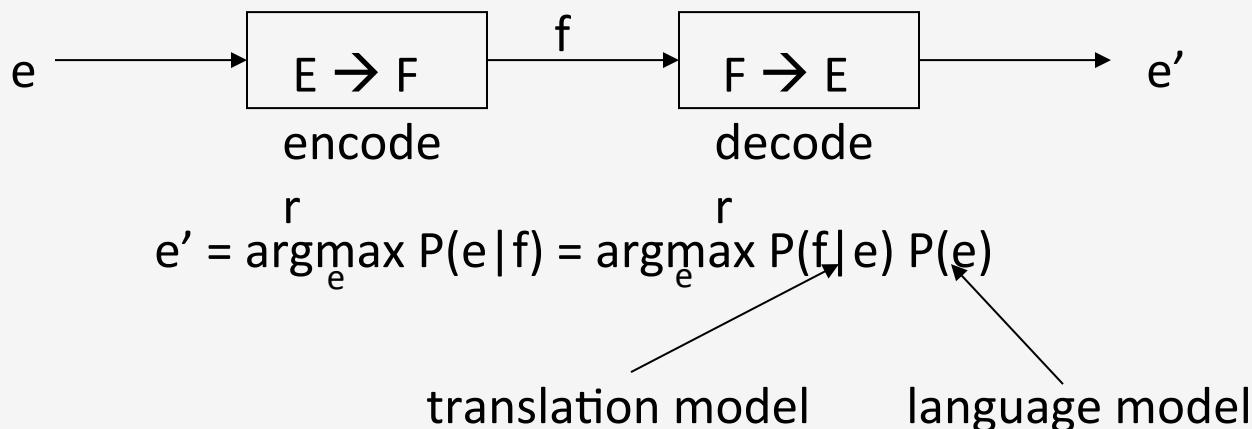
Les éléphants sont des animaux sociaux. Ils **vivent** avec leur famille, **faire** des câlins et **appeler** les uns les autres en utilisant leurs troncs trompettes. Ils pourraient également savoir comment aider les uns les autres.

Dans une étude récente d'éléphants par des chercheurs des États-Unis et la Thaïlande, des paires d'animaux géants ont appris à travailler ensemble pour obtenir des épis de maïs. D'autres animaux, en particulier des primates, sont déjà connus pour travailler ensemble pour accomplir des tâches, mais maintenant, les éléphants ont rejoint le club. Peut-être le résultat n'est pas trop surprenant: Les scientifiques soupçonnent que **les éléphants**, avec leurs gros cerveaux et de bon sens de survie, **peut-être** parmi les plus intelligents des animaux sur la planète.

Joshua Plotnick, qui a travaillé sur l'étude, **dit Nouvelles de la Science** que les animaux n'ont pas seulement appris un truc. Au lieu de cela, les moyens les éléphants se comportent montrent qu'ils comprennent comment travailler ensemble apporte des avantages à toutes **les personnes** impliquées. Plotnik est un psychologue **comparative** maintenant à l'Université de Cambridge en Angleterre. La psychologie est l'étude des comportements et des processus mentaux, et **étude comparative des psychologues** comment les animaux autres que les humains se comportent.

Machine Translation

- Noisy channel model (“Chinese Whispers”)



Machine Translation

- IBM Method



Text Summarization

Health Benefits

- Eating a diet rich in vegetables and fruits as part of an overall healthy diet may reduce risk for heart disease, including heart attack and stroke.
- Eating a diet rich in some vegetables and fruits as part of an overall healthy diet may protect against certain types of cancers.
- Diets rich in foods containing fiber, such as some vegetables and fruits, may reduce the risk of heart disease, obesity, and type 2 diabetes.
- Eating vegetables and fruits rich in potassium as part of an overall healthy diet may lower blood pressure, and may also reduce the risk of developing kidney stones and help to decrease bone loss.
- Eating foods such as vegetables that are lower in calories per cup instead of some other higher-calorie food may be useful in helping to lower calorie intake.

Nutrients

- Most vegetables are naturally low in fat and calories. None have cholesterol. (Sauces or seasonings may add fat, calories, or cholesterol.)
- Vegetables are important sources of many nutrients, including potassium, dietary fiber, folate (folic acid), vitamin A, and vitamin C.
- Diets rich in potassium may help to maintain healthy blood pressure. Vegetable sources of potassium include sweet potatoes, white potatoes, white beans, tomato products (paste, sauce, and juice), beet greens, soybeans, lima beans, spinach, lentils, and kidney beans.
- Dietary fiber from vegetables, as part of an overall healthy diet, helps reduce blood cholesterol levels and may lower risk of heart disease. Fiber is important for proper bowel function. It helps reduce constipation and diverticulosis. Fiber-containing foods such as vegetables help provide a feeling of fullness with fewer calories.
- Folate (folic acid) helps the body form red blood cells. Women of childbearing age who may become pregnant should consume adequate folate from foods, and in addition 400 mcg of synthetic folic acid from fortified foods or supplements. This reduces the risk of neural tube defects, spina bifida, and anencephaly during fetal development.

Summary

Eating vegetables is healthy.

NewsInEssence: Web-based News Summarization - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Home Search Favorites History

Address Go

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

Interactive Multi-source News Summarization

Pressure grows on Bush to globalise Iraq effort

Bangkok Post Friday 1 August, 2003 - US sets mid-2004 target for Iraq elections, two soldiers killed. BBC NEWS World Middle East Bush under fire over Iraq. President George W Bush is coming under increasing pressure from his own Republican Party to disclose how much financing will be required to cover the costs of occupying and reconstructing Iraq.

[7 Articles from 6 Sources] [5 Summaries]

Recent NIE News Clusters (more)

- ['Ananova - Tensions high as cortège approaches Najaf'](#)
24 articles, 4 summaries: 09/02, 6:10 AM
- ['Bush Makes Push for Manufacturing Jobs'](#)
18 articles, 4 summaries: 09/02, 6:10 AM
- ['Israeli Strike Kills Hamas Member September 1, 2003 21:44:37'](#)
12 articles, 4 summaries: 09/02, 6:10 AM

NIE Headlines

Build your own cluster of articles.

NewsTroll from URL:
URL must be from CNN, Yahoo!, MSNBC, BBC, or USA Today.
<http://www.hinduonnet.com>

NewsTroll from query:
[jugar bremer bush iraq]

Advanced Options

NIE News Clusters (Archive)

- ['Ananova - Tensions high as cortège approaches Najaf'](#)
24 articles, 4 summaries: 09/02, 6:10 AM
- ['Bush Makes Push for Manufacturing Jobs'](#)
18 articles, 4 summaries: 09/02, 6:10 AM
- ['Israeli Strike Kills Hamas Member September 1, 2003 21:44:37'](#)
12 articles, 4 summaries: 09/02, 6:10 AM
- ['Taliban ambush two Afghan patrols'](#)
9 articles, 4 summaries: 09/02, 6:10 AM
- ['FOXNews.com'](#)
7 articles, 4 summaries: 09/02, 6:10 AM
- ['BBC SPORT Cricket Cricket to return to Kashmir'](#)
7 articles, 4 summaries: 09/02, 6:10 AM
- ['Vivendi mulling Bronfman, GE offers; could decide Tuesday'](#)

Pressure grows on Bush to globalise Iraq effort

produced on 09/02, 6:10 AM

2% Summary

Bangkok Post Friday 1 August, 2003 - US sets mid-2004 target for Iraq elections, two soldiers killed (1:1) BBC NEWS World Middle East Bush under fire over Iraq (2:1) President George W Bush is coming under increasing pressure from his own Republican Party to disclose how much financing will be required to cover the costs of occupying and reconstructing Iraq. (2:2) The Bush administration, which is already spending \$4bn a month of US taxpayers' money on the military costs in Iraq alone, looks to be gearing up for a big international appeal for contributions towards the cost of rebuilding Iraq. (2:10)

Bush is facing growing calls from within his own party and from men running for his job to bring more international troops into Iraq amid mounting US casualties and costs. (4:3) A peacekeeping battalion from Kazakhstan has been deployed

Text to Speech

The screenshot shows a web browser window for the AT&T Natural Voices Text-to-Speech Demo. The page features a large AT&T logo and navigation links for Home, Demo, FAQ, Publications, Contact, Wizard Software, and How-To-Buy. A banner at the top states, "TTS is just one of the projects at AT&T Labs – Research. Visit our public website." Below this, a message says, "The previous listening experiment is complete. Thanks to all who participated." The main area is divided into three steps:

- STEP 1**: Voice & Language: Crystal US English
- STEP 2**: Text: Cleared text — No translation — Length limit — Special characters. The text input field contains: "The president signed the bill."
- STEP 3**: Click: SPEAK or DOWNLOAD [restrictions apply**]

At the bottom, there are feedback and diagnostic links, a note about restrictions on use of audio, and legal notices. The browser's status bar at the bottom shows a search for "children".

<http://www2.research.att.com/~ttsweb/tts/demo.php>

Text to Speech

- www.ivona.com

Entailment and Paraphrasing

ID	TEXT	HYPOTHESIS	TASK	VALUE
1586	<i>The Republic of Yemen is an Arab, Islamic and independent sovereign state whose integrity is inviolable, and no part of which may be ceded.</i>	<i>The national language of Yemen is Arabic.</i>	QA	True
1076	<i>Most Americans are familiar with the Food Guide Pyramid—but a lot of people don't understand how to use it and the government claims that the proof is that two out of three Americans are fat.</i>	<i>Two out of three Americans are fat.</i>	RC	True
1667	<i>Regan attended a ceremony in Washington to commemorate the landings in Normandy.</i>	<i>Washington is located in Normandy.</i>	IE	False
2016	<i>Google files for its long awaited IPO.</i>	<i>Google goes public.</i>	IR	True
2097	<i>The economy created 228,000 new jobs after a disappointing 112,000 in June.</i>	<i>The economy created 228,000 jobs after disappointing the 112,000 of June.</i>	MT	False
893	<i>The first settlements on the site of Jakarta were established at the mouth of the Ciliwung, perhaps as early as the 5th century AD.</i>	<i>The first settlements on the site of Jakarta were established as early as the 5th century AD.</i>	CD	True
1960	<i>Bush returned to the White House late Saturday while his running mate was off campaigning in the West.</i>	<i>Bush left the White House.</i>	PP	False
586	<i>The two suspects belong to the 30th Street gang, which became embroiled in one of the most notorious recent crimes in Mexico: a shootout at the Guadalajara airport in May, 1993, that killed Cardinal Juan Jesus Posadas Ocampo and six others.</i>	<i>Cardinal Juan Jesus Posadas Ocampo died in 1993.</i>	QA	True

Table 1. Examples of Text-Hypothesis pairs

Discourse Analysis

- Anaphoric relations:

1. *Mary helped Peter get out of the car. He thanked her.*

2. *Mary helped the other passenger out of the car.*

The man had asked her for help because of his foot injury.

Tom appeared on the sidewalk with a bucket of whitewash and a long-handled brush. He surveyed the fence, and all gladness left him and a deep melancholy settled down upon his spirit. (Tom Sawyer)

Dialogue Systems

- I would like to make a reservation at Sorrento.
- For when?
- 8 pm Friday night.
- We only have availability for 7 pm and 10 pm.
- Sorry, these don't work for me.

Other Applications

- Spelling Correction
- Web search
- Natural language interfaces to databases
- Parsing job postings
- Summarizing medical records
- Information extraction for databases
- Social network extraction from text
- Alignment of text w/ other signal (time series)
- Essay grading
- Generating weather reports, sports reports, and news stories

NLP

NLP

Introduction to NLP

Syntax

Syntax

- Is language more than just a “bag of words”?
- Grammatical rules apply to categories and groups of words, not individual words.
- Example – a sentence includes a subject and a predicate. The subject is a noun phrase and the predicate is a verb phrase.
 - Noun phrase: The cat, Samantha, She
 - Verb phrase: arrived, went away, had dinner
- When people learn a new word, they learn its syntactic usage.
 - Examples: wug (n), cluvious (adj) – use them in sentences
 - Hard to come up with made up words: forkle, vleer, etc. all taken.

Defining Parts of Speech

- What do nouns typically have in common?
 - E.g., *can* be preceded by “the”.
- Verbs can be preceded by “can’t”.
- Adjectives can come between “the” and a noun.
- How is this different from grade school definitions?
- Determiners: a, the, many, no, five
- Prepositions: for, to, in, without, before

The Lexicon

- How do we think of words like cat, run, five?
 - pronunciation, part of speech, meaning
- Five: /faɪv/, numeral, “5”
- Ambiguity

Constituents

- Constituents are continuous
- Constituents are non-crossing
 - if two constituents share one word, then one of them must completely contain the other.
- Each word is a constituent

Constituent Tests

- “coordination” test
- “pronoun” test
 - A small dog is barking in the park.
 - It is barking in the park
- “question by repetition” test:
 - I have seen blue elephants
 - Blue elephants?
 - * Seen blue?
 - Seen blue elephants?
- “topicalization” test:
 - Blue elephants, I have seen.
- “question” test:
 - *What* have I seen?
- “deletion” test
 - Last year I saw a blue elephant in the zoo.
- “semantic” test
- “intuition” test

How To Generate Sentences

- One way: tree structure
 - Generate the tree structure first
 - Then fill the leaf nodes with terminals

A Simple Syntactic Rule

- The simplest rule for a sentence, e.g. “Birds fly”

$$S \rightarrow N \ V$$

Simplest Grammar

S → N V

N → Samantha | Min | Jorge

V → left | sang | walked

Sample sentences:

Samantha sang

Jorge left

Syntax

- The verbs so far were intransitive (no direct object)
- What rules are needed next?
 - Transitive verbs and direct objects (“Jorge saw Samantha”)
 - Determiners (“the cats”)
- Combinatorial explosion (even for the simplest form of sentences)
- Need for noun phrases
- Ditto for verb phrases

Latest Grammar

S → NP VP

NP → DT N

VP → V NP

DT → the | a

N → child | cat | dog

V → took | saw | liked | scared | chased

Sample sentences:

a dog chased the cat

the child saw a dog

Alternatives

- Different expansions of a category are delineated with “|”
 - $\text{NP} \rightarrow \text{PN} \mid \text{DT CN}$
- One rule for proper nouns and another for common nouns

Latest Grammar

S → NP VP

NP → DT CN

NP → PN

VP → V NP

DT → the | a

CN → child | cat | dog

PN → Samantha | Jorge | Min

V → took | saw | liked | scared | chased

Sample sentences:

a child scared Jorge

Min took the child

Optional Categories

- Wherever N is allowed in a sentence,
 - $DT\ N$
 - $JJ\ N$
 - $DT\ JJ\ N$are also allowed
- We can use the notation for alternatives
 - $NP \rightarrow N \mid DT\ N \mid JJ\ N \mid DT\ JJ\ N$
- Optional categories can be also marked using parentheses:
 - $NP \rightarrow (DT) (JJ) N$

Verb Phrases

- Samantha ran.
- Samantha ran to the park.
- Samantha ran away.
- Samantha bought a cookie.
- Samantha bought a cookie for John.
- Overall structure: $\text{VP} \rightarrow \text{V}(\text{NP})(\text{P})(\text{NP})$

Latest Grammar

S → NP VP

NP → DT CN

NP → PN

VP → V (NP) (P) (NP)

DT → the | a

CN → child | cat | dog

PN → Samantha | Jorge | Min

P → to | for | from | in

V → took | saw | liked | scared | chased | gave

Sample sentences:

Samantha saw the cat

Jorge gave the cat to Min

Prepositional Phrases

- Examples:
 - Mary bought a book for John **in a bookstore**.
 - The bookstore sells magazines.
 - The bookstore **on Main St.** sells magazines.
 - Mary ran away.
 - Mary ran **down the hill**.
- Changes are needed to both NP and VP to accommodate prepositional phrases
 - Wherever a preposition is allowed, it can be followed by a noun phrase.
 - Run up
 - NP can contain any number of PPs but only up to two NPs.
- How do we revise the grammar accordingly?

The Rules So Far

- $S \rightarrow NP\ VP$
- $NP \rightarrow (DT)\ (JJ)\ N\ (PP)$
- $VP \rightarrow V\ (NP)\ (PP)$
- $PP \rightarrow P\ (NP)$

PP Ambiguity

- The boy saw the woman with the telescope.

PP → PREP NP

VP → V NP PP

VP → V NP

NP → DT N

NP → DT N PP

Repetition (*)

- (JJ^*) = a sequence of zero or more JJ
- Are all sequences of adjectives allowed?
 - a big red house
 - * a red big house
- Adjective ordering in English depends on semantics!

Exercise

- The Little Red Riding Hood
- Three Little Pigs
- The Three Musketeers
- The Steadfast Tin Soldier
- The French Connection
- Old Macdonald
- Five Golden Rings
- The Ancient Mariner

Adjective Ordering

- **Det**
 - Number
 - Strength
 - Size
 - Age
 - Shape
 - Color
 - Origin
 - Material
 - Purpose
 - **Noun**
-
- det < number < size < color < purpose < noun
 - strength < material < noun
 - origin < noun

Nested Sentences

- Examples:
 - I don't recall whether I took the dog out.
 - Do you know if the mall is still open?
- $VP \rightarrow V (NP) (NP) (C S) (PP^*)$
- Can (C S) appear inside an NP?
 - Whether he will win the elections remains to be seen.

Recursion

- S can generate VP, VP can generate S
- NP can generate PP, PP can generate NP
- What does recursion allow?
- Is there a longest sentence in English?
- Conjunction of NPs:

$\text{NP} \rightarrow \text{NP} \text{ and } \text{NP}$

- Conjunction of PPs:

$\text{PP} \rightarrow \text{PP} \text{ and } \text{PP}$

- Conjunction of VPs:

$\text{VP} \rightarrow \text{VP} \text{ and } \text{VP}$

Meta-patterns

- $S \rightarrow NP\ VP$
 - $NP \rightarrow (DT)\ (JJ)\ N\ (PP)$
 - $VP \rightarrow V\ (NP)\ (PP)$
 - $PP \rightarrow P\ (NP)$
- Is there a meta-pattern here?
 - $XP \rightarrow (\text{specifier})\ X'$
 - $X' \rightarrow X\ (\text{complement})$
- Example: $NP \rightarrow DT\ N'$
- X-bar Theory
 - [http://www.unlweb.net/wiki/X-bar theory](http://www.unlweb.net/wiki/X-bar_theory)

Meta-rules for Conjunctions

- Conjunction
 - $X \rightarrow X \text{ and } X$
- This kind of rule even covers entire sentences
 - $S \rightarrow S \text{ and } S$

Auxiliaries

- Is “Aux V” a constituent?
 - I have seen blue elephants and will remember them forever.
- Recursion:
 - VP → Aux VP
 - Raj may have been sleeping.
- Is such recursion unlimited?

Exercise

- Grammar:
 - $S \rightarrow NP\ VP \mid CP\ VP$
 - $NP \rightarrow (DT)\ (JJ^*)\ N\ (CP)\ (PP^*)$
 - $VP \rightarrow V\ (NP)\ (NP)\ (PP^*) \mid V\ (NP)\ (CP)\ (PP^*)$
 - $PP \rightarrow P\ NP$
 - $CP \rightarrow C\ S$
- What rules are needed to generate these three sentences:
 - 1. The small dog of the neighbors brought me an old tennis ball.
 - 2. That wugs have three eyes is unproven by scientists.
 - 3. I saw the gift that the old man gave me at the meeting.

NLP

NLP

Introduction to NLP

Introduction to Parsing

Parsing Programming Languages

```
#include <stdio.h>

int main()
{
    int n, reverse = 0;

    printf("Enter a number to reverse\n");
    scanf("%d", &n);

    while (n != 0)
    {
        reverse = reverse * 10;
        reverse = reverse + n%10;
        n = n/10;
    }
    printf("Reverse of entered number is = %d\n", reverse);

    return 0;
}
```

Parsing Human Languages

- Rather different than computer languages
 - Can you think in which ways?

Parsing Human Languages

- Rather different than computer languages
 - No types for words
 - No brackets around phrases
 - Ambiguity
 - Words
 - Parses
 - Implied information

The Parsing Problem

- Parsing means associating tree structures to a sentence, given a grammar (often a CFG)
 - There may be exactly one such tree structure
 - There may be many such structures
 - There may be none
- Grammars (e.g., CFG) are declarative
 - They don't specify how the parse tree will be constructed

Syntactic Ambiguities

- PP attachment
 - I saw the man with the telescope
- Gaps
 - Mary likes Physics but hates Chemistry
- Coordination scope
 - Small boys and girls are playing
- Particles vs. prepositions
 - She ran up a large bill
- Gerund vs. adjective
 - Frightening kids can cause trouble

Applications Of Parsing

- Grammar checking
 - I want to return this shoes.
- Question answering
 - How many people in sales make \$40K or more per year?
- Machine translation
 - E.g., word order – SVO vs. SOV
- Information extraction
 - Breaking Bad takes place in New Mexico.
- Speech generation
- Speech understanding
- Interpretation

NLP

Introduction to NLP

Context-free grammars

Context-free Grammars

- A context-free grammar is a 4-tuple (N, Σ, R, S)
 - N : non-terminal symbols
 - Σ : terminal symbols (disjoint from N)
 - R : rules $(A \rightarrow \beta)$, where β is a string from $(\Sigma \cup N)^*$
 - S : start symbol from N

Example

```
["the", "child", "ate", "the", "cake", "with", "the", "fork"]
```

S → NP VP

NP → DT N | NP PP

PP → PRP NP

VP → V NP | VP PP

DT → 'a' | 'the'

N → 'child' | 'cake' | 'fork'

PRP → 'with' | 'to'

V → 'saw' | 'ate'

Example

["the", "child", "ate", "the", "cake", "with", "the", "fork"]

S → NP VP

NP → DT **N** | NP PP

PP → **PRP** NP

VP → **V** NP | VP PP

DT → 'a' | 'the'

N → 'child' | 'cake' | 'fork'

PRP → 'with' | 'to'

V → 'saw' | 'ate'

Heads marked in bold face

Phrase-structure Grammars (1/2)

- Sentences are not just bags of words
 - Alice bought Bob flowers
 - Bob bought Alice flowers
- Context-free view of language
 - A prepositional phrase looks the same whether it is part of the subject NP or part of the VP
- Constituent order
 - SVO (subject verb object)
 - SOV (subject object verb)

Phrase-structure Grammars (2/2)

- Auxiliary verbs
 - The dog may have eaten my homework
- Imperative sentences
 - Leave the book on the table
- Interrogative sentences
 - Did the customer have a complaint?
- Negative sentences
 - The customer didn't have a complaint

A Longer Example

```
S -> NP VP | Aux NP VP | VP
NP -> PRON | Det Nom
Nom -> N | Nom N | Nom PP
PP -> PRP NP
VP -> V | V NP | VP PP
Det -> 'the' | 'a' | 'this'
PRON -> 'he' | 'she'
N -> 'book' | 'boys' | 'girl'
PRP -> 'with' | 'in'
V -> 'takes' | 'take'
```

What changes were made to the grammar?

A Longer Example

```
S -> NP VP | Aux NP VP | VP
NP -> PRON | Det Nom
Nom -> N | Nom N | Nom PP
PP -> PRP NP
VP -> V | V NP | VP PP
Det -> 'the' | 'a' | 'this'
PRON -> 'he' | 'she'
N -> 'book' | 'boys' | 'girl'
PRP -> 'with' | 'in'
V -> 'takes' | 'take'
```

A Longer Example

S → NP VP | **Aux** **NP** **VP** | VP
NP → PRON | Det Nom
Nom → **N** | Nom N | Nom PP
PP → PRP NP
VP → V | V NP | VP PP
Det → 'the' | 'a' | 'this'
PRON → 'he' | 'she'
N → 'book' | 'boys' | 'girl'
PRP → 'with' | 'in'
V → 'takes' | 'take'

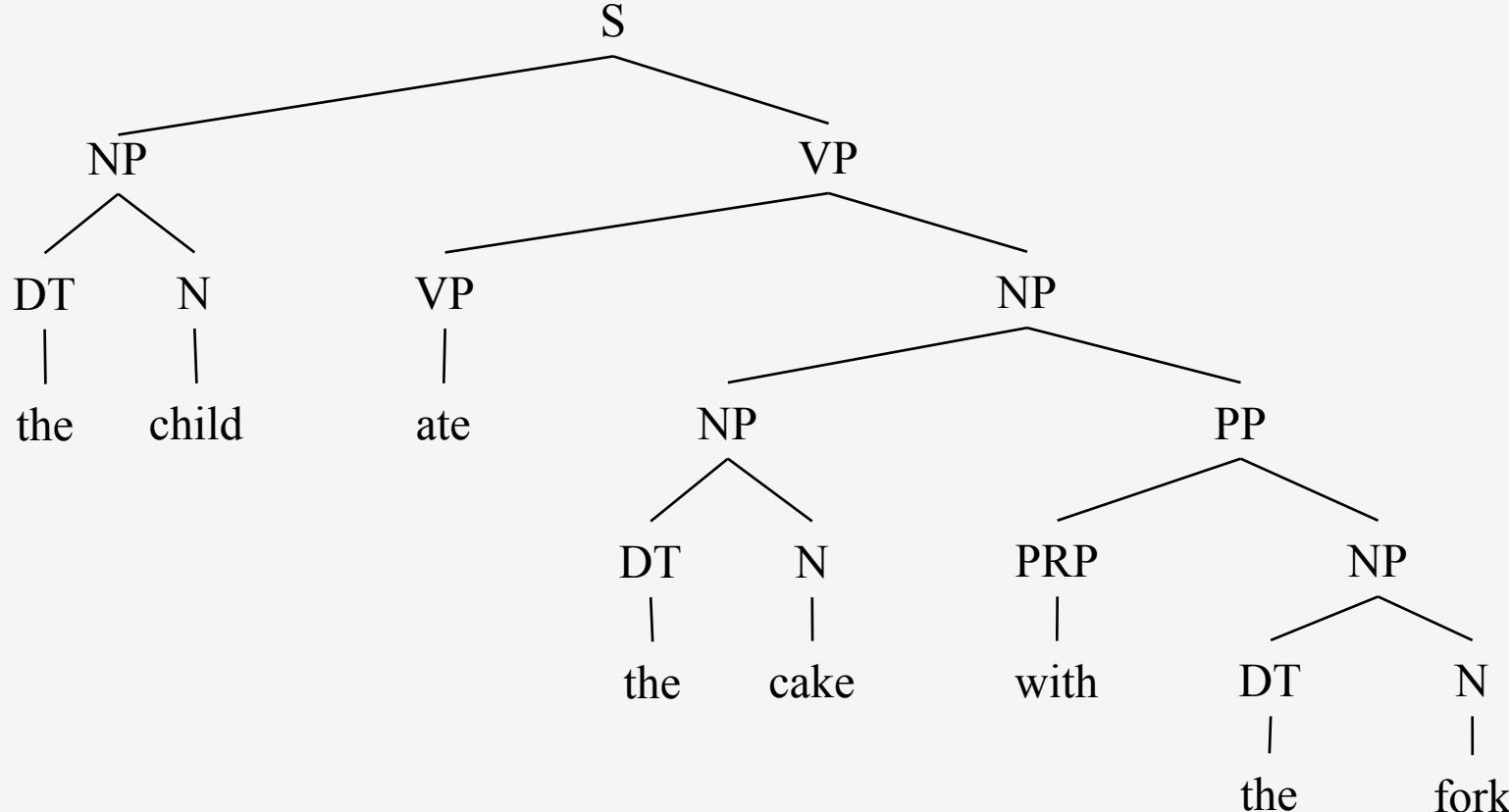
Penn Treebank Example

```
( (S
    (NP-SBJ
        (NP (NNP Pierre) (NNP Vinken) )
        (, ,)
        (ADJP
            (NP (CD 61) (NNS years) )
            (JJ old) )
        (, ,) )
    (VP (MD will)
        (VP (VB join)
            (NP (DT the) (NN board) )
            (PP-CLR (IN as)
                (NP (DT a) (JJ nonexecutive) (NN director) )))
            (NP-TMP (NNP Nov.) (CD 29) )))
        (. .) )))
( (S
    (NP-SBJ (NNP Mr.) (NNP Vinken) )
    (VP (VBZ is)
        (NP-PRD
            (NP (NN chairman) )
            (PP (IN of)
                (NP
                    (NP (NNP Elsevier) (NNP N.V.) )
                    (, ,)
                    (NP (DT the) (NNP Dutch) (VBG publishing) (NN group) )))))
        (. .) )))
```

Leftmost Derivation

- A leftmost derivation is a sequence of strings s_1, s_2, \dots, s_n
 - $s_1 = S$, the start symbol
 - s_n includes only terminal symbols
- Example:
 - [S]
 - [S] [NP VP]
 - [S] [NP VP] [DT N VP]
 - ...
 - [S] [NP VP] [DT N VP] ... [the child ate the cake with the fork]

Leftmost Derivation

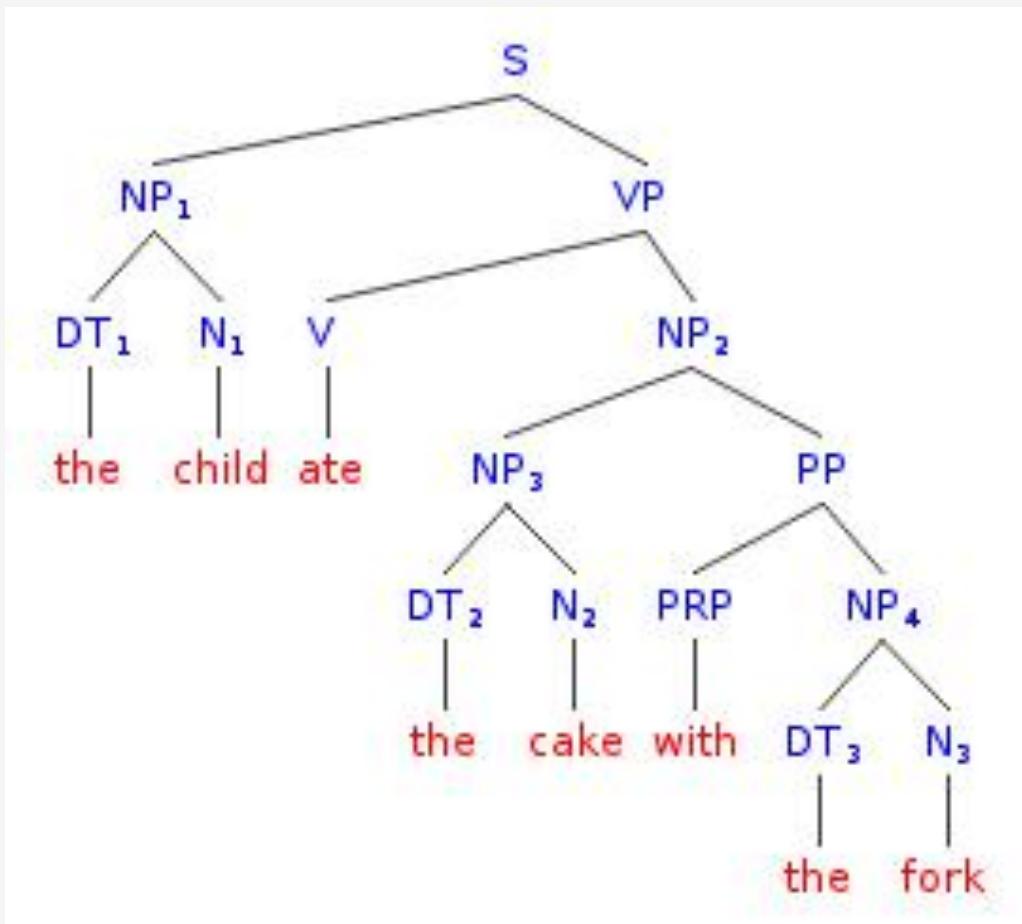


NLP

NLP

Introduction to NLP

Classic parsing methods



S → NP VP
NP → DT N | NP PP
PP → PRP NP
VP → V NP | VP PP
DT → 'a' | 'the'
N → 'child' | 'cake' | 'fork'
PRP → 'with' | 'to'
V → 'saw' | 'ate'

Parsing as Search

- There are two types of constraints on the parses
 - From the input sentence
 - From the grammar
- Therefore, two general approaches to parsing
 - Top-down
 - Bottom-up

Top Down Parsing

S

S → NP VP

NP → DT N | NP PP

PP → PRP NP

VP → V NP | VP PP

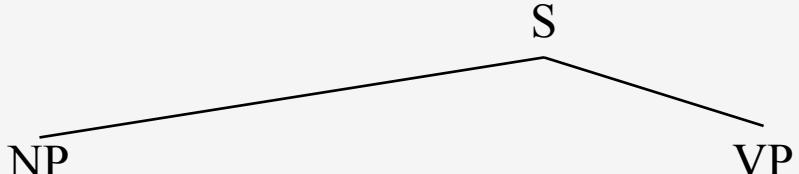
DT → 'a' | 'the'

N → 'child' | 'cake' | 'fork'

PRP → 'with' | 'to'

V → 'saw' | 'ate'

Top Down Parsing



S → NP VP
NP → DT N | NP PP
PP → PRP NP
VP → V NP | VP PP
DT → 'a' | 'the'
N → 'child' | 'cake' | 'fork'
PRP → 'with' | 'to'
V → 'saw' | 'ate'

Top Down Parsing



S → NP VP
NP → DT N | NP PP
PP → PRP NP
VP → V NP | VP PP
DT → 'a' | 'the'
N → 'child' | 'cake' | 'fork'
PRP → 'with' | 'to'
V → 'saw' | 'ate'

Top Down Parsing



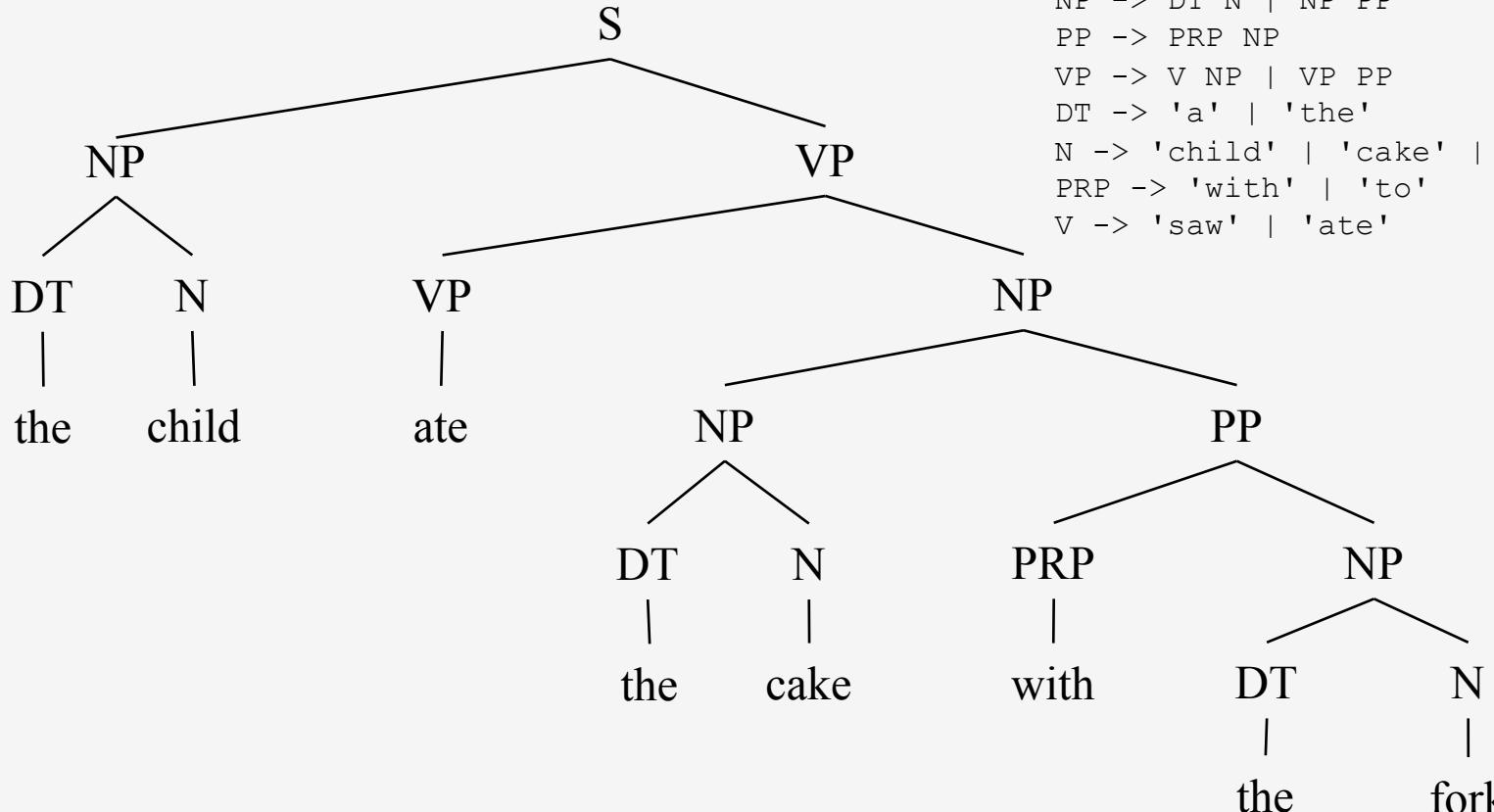
S → NP VP
NP → DT N | NP PP
PP → PRP NP
VP → V NP | VP PP
DT → 'a' | 'the'
N → 'child' | 'cake' | 'fork'
PRP → 'with' | 'to'
V → 'saw' | 'ate'

Top Down Parsing

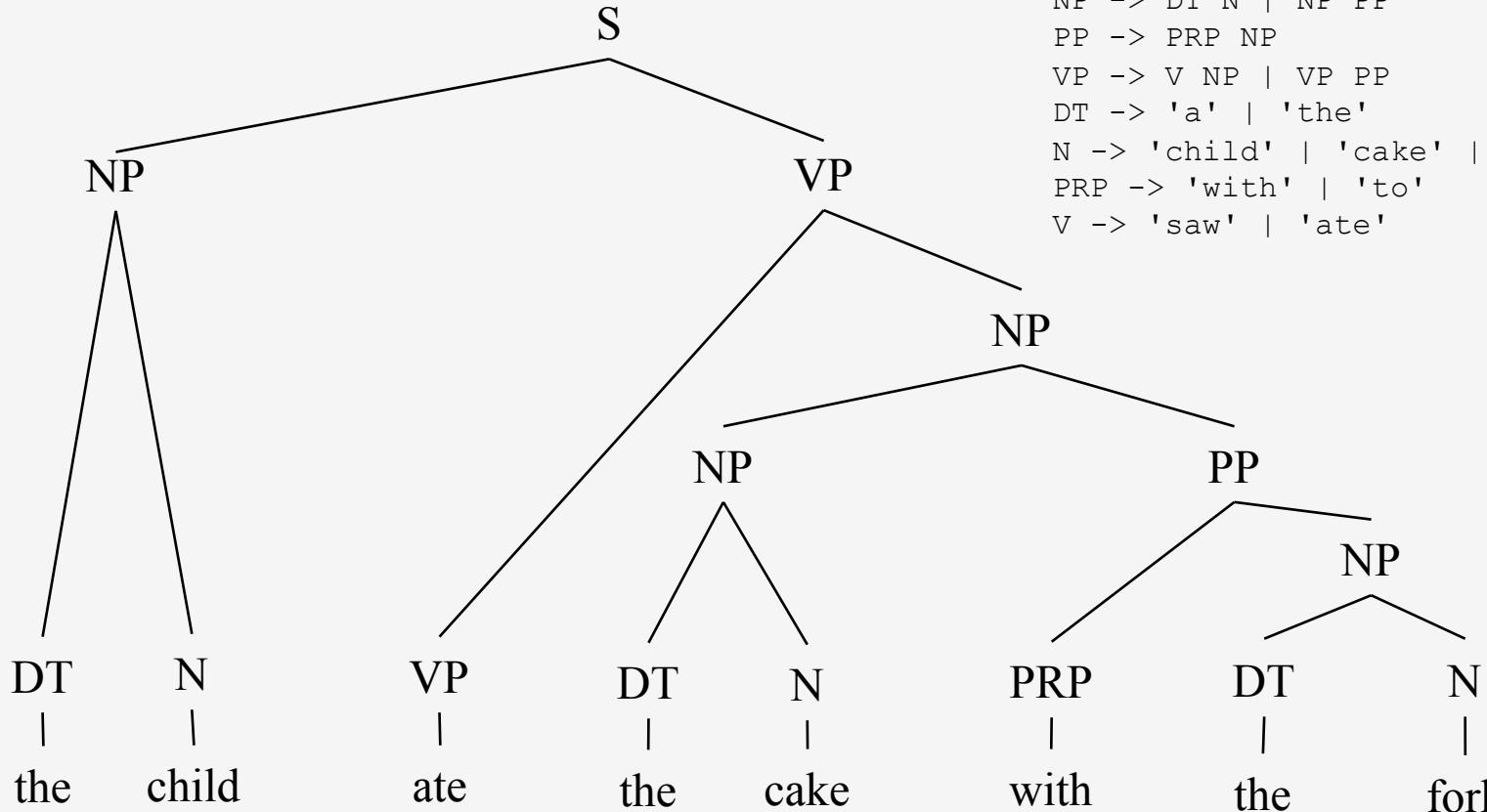


S → NP VP
NP → DT N | NP PP
PP → PRP NP
VP → V NP | VP PP
DT → 'a' | 'the'
N → 'child' | 'cake' | 'fork'
PRP → 'with' | 'to'
V → 'saw' | 'ate'

Top Down Parsing



Bottom Up Parsing



S → NP VP
NP → DT N | NP PP
PP → PRP NP
VP → V NP | VP PP
DT → 'a' | 'the'
N → 'child' | 'cake' | 'fork'
PRP → 'with' | 'to'
V → 'saw' | 'ate'

Bottom Up Vs. Top Down Methods

- Bottom up
 - explores options that won't lead to a full parse.
- Top down
 - explores options that don't match the full sentence.
- Dynamic programming
 - caches of intermediate results (memoization)
- Cocke–Kasami–Younger (CKY) parser
 - based on dynamic programming

Introduction to NLP

Shift-reduce parsing

Shift-reduce Parsing

- A bottom-up parser
 - Tries to match the RHS of a production until it can build an S
- *Shift operation*
 - Each word in the input sentence is pushed onto a stack
- *Reduce operation*
 - If the top n words on the top of the stack match the RHS of a production, then they are popped and replaced by the LHS of the production
- *Stopping condition*
 - The process stops when the input sentence has been processed and S has been popped from the stack.

Shift-reduce Parsing Example

[* the child ate the cake]

Shift-reduce Parsing Example

```
[ * the child ate the cake]  
S [ 'the' * child ate the cake]
```

Shift-reduce Parsing Example

```
[ * the child ate the cake]
S [ 'the' * child ate the cake]
R [ DT * child ate the cake]
```

Shift-reduce Parsing Example

```
[ * the child ate the cake]
S [ 'the' * child ate the cake]
R [ DT * child ate the cake]
S [ DT 'child' * ate the cake]
R [ DT N * ate the cake]
```

Shift-reduce Parsing Example

```
[ * the child ate the cake]
S [ 'the' * child ate the cake]
R [ DT * child ate the cake]
S [ DT 'child' * ate the cake]
R [ DT N * ate the cake]
R [ NP * ate the cake]
S [ NP 'ate' * the cake]
```

Shift-reduce Parsing Example

```
[ * the child ate the cake]
S [ 'the' * child ate the cake]
R [ DT * child ate the cake]
S [ DT 'child' * ate the cake]
R [ DT N * ate the cake]
R [ NP * ate the cake]
S [ NP 'ate' * the cake]
R [ NP V * the cake]
S [ NP V 'the' * cake]
R [ NP V DT * cake]
S [ NP V DT 'cake' * ]
```

Shift-reduce Parsing Example

```
[ * the child ate the cake]
S [ 'the' * child ate the cake]
R [ DT * child ate the cake]
S [ DT 'child' * ate the cake]
R [ DT N * ate the cake]
R [ NP * ate the cake]
S [ NP 'ate' * the cake]
R [ NP V * the cake]
S [ NP V 'the' * cake]
R [ NP V DT * cake]
S [ NP V DT 'cake' * ]
R [ NP V DT N * ]
R [ NP V NP * ]
R [ NP VP * ]
R [ S * ]
```

Shift-reduce Parsing Example

```
[ * the child ate the cake]
S [ 'the' * child ate the cake]
R [ DT * child ate the cake]
S [ DT 'child' * ate the cake]
R [ DT N * ate the cake]
R [ NP * ate the cake]
S [ NP 'ate' * the cake]
R [ NP V * the cake]
S [ NP V 'the' * cake]
R [ NP V DT * cake]
S [ NP V DT 'cake' * ]
R [ NP V DT N * ]
R [ NP V NP * ]
R [ NP VP * ]
R [ S * ]
```

(S (NP (DT the) (N child)) (VP (V ate) (NP (DT the) (N cake))))

NLP

Introduction to NLP

Cocke-Kasami-Younger (CKY) Parsing

Dynamic Programming

- Motivation
 - A lot of the work is repeated
 - Caching intermediate results improves the complexity
- Dynamic programming
 - Building a parse for a substring $[i,j]$ based on all parses $[i,k]$ and $[k, j]$ that are included in it.
- Complexity
 - $O(n^3)$ for recognizing an input string of length n

Dynamic Programming

- CKY (Cocke–Kasami–Younger)
 - bottom-up
 - requires a normalized (binarized) grammar
- Earley parser
 - top-down
 - more complicated

Example

```
["the", "child", "ate", "the", "cake", "with", "the", "fork"]
```

S → NP VP

NP → DT N | NP PP

PP → PRP NP

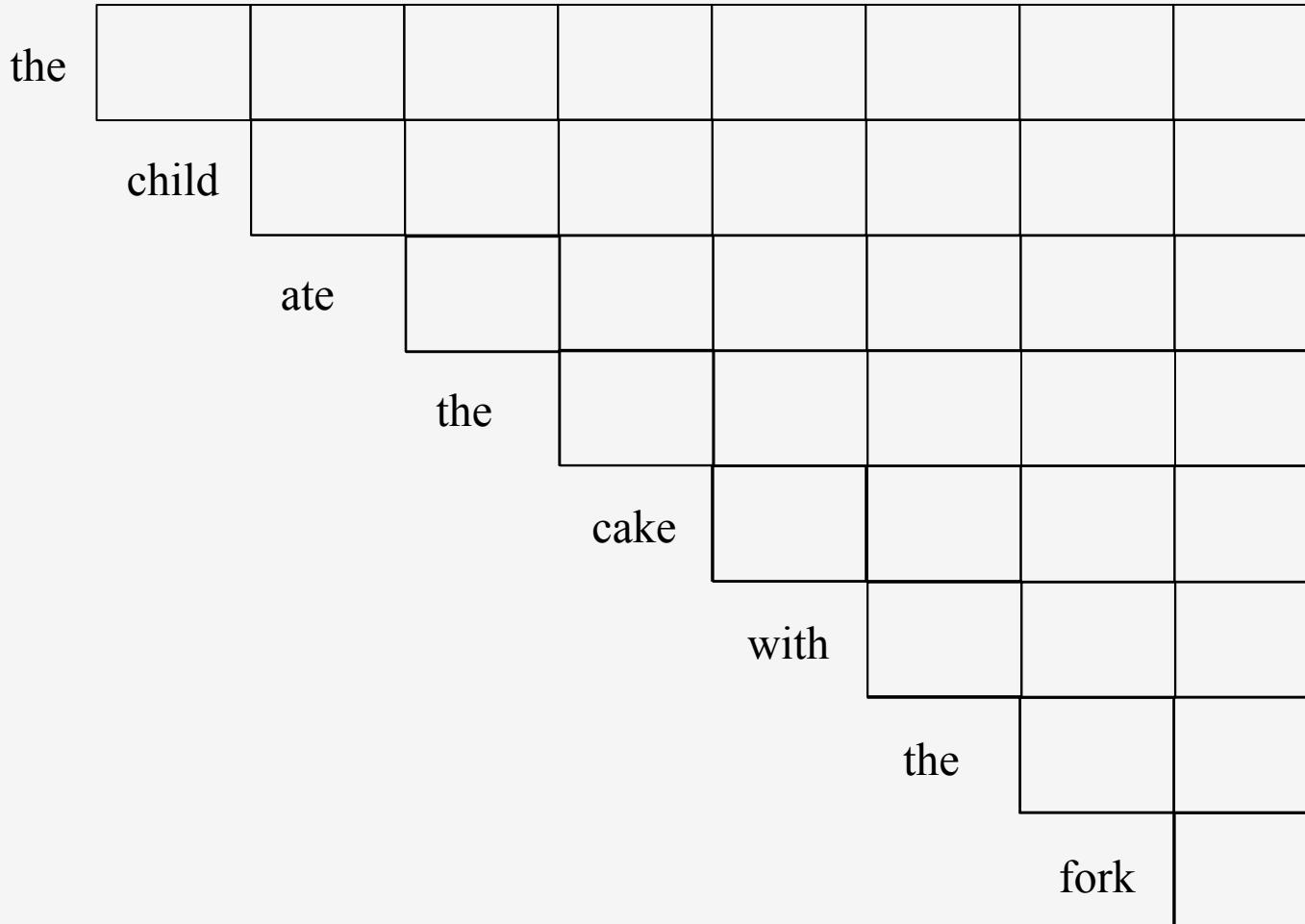
VP → V NP | VP PP

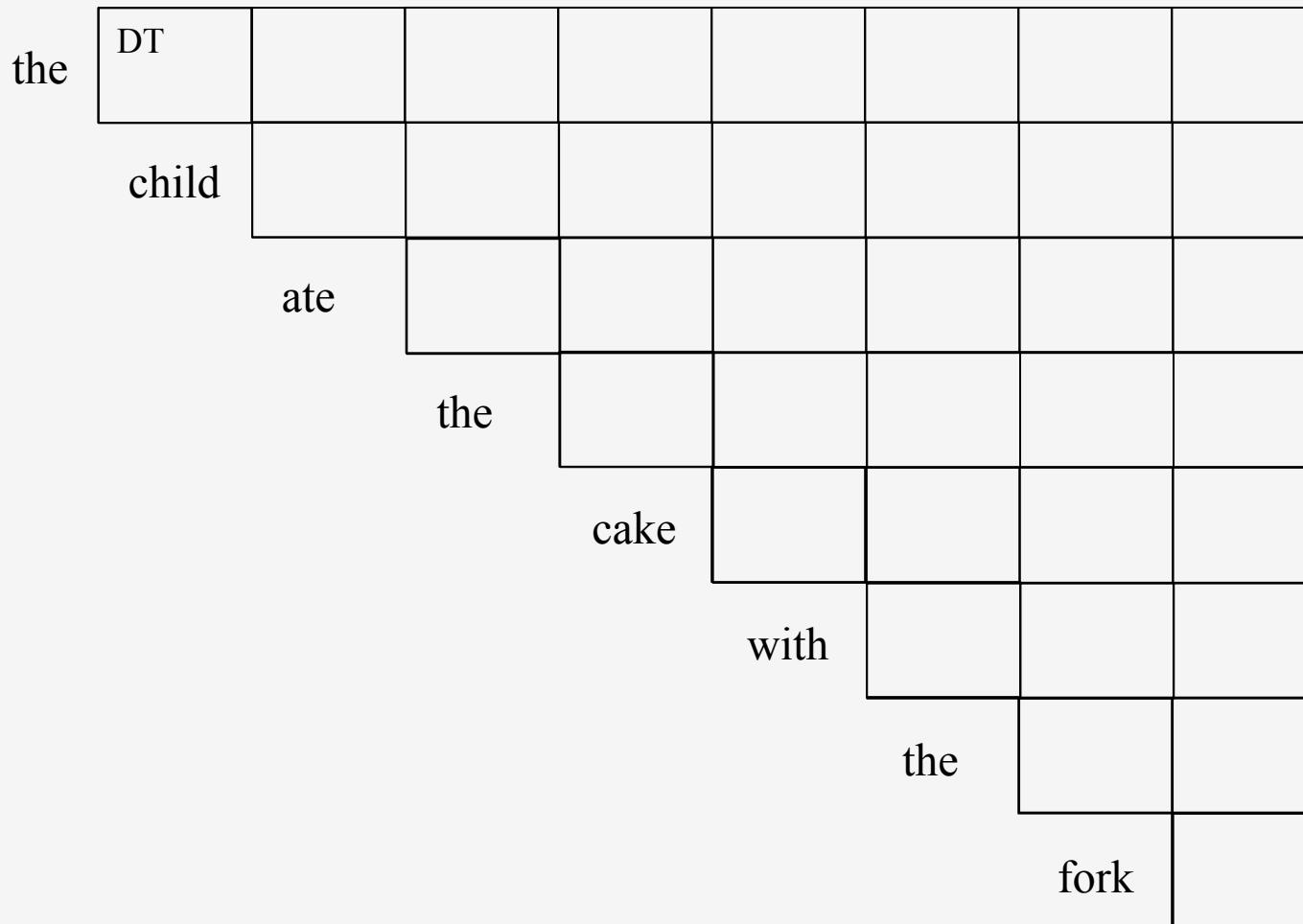
DT → 'a' | 'the'

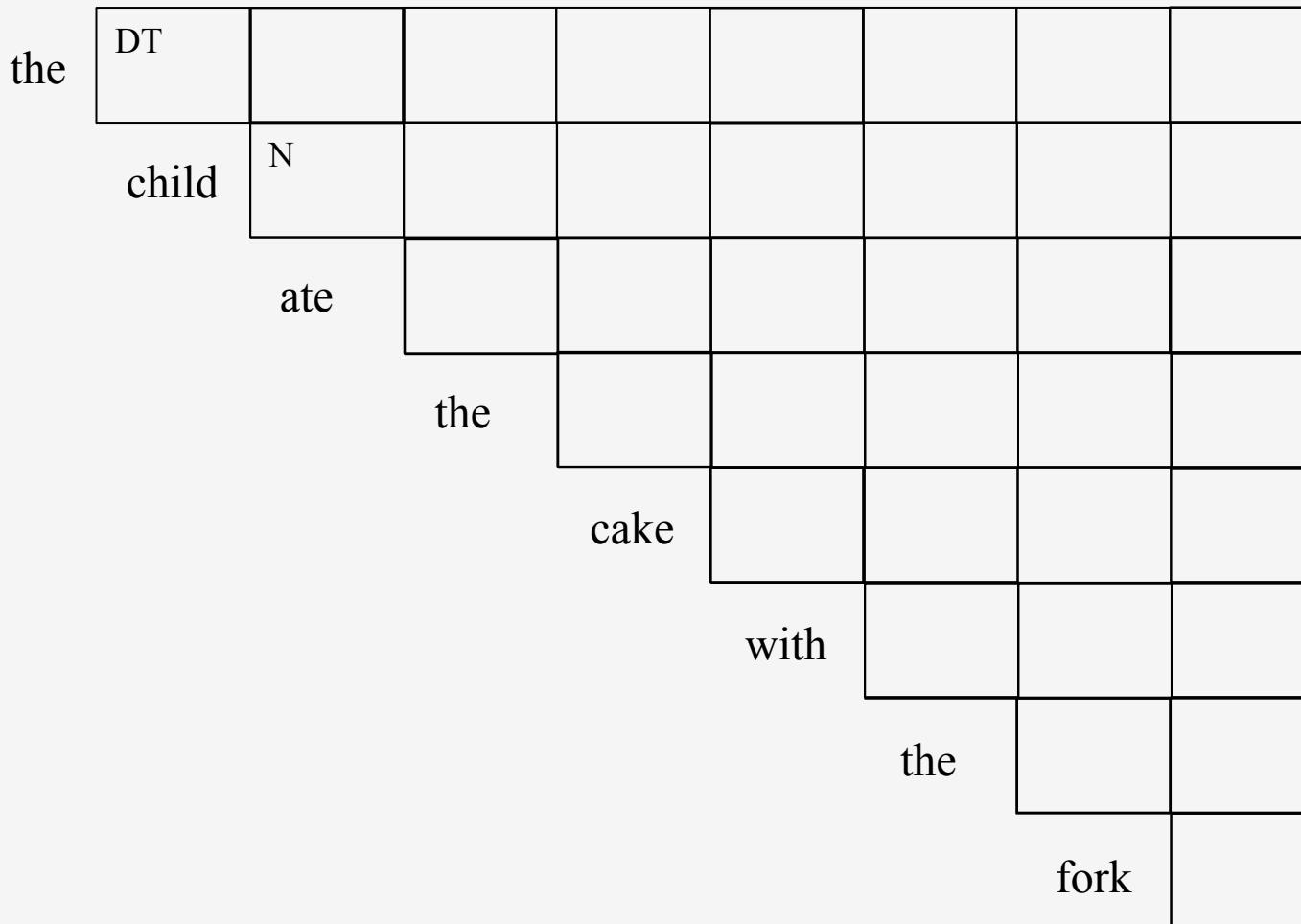
N → 'child' | 'cake' | 'fork'

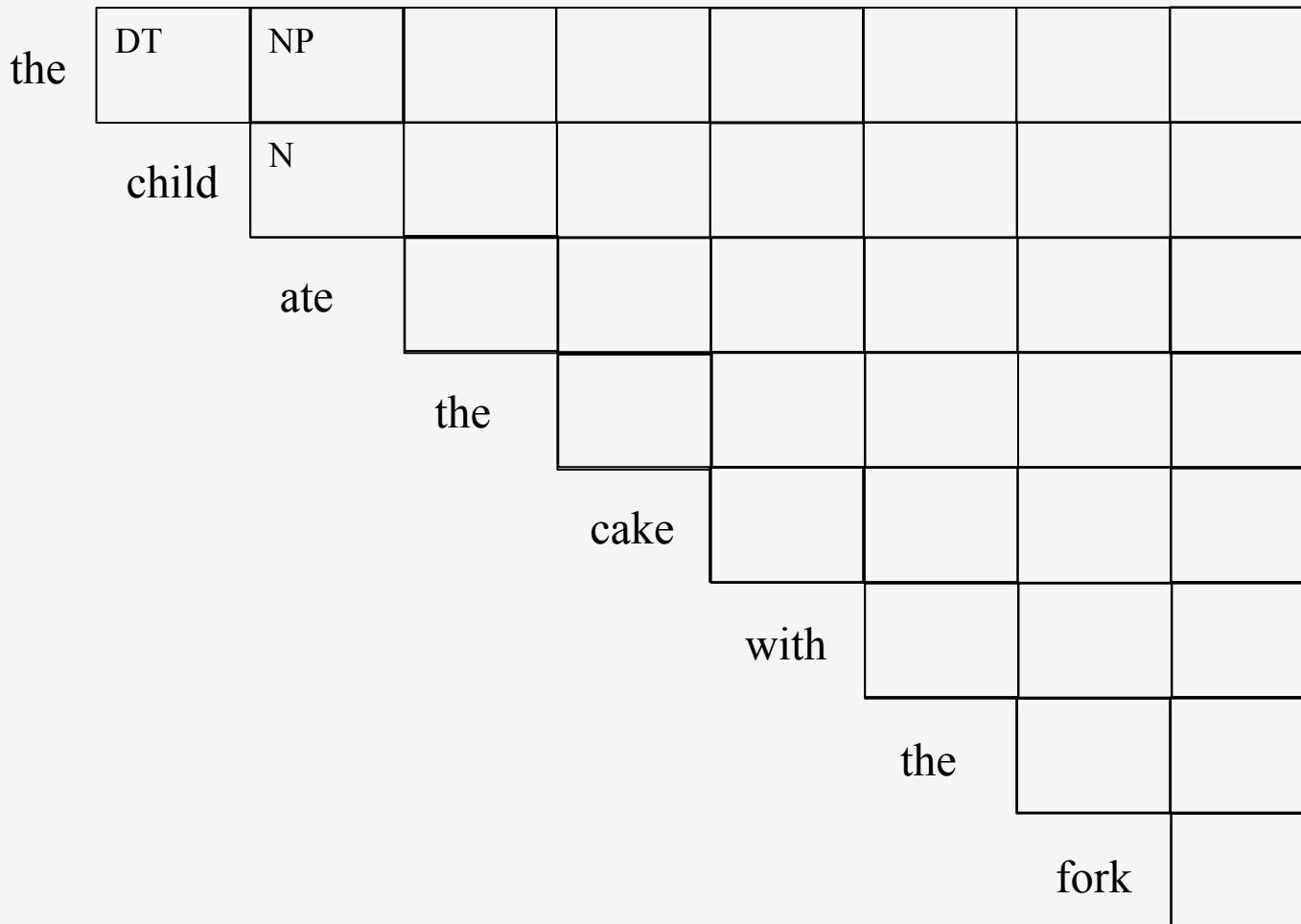
PRP → 'with' | 'to'

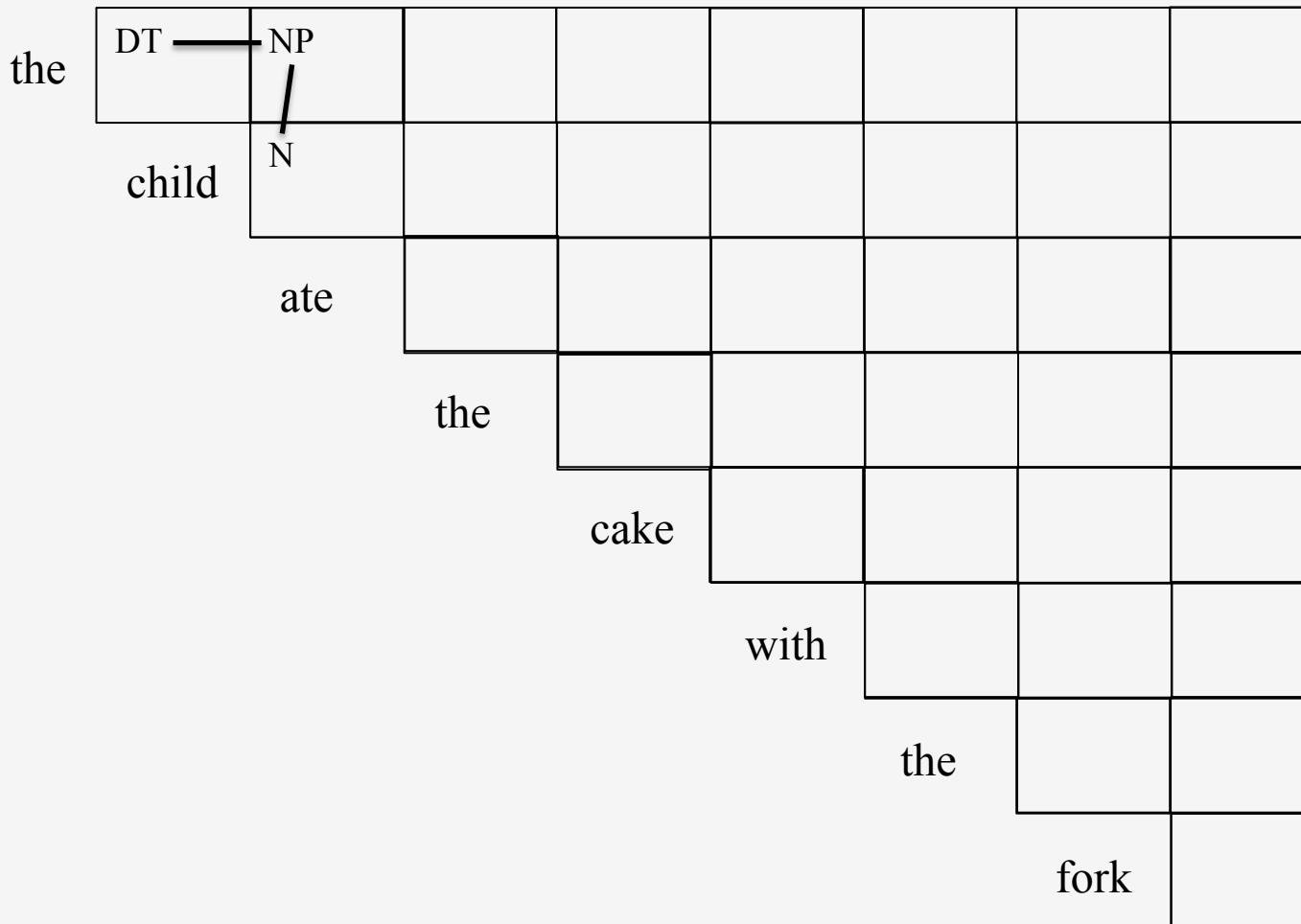
V → 'saw' | 'ate'

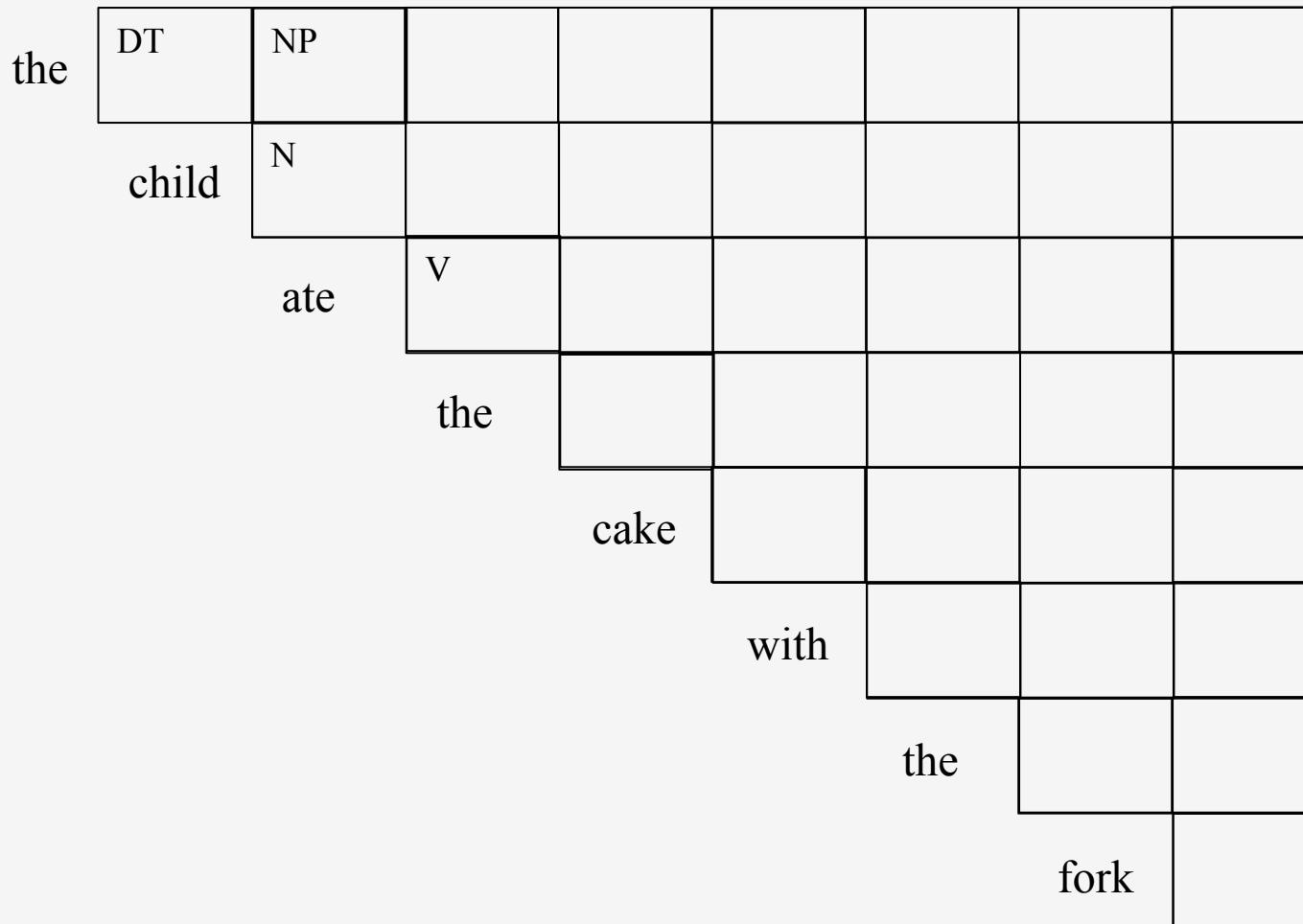


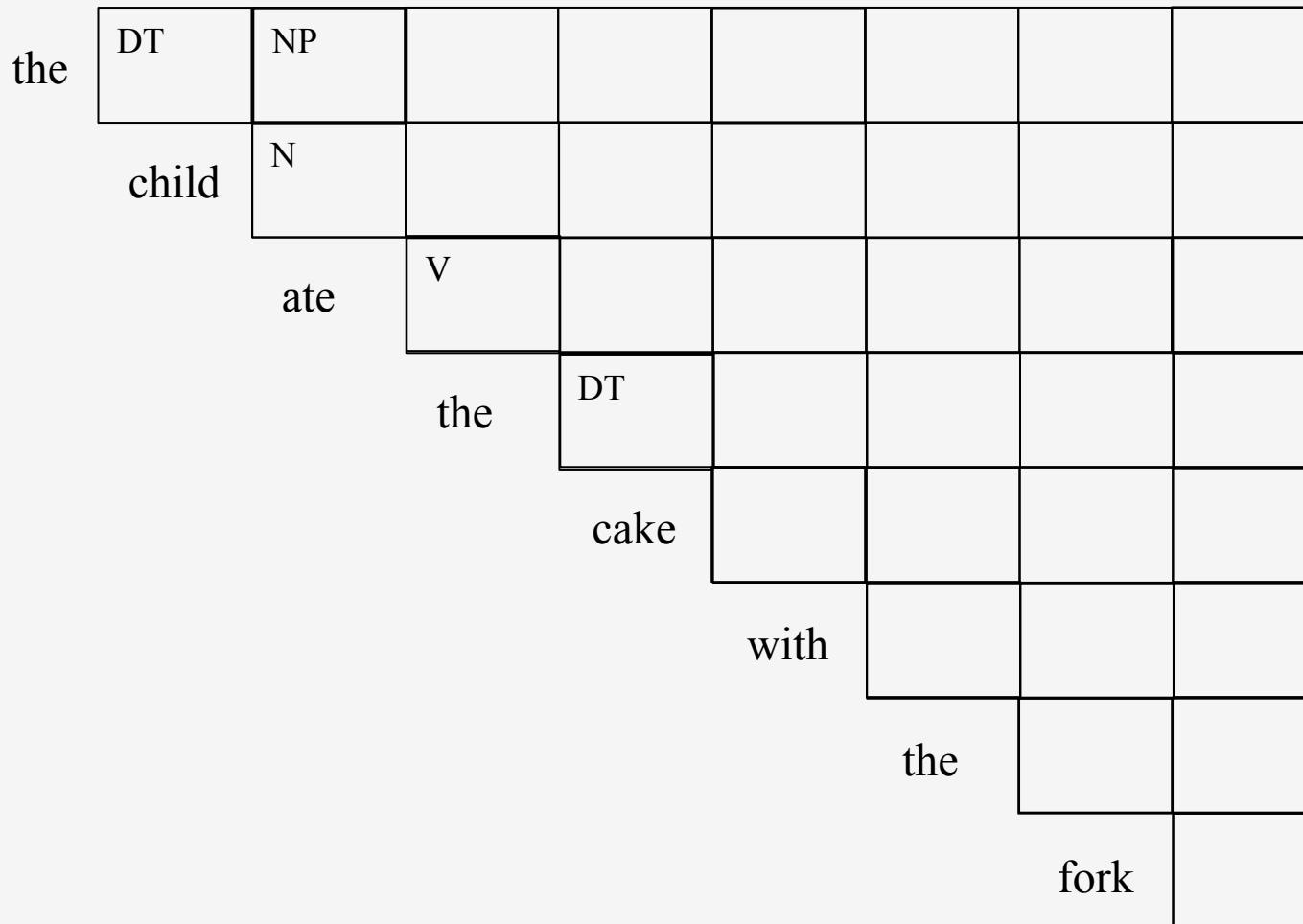






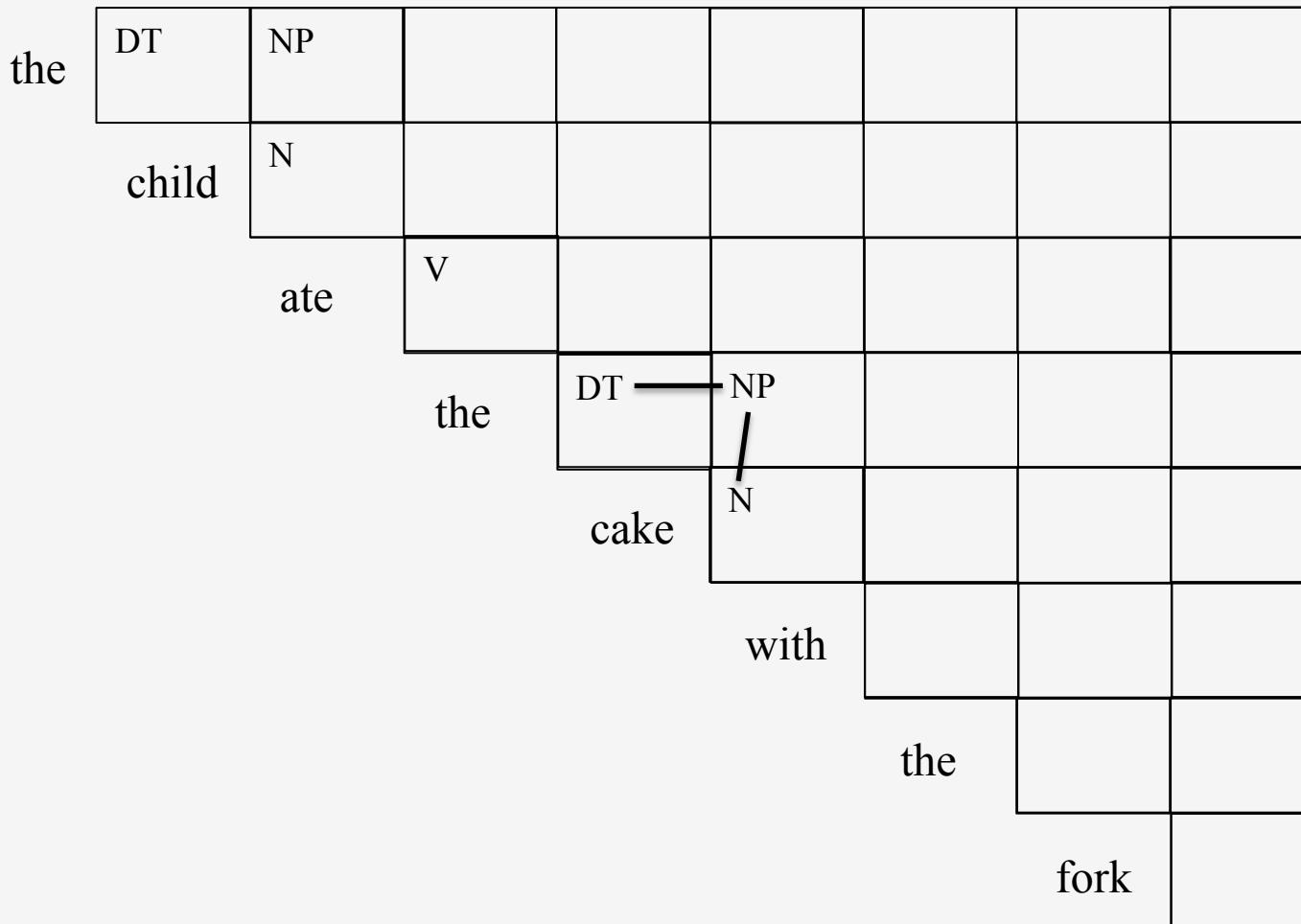


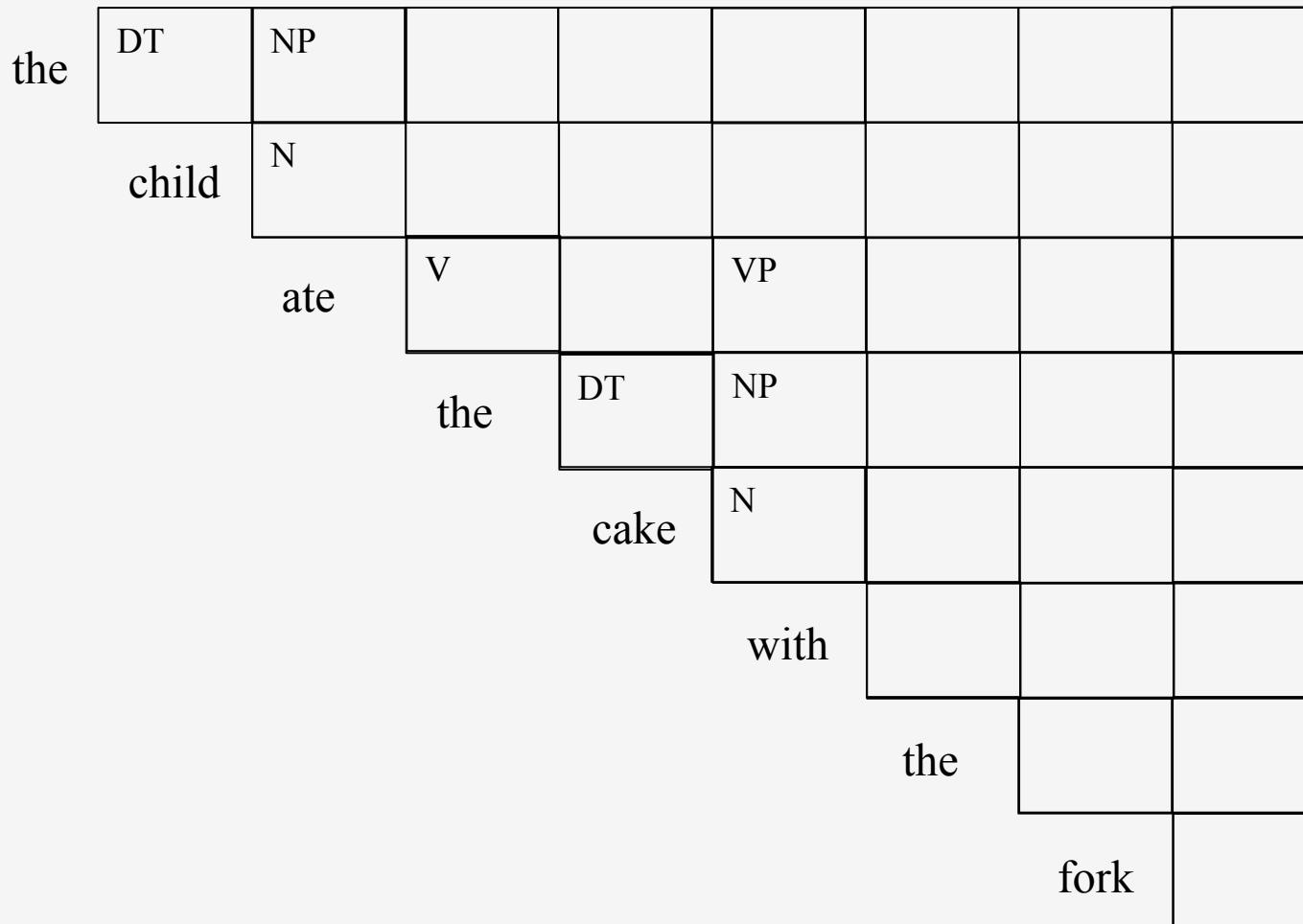


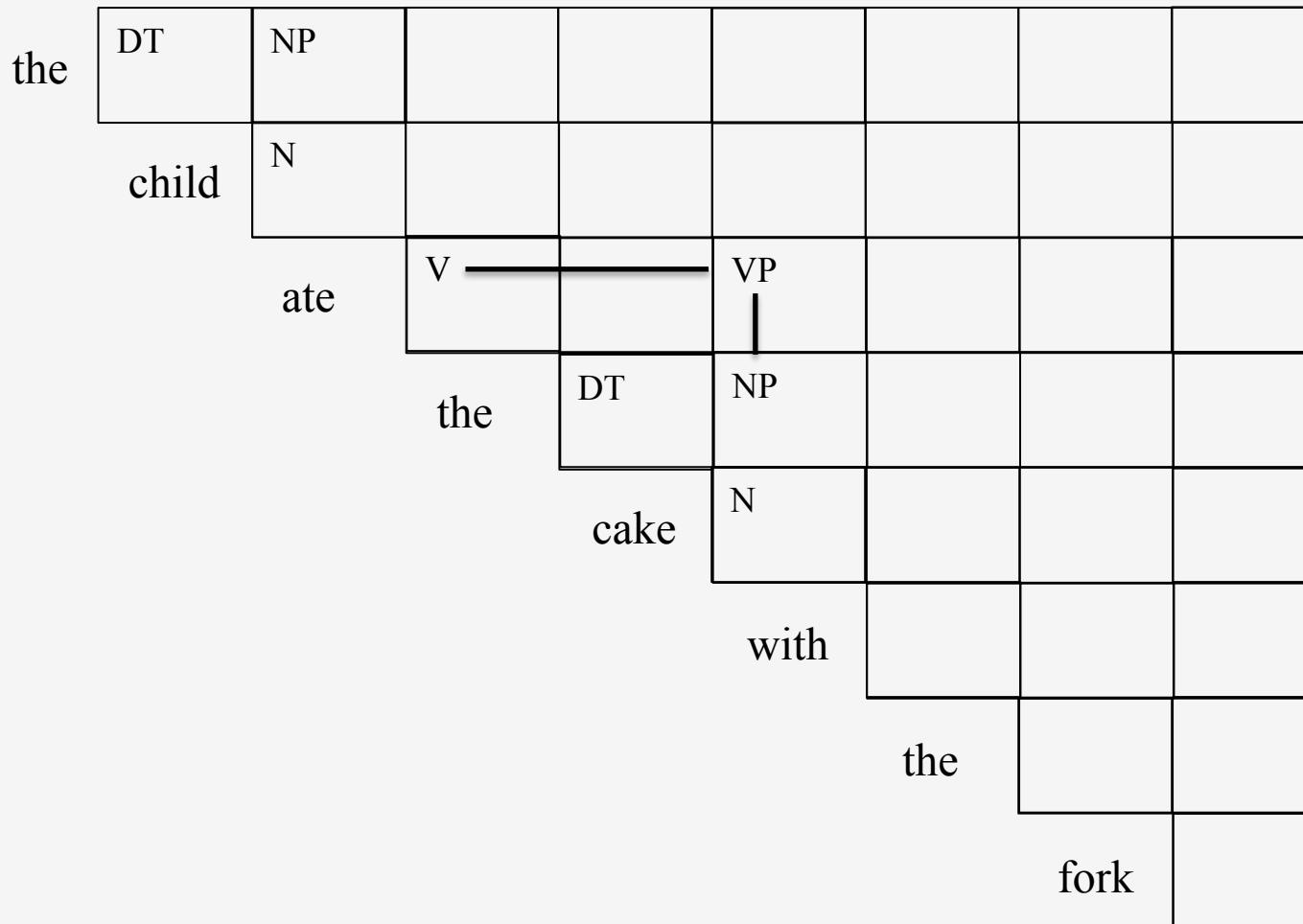


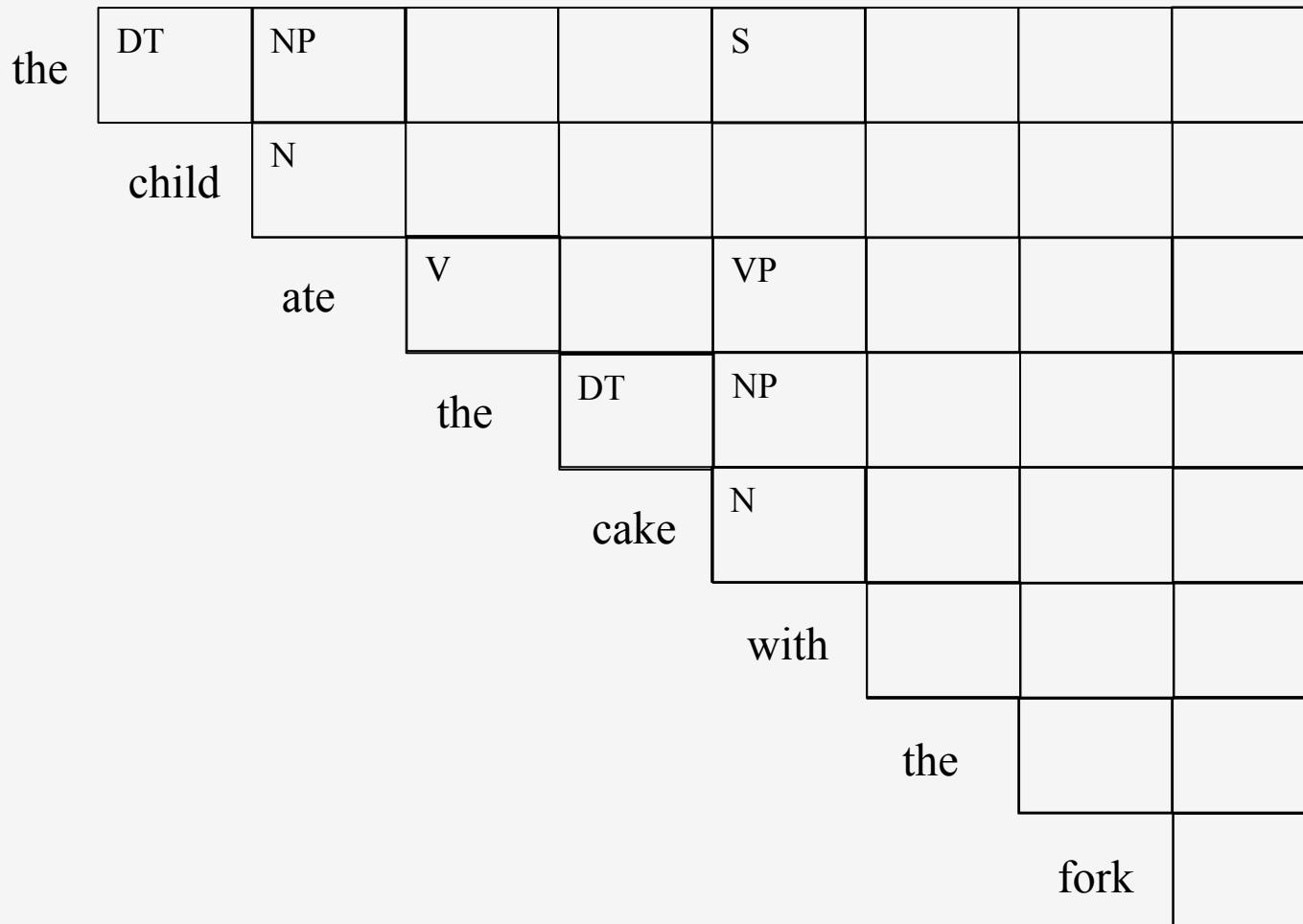
the	DT	NP						
child		N						
ate		V						
	the		DT					
	cake			N				
		with						
			the					
			fork					

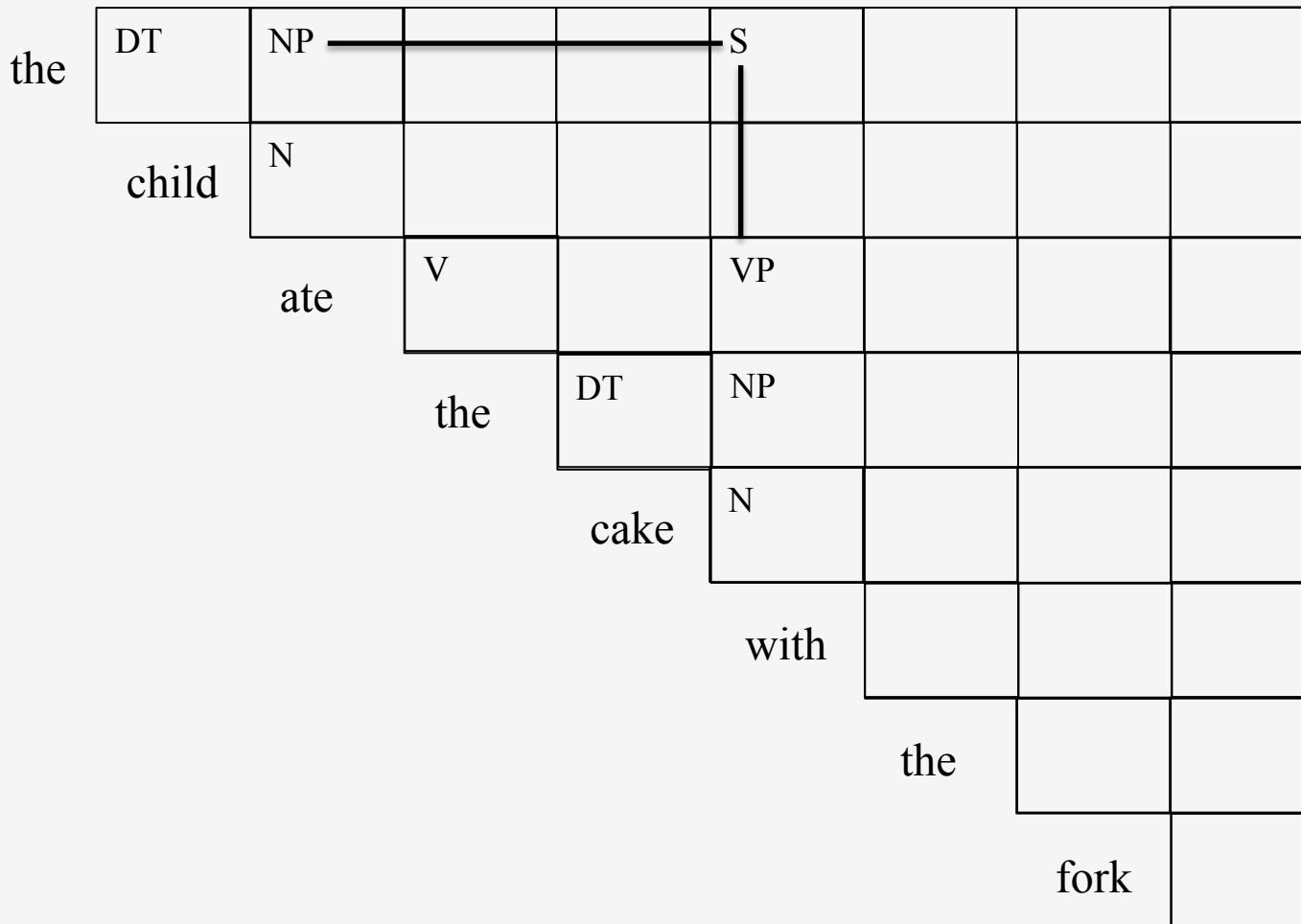
the	DT	NP						
child		N						
ate		V						
	the	DT	NP					
	cake		N					
	with							
	the							
	fork							

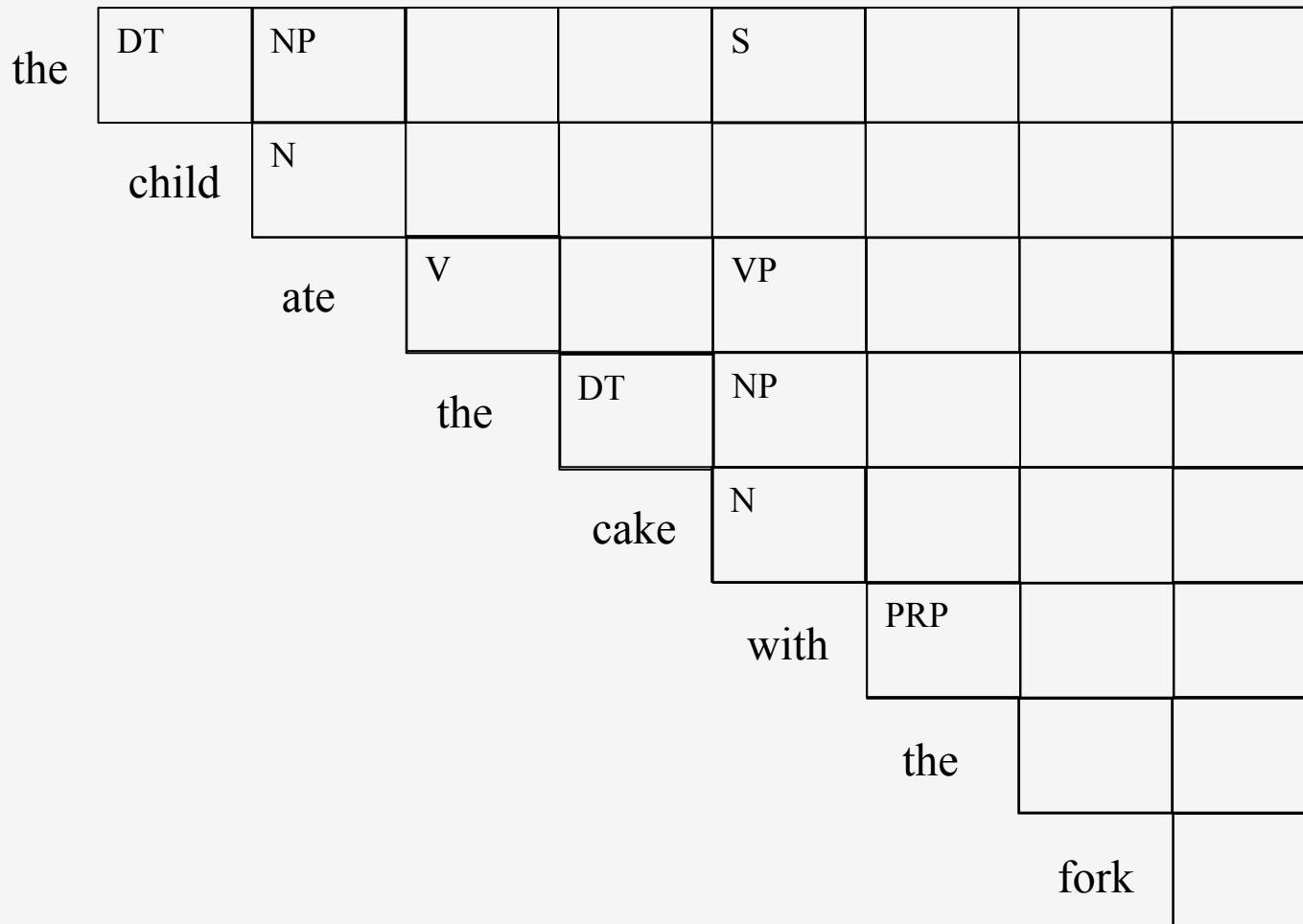








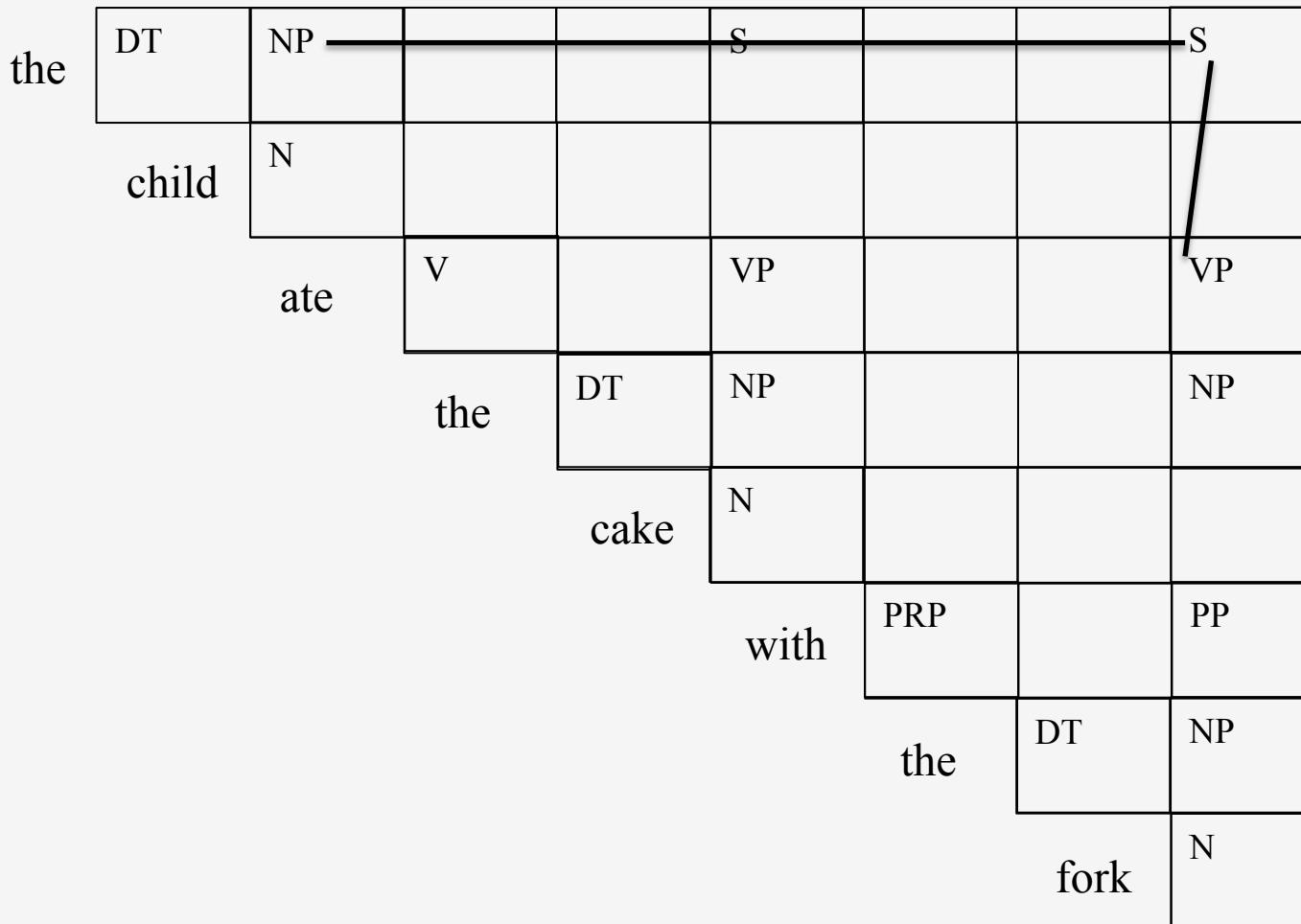


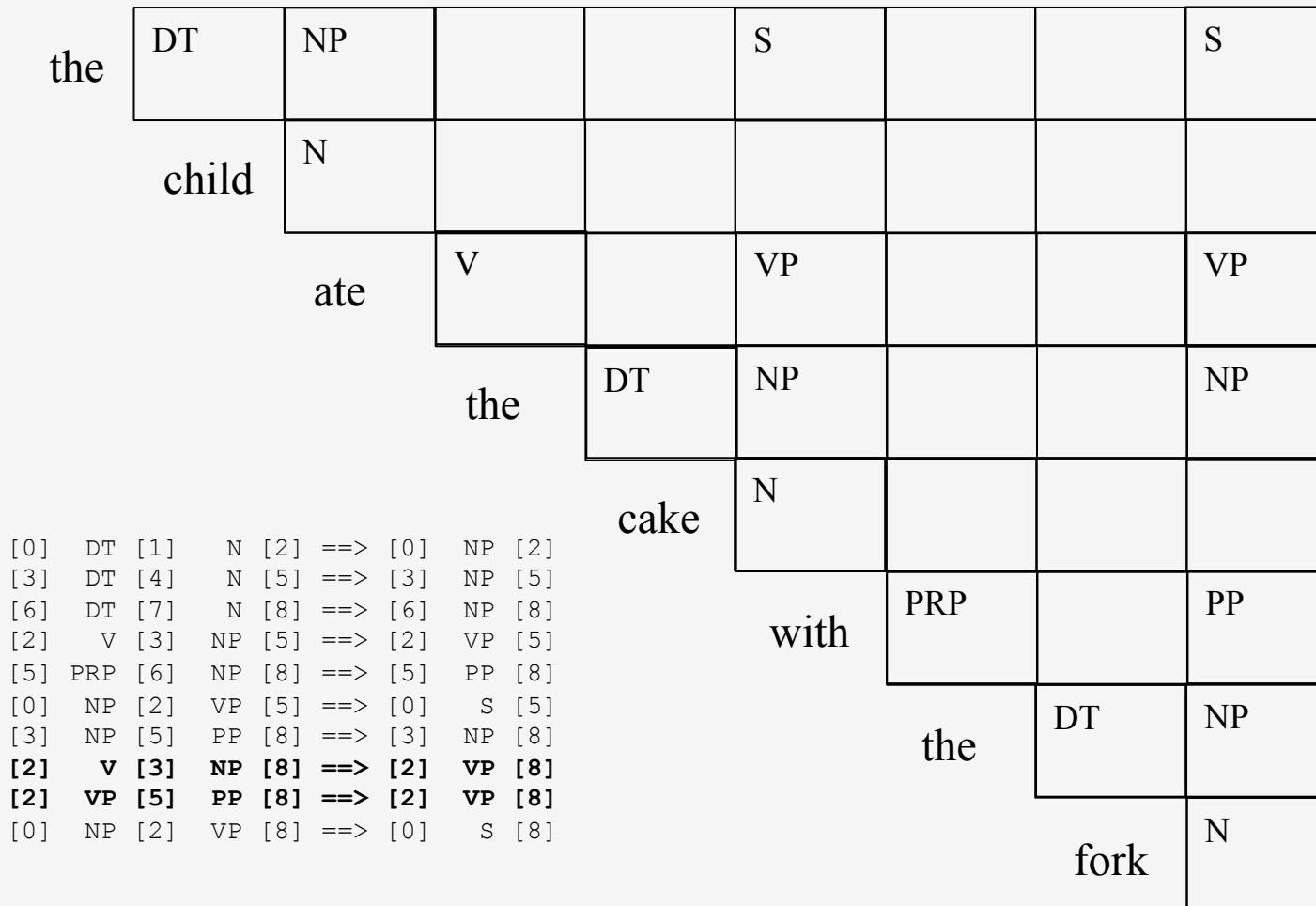


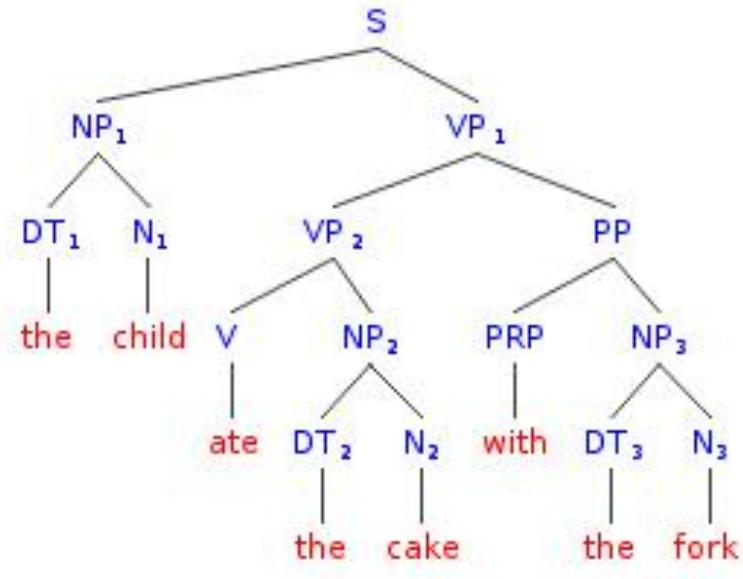
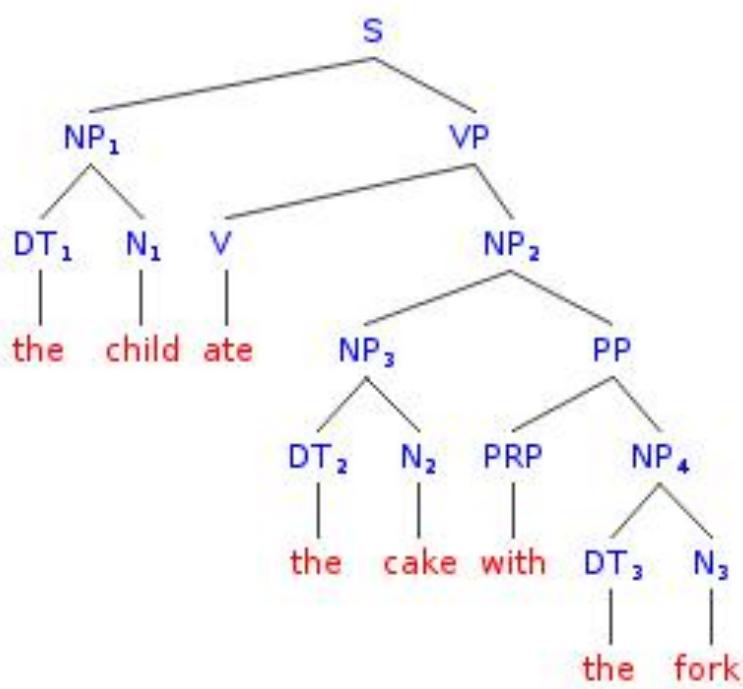
the	DT	NP			S			
child		N						
ate		V		VP				
the		DT	NP				NP	
cake		N						
with			PRP				PP	
the				DT	NP			
fork					N			

the	DT	NP			S			
child		N						
ate		V		VP			VP	
	the	DT	NP				NP	
cake		N						
	with		PRP				PP	
	the		DT		NP			
	fork				N			

the	DT	NP			S			
child		N						
ate	V		VP		VP		/	
the		DT	NP				NP	
cake		N						
with			PRP				PP	
the				DT	NP			
fork					N			







What is the *meaning* of each of these sentences?

```
(S
  (NP (DT the) (N child) )
  (VP
    (VP (V ate) (NP (DT the) (N cake) ))
    (PP (PRP with) (NP (DT the) (N fork))))))
```

(S
 (NP (DT the) (N child))
 (VP
 (VP (V ate) (NP (DT the) (N cake)))
 (PP (PRP with) (NP (DT the) (N fork)))))

(S
 (NP (DT the) (N child))
 (VP
 (V ate)
 (NP
 (NP (DT the) (N cake))
 (PP (PRP with) (NP (DT the) (N fork))))))

Online Demo

- [http://www.diotavelli.net/people/void/
demos/cky.html](http://www.diotavelli.net/people/void/demos/cky.html)

Complexity of CKY

- There are $O(n^2)$ cells in the table
- Single parse
 - Each cell requires a linear lookup.
 - Total time complexity is $O(n^3)$
- All parses
 - Total time complexity is exponential

A Longer Example

["take", "this", "book"]

S → NP VP | Aux NP VP | VP
NP → PRON | Det Nom
Nom → N | Nom N | Nom PP
PP → PRP NP
VP → V | V NP | VP PP
Det → 'the' | 'a' | 'this'
PRON → 'he' | 'she'
N → 'book' | 'boys' | 'girl'
PRP → 'with' | 'in'
V → 'takes' | 'take'

Non-binary Productions

["take", "this", "book"]

S → NP VP | **Aux** NP VP | VP
NP → **PRON** | Det Nom
Nom → **N** | Nom N | Nom PP
PP → PRP NP
VP → **V** | V NP | VP PP
Det → 'the' | 'a' | 'this'
PRON → 'he' | 'she'
N → 'book' | 'boys' | 'girl'
PRP → 'with' | 'in'
V → 'takes' | 'take'

Chomsky Normal Form

- All rules have to be in binary form:
 - $X \rightarrow Y Z$ or $X \rightarrow w$
- This introduces new non-terminals for
 - hybrid rules
 - n-ary rules
 - unary rules

ATIS Grammar

Original version

$S \rightarrow NP\ VP$

$S \rightarrow Aux\ NP\ VP$

$S \rightarrow VP$

$NP \rightarrow Pronoun$

$NP \rightarrow Proper-Noun$

$NP \rightarrow Det\ Nominal$

$Nominal \rightarrow Noun$

$Nominal \rightarrow Nominal\ Noun$

$Nominal \rightarrow Nominal\ PP$

$VP \rightarrow Verb$

$VP \rightarrow Verb\ NP$

$VP \rightarrow VP\ PP$

$PP \rightarrow Prep\ NP$

ATIS Grammar In CNF

Original version

$S \rightarrow NP\ VP$
 $S \rightarrow Aux\ NP\ VP$

 $S \rightarrow VP$

$NP \rightarrow Pronoun$
 $NP \rightarrow Proper-Noun$
 $NP \rightarrow Det\ Nominal$
 $Nominal \rightarrow Noun$
 $Nominal \rightarrow Nominal\ Noun$
 $Nominal \rightarrow Nominal\ PP$
 $VP \rightarrow Verb$
 $VP \rightarrow Verb\ NP$
 $VP \rightarrow VP\ PP$
 $PP \rightarrow Prep\ NP$

CNF version

$S \rightarrow NP\ VP$
 $S \rightarrow X1\ VP$
 $X1 \rightarrow Aux\ NP$
 $S \rightarrow book \mid include \mid prefer$
 $S \rightarrow Verb\ NP$
 $S \rightarrow VP\ PP$
 $NP \rightarrow I \mid he \mid she \mid me$
 $NP \rightarrow Houston \mid NWA$
 $NP \rightarrow Det\ Nominal$
 $Nominal \rightarrow book \mid flight \mid meal \mid money$
 $Nominal \rightarrow Nominal\ Noun$
 $Nominal \rightarrow Nominal\ PP$
 $VP \rightarrow book \mid include \mid prefer$
 $VP \rightarrow Verb\ NP$
 $VP \rightarrow VP\ PP$
 $PP \rightarrow Prep\ NP$

ATIS Grammar In CNF

Original version

S → NP VP
S → Aux NP VP

S → VP

NP → Pronoun
NP → Proper-Noun
NP → Det Nominal
Nominal → Noun
Nominal → Nominal Noun
Nominal → Nominal PP
VP → Verb
VP → Verb NP
VP → VP PP
PP → Prep NP

CNF version

S → NP VP
S → X1 VP
X1 → Aux NP
S → book | include | prefer
S → Verb NP
S → VP PP
NP → I | he | she | me
NP → Houston | NWA
NP → Det Nominal
Nominal → book | flight | meal | money
Nominal → Nominal Noun
Nominal → Nominal PP
VP → book | include | prefer
VP → Verb NP
VP → VP PP
PP → Prep NP

Chomsky Normal Form

- All rules have to be in binary form:
 - $X \rightarrow Y Z$ or $X \rightarrow w$
- New non-terminals for hybrid rules, n-ary and unary rules:
 - $\text{INF-VP} \rightarrow \text{to VP}$ *becomes*
 - $\text{INF-VP} \rightarrow \text{TO}$
 - $\text{TO} \rightarrow \text{to}$
 - $S \rightarrow \text{Aux NP VP}$ *becomes*
 - $S \rightarrow R1 \text{ VP}$
 - $R1 \rightarrow \text{Aux NP}$
 - $S \rightarrow \text{VP VP} \rightarrow \text{Verb VP} \rightarrow \text{Verb NP VP} \rightarrow \text{Verb PP}$ *becomes*
 - $S \rightarrow \text{book}$
 - $S \rightarrow \text{buy}$
 - $S \rightarrow R2 \text{ PP}$
 - $S \rightarrow \text{Verb PP}$
 - etc.

Issues with CKY

- Weak equivalence only
 - Same language, different structure
 - If the grammar had to be converted to CNF, then the final parse tree doesn't match the original grammar
 - However, it can be converted back using a specific procedure
- Syntactic ambiguity
 - (Deterministic) CKY has no way to perform syntactic disambiguation

NLP

NLP

Introduction to NLP

Earley Parser

Earley's Parser

- **Background**
 - Developed by Jay Earley in 1970
 - No need to convert the grammar to CNF
 - Left to right
- **Complexity**
 - Faster than $O(n^3)$ in many cases

Earley's Parser

- Looks for both full and partial constituents
- Example:
 - $S \rightarrow \text{Aux} . \text{NP VP}$
- When reading word k , it has already identified all hypotheses that are consistent with words 1 to $k-1$
- Example:
 - If the parser matches NP in the example above
 - $S \rightarrow \text{Aux NP} . \text{VP}$

Earley's Parser

- It uses a dynamic programming table, just like CKY
- Example entry in column 1
 - $[0:1] \text{ VP } \rightarrow \text{ VP } . \text{ PP}$
 - Created when processing word 1
 - Corresponds to words 0 to 1 (these words correspond to the VP part of the RHS of the rule)
 - The dot separates the completed (known) part from the incomplete (and possibly unattainable) part

Earley's Parser

- Three types of entries
 - ‘scan’ – for words
 - ‘predict’ – for non-terminals
 - ‘complete’ – otherwise

S → NP VP
S → Aux NP VP
S → VP
NP → PRON
NP → Det Nom
Nom → N
Nom → Nom N
Nom → Nom PP
PP → PRP NP
VP → V
VP → V NP
VP → VP PP
Det → 'the'
Det → 'a'
Det → 'this'
PRON → 'he'
PRON → 'she'
N → 'book'
N → 'boys'
N → 'girl'
PRP → 'with'
PRP → 'in'
V → 'takes'
V → 'take'

```
| .      take   .      this   .      book   . |
| [-----]           .           . | [0:1] 'take'
| .           [-----]           . | [1:2] 'this'
| .           .           [-----] | [2:3] 'book'
```

Example created using NLTK

```
| .      take   .      this   .      book   .|
| [-----]           .                   .| [0:1] 'take'
| .      [-----]           .                   .| [1:2] 'this'
| .      .      [-----] | [2:3] 'book'
|>      .      .      .| [0:0] S  -> * NP VP
|>      .      .      .| [0:0] S  -> * Aux NP VP
|>      .      .      .| [0:0] S  -> * VP
|>      .      .      .| [0:0] VP -> * V
|>      .      .      .| [0:0] VP -> * V NP
|>      .      .      .| [0:0] VP -> * VP PP
|>      .      .      .| [0:0] V  -> * 'take'
|>      .      .      .| [0:0] NP -> * PRON
|>      .      .      .| [0:0] NP -> * Det Nom
```

```
| .      take   .      this   .      book   . |
| [-----]           .           . | [0:1] 'take'
| .      [-----]           . | [1:2] 'this'
| .      .      [-----] | [2:3] 'book'
| >      .      .      . | [0:0] S -> * NP VP
| >      .      .      . | [0:0] S -> * Aux NP VP
| >      .      .      . | [0:0] S -> * VP
| >      .      .      . | [0:0] VP -> * V
| >      .      .      . | [0:0] VP -> * V NP
| >      .      .      . | [0:0] VP -> * VP PP
| >      .      .      . | [0:0] V -> * 'take'
| >      .      .      . | [0:0] NP -> * PRON
| >      .      .      . | [0:0] NP -> * Det Nom
| [-----]           . | [0:1] V -> 'take' *
| [-----]           . | [0:1] VP -> V *
| [----->]           . | [0:1] VP -> V * NP
| .      >      . | [1:1] NP -> * PRON
| .      >      . | [1:1] NP -> * Det Nom
```

(S (VP (V take) (NP (Det this) (Nom (N book)))))

```

| .      take   .      this   .      book   . |
| [-----]           .           . | [0:1] 'take'
| .      [-----]           . | [1:2] 'this'
| .      .      [-----] | [2:3] 'book'
| >      .      .      . | [0:0] S -> * NP VP
| >      .      .      . | [0:0] S -> * Aux NP VP
| >      .      .      . | [0:0] S -> * VP
| >      .      .      . | [0:0] VP -> * V
| >      .      .      . | [0:0] VP -> * V NP
| >      .      .      . | [0:0] VP -> * VP PP
| >      .      .      . | [0:0] V -> * 'take'
| >      .      .      . | [0:0] NP -> * PRON
| >      .      .      . | [0:0] NP -> * Det Nom
| [-----]           . | [0:1] V -> 'take' *
| [-----]           . | [0:1] VP -> V *
| [-----]>          . | [0:1] VP -> V * NP
| .      >          . | [1:1] NP -> * PRON
| .      >          . | [1:1] NP -> * Det Nom

```

```

| .      >          .      . | [1:1] Det -> * 'this'
| [-----]           .           . | [0:1] S -> VP *
| [-----]>          .      . | [0:1] VP -> VP * PP
| .      >          .      . | [1:1] PP -> * PRP NP
| .      [-----]           . | [1:2] Det -> 'this' *
| .      [-----]>          . | [1:2] NP -> Det * Nom
| .      .      >          . | [2:2] Nom -> * N
| .      .      >          . | [2:2] Nom -> * Nom N
| .      .      >          . | [2:2] Nom -> * Nom PP
| .      .      >          . | [2:2] N -> * 'book'
| .      .      [-----] | [2:3] N -> 'book' *
| .      .      [-----] | [2:3] Nom -> N *
| .      [-----] | [1:3] NP -> Det Nom *
| .      .      [-----]> | [2:3] Nom -> Nom * N
| .      .      [-----]> | [2:3] Nom -> Nom * PP
| .      .      .      > | [3:3] PP -> * PRP NP
| [=====] | [0:3] VP -> V NP *
| [=====] | [0:3] S -> VP *
| [-----]> | [0:3] VP -> VP * PP

```

NLP

Introduction to NLP

Issues with Context-free grammars

Agreement

- Number
 - Chen is/people are
- Person
 - I am/Chen is
- Tense
 - Chen was reading/Chen is reading/Chen will be reading
- Case
 - not in English but in many other languages such as German, Russian, Greek
- Gender
 - not in English but in many other languages such as German, French, Spanish

Combinatorial Explosion

- Many combinations of rules are needed to express agreement
 - $S \rightarrow NP\ VP$
 - $S \rightarrow 1sgNP\ 1sgVP$
 - $S \rightarrow 2sgNP\ 2sgVP$
 - $S \rightarrow 3sgNP\ 3sgVP$
 - ...
 - $1sgNP \rightarrow 1sgN$
 - ...

Subcategorization Frames

- Direct object
 - The dog ate a sausage
- Prepositional phrase
 - Mary left the car in the garage
- Predicative adjective
 - The receptionist looked worried
- Bare infinitive
 - She helped me buy this place
- To-infinitive
 - The girl wanted to be alone
- Participial phrase
 - He stayed crying after the movie ended
- That-clause
 - Ravi doesn't believe that it will rain tomorrow
- Question-form clauses
 - She wondered where to go

CFG independence Assumptions

- Non-independence
 - All NPs
 - 11% NP PP, 9% DT NN, 6% PRP
 - NPs under S
 - 9% NP PP, 9% DT NN, 21% PRP
 - NPs under VP
 - 23% NP PP, 7% DT NN, 4% PRP
 - (example from Dan Klein)
- Lexicalized grammars
 - later

Conclusions

- Syntax helps understand the meaning of a sentence.
 - Bob gave Alice a flower
 - Who gave a flower to Alice?
 - What did Bob give to Alice?
- Context-free grammars are an appropriate representation for syntactic information
- Dynamic programming is needed for efficient parsing
 - Cubic time to find one parse
 - Still exponential time to find all parses
 - Why?

Answer

- Why does it still take an exponential time to find all parses?
 - Very simple – because the number of parses can be exponential

NLP

NLP

Introduction to NLP

The Penn Treebank

Description

- **Background**
 - From the early 90's
 - Developed at the University of Pennsylvania
 - (Marcus, Santorini, and Marcinkiewicz 1993)
- **Size**
 - 40,000 training sentences
 - 2400 test sentences
- **Genre**
 - Mostly Wall Street Journal news stories and some spoken conversations
- **Importance**
 - Helped launch modern automatic parsing methods

External Links

- Treebank-3
 - <http://catalog.ldc.upenn.edu/LDC99T42>
- Original version
 - <http://catalog.ldc.upenn.edu/LDC95T7>
- Tokenization guidelines
 - <http://www.cis.upenn.edu/~treebank/tokenization.html>
- The American National Corpus
 - <http://www.americannationalcorpus.org/OANC/penn.html>

Penn Treebank tagset (1/2)

Tag	Description	Example
CC	coordinating conjunction	and
CD	cardinal number	1, third
DT	determiner	the
EX	existential there	<i>there</i> is
FW	foreign word	d'oeuvre
IN	preposition/subordinating conjunction	in, of, like
JJ	adjective	green
JJR	adjective, comparative	greener
JJS	adjective, superlative	greenest
LS	list marker	1)
MD	modal	could, will
NN	noun, singular or mass	table
NNS	noun plural	tables
NNP	proper noun, singular	John
NNPS	proper noun, plural	Vikings
PDT	predeterminer	<i>both</i> the boys
POS	possessive ending	friend's

Penn Treebank tagset (2/2)

Tag	Description	Example
PRP	personal pronoun	I, he, it
PRP\$	possessive pronoun	my, his
RB	adverb	however, usually, naturally, here, good
RBR	adverb, comparative	better
RBS	adverb, superlative	best
RP	particle	give <i>up</i>
TO	to	<i>to go, to him</i>
UH	interjection	uhhuhhuhh
VB	verb, base form	take
VBD	verb, past tense	took
VBG	verb, gerund/present participle	taking
VBN	verb, past participle	taken
VBP	verb, sing. present, non-3d	take
VBZ	verb, 3rd person sing. present	takes
WDT	wh-determiner	which
WP	wh-pronoun	who, what
WP\$	possessive wh-pronoun	whose
WRB	wh-abverb	where, when

Example Sentence

- `WSJ/12/WSJ_1273.MRG`, `sentence 11`
- Because the CD had an effective yield of 13.4 % when it was issued in 1984 , and interest rates in general had declined sharply since then , part of the price Dr. Blumenfeld paid was a premium -- an additional amount on top of the CD 's base value plus accrued interest that represented the CD 's increased market value .

Parsed sentence

```
(S
  (SBAR-PRP
    (IN Because)
    (S
      (S
        (NP-SBJ (DT the) (NNP CD))
        (VP
          (VBD had)
          (NP
            (NP (DT an) (JJ effective) (NN yield))
            (PP (IN of) (NP (CD 13.4) (NN %))))
          (SBAR-TMP
            (WHADVP-4 (WRB when))
            (S
              (NP-SBJ-1 (PRP it))
              (VP
                (VBD was)
                (VP
                  (VBN issued)
                  (NP (-NONE- *-1))
                  (PP-TMP (IN in) (NP (CD 1984)))
                  (ADVP-TMP (-NONE- *T*-4)))))))
            ...
          )
        )
      )
    )
  )
)
```



(S
 (SBAR-PRP
 (IN Because)
 (S
 (S
 (NP-SBJ (DT the) (NNP CD))
 (VP
 (VBD had)
 (NP
 (NP (DT an) (JJ effective) (NN yield))
 (PP (IN of) (NP (CD 13.4) (NN %))))
 (SBAR-TMP
 (WHADVP-4 (WRB when))
 (S
 (NP-SBJ-1 (PRP it))
 (VP
 (VBD was)
 (VP
 (VBN issued)
 (NP (-NONE- *-1))
 (PP-TMP (IN in) (NP (CD 1984)))
 (ADVP-TMP (-NONE- *T*-4)))))))
 (, ,)
 (CC and)
 (S
 (NP-SBJ
 (NP (NN interest) (NNS rates))
 (PP (IN in) (ADJP (JJ general))))
 (VP
 (VBD had)
 (VP
 (VBN declined)
 (ADVP-MNR (RB sharply))
 (PP-TMP (IN since) (NP (RB
 then)))))))
 (, ,)
 (NP-SBJ
 (NP (NN part))
 (PP
 (IN of)
 (NP
 (NP (DT the) (NN price))
 (SBAR
 (WHNP-3 (-NONE- 0))
 (S
 (NP-SBJ (NNP Dr.) (NNP Blumenfeld))
 (VP (VBD paid) (NP (-NONE- *T*-3))))))))
 (, .))
 (VP
 (VBD was)
 (NP-PRD
 (NP (DT a) (NN premium))
 (: --)
 (NP
 (NP
 (NP (DT an) (JJ additional) (NN amount))
 (PP-LOC
 (IN on)
 (NP
 (NP (NN top))
 (PP
 (IN of)
 (NP
 (NP (DT the) (NNP CD) (POS 's))
 (NN base)
 (NN value))))))
 (CC plus)
 (NP (VBN accrued) (NN interest))
 (SBAR
 (WHNP-2 (WDT that))
 (S
 (NP-SBJ (-NONE- *T*-2))
 (VP
 (VBD represented)
 (NP
 (NP (DT the) (NNP CD) (POS 's))
 (VBN increased)
 (NN market)
 (NN value))))))

(S
(SBAR-PRP
 (IN Because)
(S
 (NP-SBJ (DT the) (NNP CD))
 (VP
 (VBD had)
 (NP
 (NP (DT an) (JJ effective) (NN yield))
 (PP (IN of) (NP (CD 13.4) (NN %))))
(SBAR-TMP
 (WHADVP-4 (WRB when))
(S
 (NP-SBJ-1 (PRP it))
 (VP
 (VBD was)
 (VP
 (VBN issued)
 (NP (-NONE- *1))
(PP-TMP (IN in) (NP (CD 1984)))
 (ADVP-TMP (-NONE- *T*-4)))))))
(, ,)
(CC and)
(S
 (NP-SBJ
 (NP (NN interest) (NNS rates))
 (PP (IN in) (ADJP (JJ general))))
 (VP
 (VBD had)
 (VP
 (VBN declined)
 (ADVP-MNR (RB sharply))
 (PP-TMP (IN since) (NP (RB
 then)))))))

(VP
(VP
 (VBD was)
 (NP-PRD
 (NP (DT a) (NN premium))
 (: --)
 (NP
 (NP
 (NP (DT an) (JJ additional) (NN amount))
 (PP-LOC
 (IN on)
 (NP
 (NP (NN top))
 (PP
 (IN of)
 (NP
 (NP-SBJ (NNP Dr.) (NNP Blumenfeld))
 (VP (VBD paid) (NP (-NONE- *T*-3)))))))
 (CC plus)
 (NP (VBN accrued) (NN interest)))
(SBAR
 (WHNP-2 (WDT that))
(S
 (NP-SBJ (-NONE- *T*-2))
 (VP
 (VBD represented)
 (NP
 (NP (DT the) (NNP CD) (POS 's))
 (VBN increased)
 (NN market)
 (NN value)))))))
 (. .))

Peculiarities

- Complementizers
 - e.g., “that”
- Gaps
 - *NONE*
- SBAR
 - SBAR → COMP S
 - E.g., “that *NONE* represented the CD’ market value”

tgrep

A < B	A immediately dominates B
A << B	A dominates B
A <- B	B is the last child of A
A <<, B	B is a leftmost descendant of A
A <<` B	B is a rightmost descendant of A
A . B	A immediately precedes B
A .. B	A precedes B
A \$ B	A and B are sisters
A \$. B	A and B are sisters and A immediately precedes B
A \$.. B	A and B are sisters and A precedes B

The Use Of Treebanks

- Disadvantages
 - A lot more work to annotate 40K+ sentences than to write a grammar.
- Advantages
 - Statistics about different constituents and phenomena
 - Training systems
 - Evaluating systems
 - Multilingual extensions

Introduction to NLP

Parsing evaluation

Evaluation Methodology (1/2)

- Classification tasks
 - Document retrieval
 - Part of speech tagging
 - Parsing
- Data split
 - Training
 - Dev-test
 - Test

Evaluation Methodology (2/2)

- Baselines
 - Dumb baseline
 - Intelligent baseline
 - Human performance (ceiling)
- New method
- Evaluation methods
 - Accuracy
 - Precision and Recall
- Multiple references
 - Interjudge agreement

Kappa

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

- Agreement vs. expected agreement
 - $P(A)$ is the level of agreement of the judges
 - $P(E)$ is the expected probability of agreement by chance
- When $\kappa > .7$ – agreement is considered high
- Question
 - Judge agreement on a binary classification task is 60%, is this high?

Answer

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

- Data
 - $P(A) = .6$
 - $P(E) = .5$
- Kappa
 - $\kappa = .1/.5 = .2$
 - not high

Parsing Evaluation

- Precision and recall
 - get the proper constituents
- Labeled precision and recall
 - also get the correct non-terminal labels
- F1
 - harmonic mean of precision and recall
- Crossing brackets
 - $(A (B C))$ vs $((A B) C)$
- PTB corpus
 - training 02-21, development 22, test 23

Evaluation Example

GOLD = (S (NP (DT The) (JJ Japanese) (JJ industrial) (NNS companies))
(VP (MD should) (VP (VB know) (ADVP (JJR better)))) (. .)

CHAR = (S (NP (DT The) (JJ Japanese) (JJ industrial) (NNS companies))
(VP (MD should) (VP (VB know)) ((ADVP (**RBR** better)))) (. .))

Bracketing Recall	=	80.00
Bracketing Precision	=	66.67
Bracketing FMeasure	=	72.73
Complete match	=	0.00
No crossing	=	100.00
Tagging accuracy	=	87.50

NLP

NLP

Parsing

Introduction and recap

Parsing Programming Languages

```
#include <stdio.h>

int main()
{
    int n, reverse = 0;

    printf("Enter a number to reverse\n");
    scanf("%d", &n);

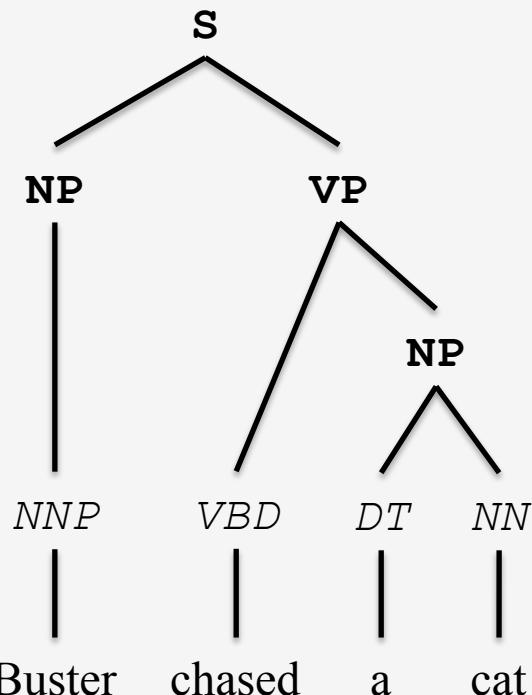
    while (n != 0)
    {
        reverse = reverse * 10;
        reverse = reverse + n%10;
        n = n/10;
    }
    printf("Reverse of entered number is = %d\n", reverse);

    return 0;
}
```

Parsing Human Language

- **Coordination scope:** Small boys and girls are playing.
- **Prepositional phrase attachment:** I saw the man with the telescope.
- **Gaps:** Mary likes Physics but hates Chemistry.
- **Particles vs. prepositions:** She ran up a large bill.
- **Gerund vs. adjective:** Playing cards can be expensive.

Phrase Structure



Parsing

Parsing noun sequences

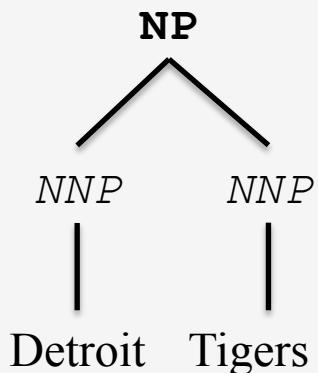
Noun-noun Compounds

- Fish tank = tank that holds fish
- Fish net = net used to catch fish
- Fish soup = soup made with fish
- Fish oil = oil extracted from fish
- Fish sauce = sauce for fish dishes? sauce made of fish?

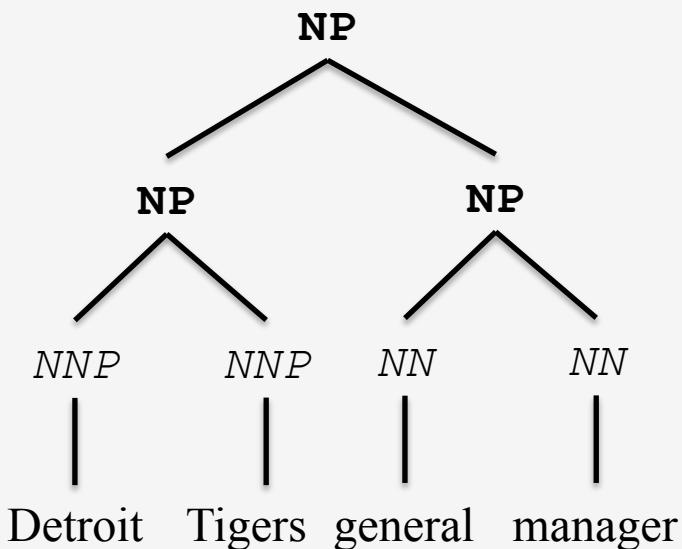
Noun-noun Compounds

- Head of the compound
 - College junior – a kind of junior
 - Junior college – a kind of college
- Head first?
 - Attorney general
- Adjectives?
 - New Mexico, general manager
- More than two nouns?
 - luxury car dealership

Noun Phrase Consisting Of Two Nouns



Noun Phrase Consisting Of Four Nouns



Representation Using Parentheses

- ((Salt Lake) City)
- (Salt (Lake City))
- Salt Lake City mayor?

Solution

- (((Salt Lake) City) mayor)

Representation Using Parentheses

- ((Salt Lake) City) mayor)
- ((Detroit Tigers) (general manager))
- Leland Stanford Junior University?

Solution

- (((Leland Stanford) Junior) University)

Combinatorics

- $n=2$
(A B)
- $n=3$
((A B) C)
(A (B C))
- $n=4$
((A B)(C D))

Solution

- $n=4$

((A B)(C D))

(A (B (C D)))

(A ((B C) D))

((A (B C)) D)

(((A B) C) D)

What About $n > 4$?

- $n=5$

$((A\ B)((C\ D)E))$

...

Solution

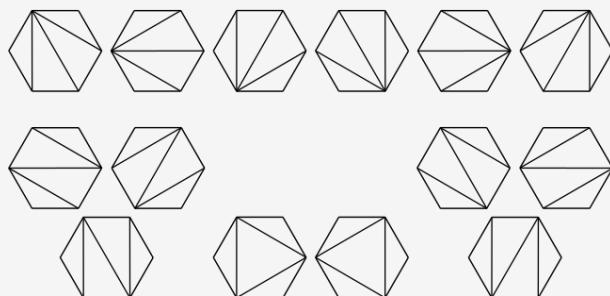
- The general solution is $C(n)$, a notation for the n^{th} Catalan number

$$C \downarrow n = 1/n+1 (\boxed{2n} @ n), \text{ for } n \geq 0$$

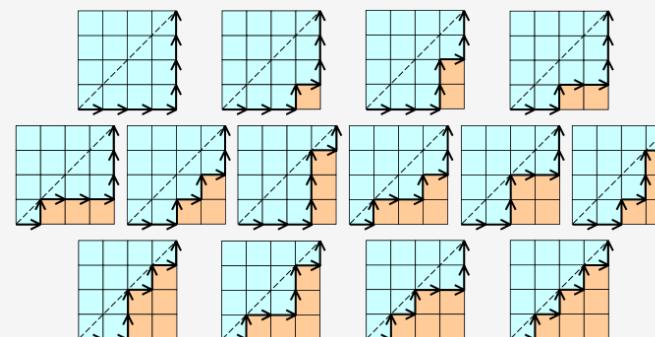
- 1, 1, 2, 5, 14, 42, 132, 429, 1430, 4862, 16796, 58786, 208012, 742900, ...
- Sequence A000108 in the On-Line Encyclopedia of Integer Sequences® (OEIS®)
- <https://oeis.org/>

Other Uses Of Catalan Numbers

- the number of different ways a convex polygon with $n + 2$ sides can be cut into triangles by connecting vertices with straight lines.
- the number of monotonic paths along the edges of a grid with $n \times n$ square cells, which do not pass above the diagonal.



<http://en.wikipedia.org/wiki/File:Catalan-Hexagons-example.svg>



http://en.wikipedia.org/wiki/File:Catalan_number_4x4_grid_example.svg

NLP

NLP

Parsing

Prepositional Phrase Attachment
(1)

Penn Treebank Representation

```
( (S
  (NP-SBJ
    (NP (NNP Pierre) (NNP Vinken) )
    (, ,)
    (ADJP
      (NP (CD 61) (NNS years) )
      (JJ old) )
    (, ,) )
  (VP (MD will)
    (VP (VB join)
      (NP (DT the) (NN board) )
      (PP-CLR (IN as)
        (NP (DT a) (JJ nonexecutive) (NN director) )))
    (NP-TMP (NNP Nov.) (CD 29) )))
  (. .) ))
```

Penn Treebank Representation

```
( (S
  (NP-SBJ (NNP Mr.) (NNP Vinken) )
  (VP (VBZ is)
    (NP-PRD
      (NP (NN chairman) )
      (PP (IN of)
        (NP
          (NP (NNP Elsevier) (NNP N.V.) )
          (, ,)
          (NP (DT the) (NNP Dutch) (VBG publishing) (NN group) )))))
    (. .) )))
```

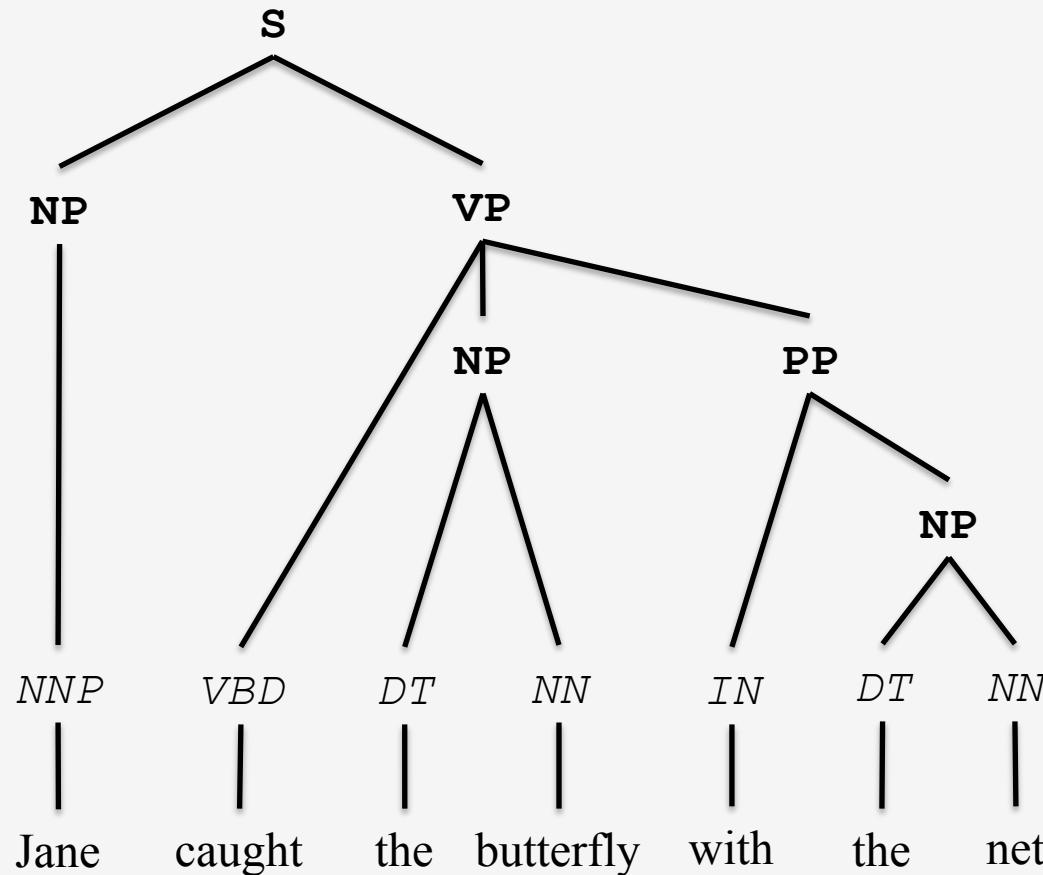
Prepositional Phrase Attachment

- High (verbal):

join board as director

- Low (nominal):

is chairman of Elsevier



Examples

- Examples:
 - Lucy's plane leaves Detroit **on Monday**. – high
 - Jenna met Mike **at the concert**. – high
 - This painting must cost millions **of dollars**. – low
- High or low attachment?
 - Alicia ate spaghetti **from Italy**.
 - Alicia ate spaghetti **with meatballs**.
 - Alicia ate spaghetti **with a fork**.
 - Alicia ate spaghetti **with Justin**.
 - Alicia ate spaghetti **with delight**.
 - Alicia ate spaghetti **on Friday**.

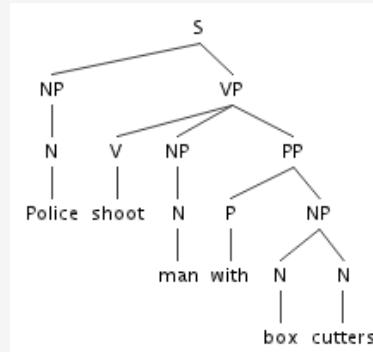
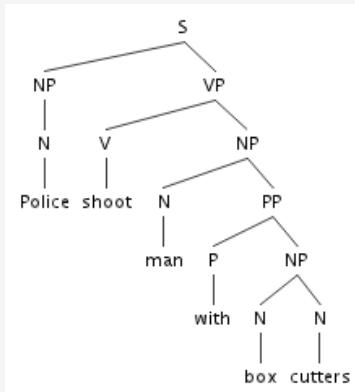
Solution

- High or low attachment?
 - Alicia ate spaghetti **from Italy**. – low
 - Alicia ate spaghetti **with meatballs**. – low
 - Alicia ate spaghetti **with a fork**. – high
 - Alicia ate spaghetti **with Justin**. – high
 - Alicia ate spaghetti **with delight**. – high
 - Alicia ate spaghetti **on Friday**. – high

Actual Headline

- Police shoot man with box cutters.

(S (NP (N Police)) (VP (V shoot) (NP (N man) (PP (P with) (NP (N box) (N cutters))))))
 (?) (S (NP (N Police)) (VP (V shoot) (NP (N man)) (PP (P with) (NP (N box) (N cutters))))))



Prepositional Phrase Attachment

- Input: a prepositional phrase and the surrounding context
- Output: a binary label: 0(high) or 1(low)
- In practice: the context consists only of four words: the preposition, the verb before the preposition, the noun before the preposition, and the noun after the preposition
- Example: *join board as director*
- Why?

Answer

- Because almost all the information needed to classify a prepositional phrase's attachment as high or low is contained in these four features.
- Furthermore, using only these *tuples* of four features allows for a consistent and scaleable approach.

Sample Tuples

Sent #	Verb	Noun ₁	Preposition	Noun ₂	Class
0	join	board	as	director	V
2	named	director	of	conglomerate	N
3	caused	percentage	of	deaths	N
6	bring	attention	to	problem	V
12	led	team	of	researchers	N
16	including	three	with	cancer	N
24	imposed	ban	on	uses	N
26	made	paper	for	filters	N
28	dumped	sacks	of	material	N
28	dumped	sacks	into	bin	V

Sidebar (1/2)

- The linguistics (and psycholinguistics) literature offers competitive explanations for attachment.
- One theory (Kimball 1973) favors the so-called *right association* rule. It says that, given a new phrase and two choices for attachment, people tend to attach the new phrase with the more recent (“rightmost” within the sentence) of the candidate nodes, resulting in low attachment.
- Alternatively, the *minimal attachment* principle (Frazier 1978) favors an attachment that results in the syntax tree having fewer additional syntactic nodes (in this case, favoring high attachment).
- As one can see from the statistics, none of these methods alone can explain the high prevalence of both types of attachment.

Sidebar (2/2)

- Some other observations can be made by performing statistical analysis of a training set.
- The standard corpus used for this sort of analyses comes from (RRR 1994) and includes 27,937 prepositional phrases extracted from the Penn Treebank (Marcus et al. 1993), divided into three groups (20,801 training, 4039 development, and 3097 test).
- This data representation makes the assumption that additional context is only marginally more useful for classification purposes compared to the four features in the table (verb, noun1, preposition, and noun2).
- For comparison, the sentence matching the data point “bring attention to problem” is actually “Although preliminary findings were reported more than a year ago, the latest results appear in today’s New England Journal of Medicine, a forum likely to bring new attention to the problem.” It is unlikely that the information in the first $\frac{3}{4}$ of the sentence will affect the classification of the prepositional phrase “to the problem”.

NLP

NLP

Parsing

*Prepositional Phrase Attachment
(2)*

Supervised Learning: Evaluation

- Manually label a set of instances.
- Split the labeled data into training and testing sets.
- Use the training data to find patterns.
- Apply these patterns on the testing data set.
- For evaluation: use **accuracy** (the percentage of correct labels that a given algorithm has assigned on the testing data).
- Compare with a simple baseline method.
- What is the simplest baseline method?

Answer

- The simplest supervised baseline method is to find the more common class (label) in the training data and assign it to **all** instances of the testing data set.

Algorithm 1

label the tuple as “low” (default)

Random baseline

- A **random** unsupervised baseline would have been to label each instance in the testing data set with a random label, 0 or 1.
- Practically, random performance is the lower bound against which any non-random methods should be compared.

Measuring The Accuracy Of The Supervised Baseline Method (Algorithm 1)

- In the official training data set (RRR94), the rate of occurrence of the 1 label (low attachment) is 52.2% ($10,865/20,801$).
- Is the accuracy of this baseline method then equal to 52.2%?

Solution

- No, this is not how accuracy is computed. It has to be computed on the testing set, not the training set.
- Using the official split, the accuracy of this method on the testing set is 59.0% ($1,826/3,097$). One shouldn't think of this number as a good result. The difference (+6.8% going from training to testing) could have been in the opposite direction, resulting in a performance below random.

Observations

- If the baseline method is simple and if the testing set is randomly drawn from the full data set and the data set is large enough, one could expect that the accuracy on the testing set is comparable to the one on the training set. Note that the PTB data set is drawn from business news stories. If one were to train a method on this data and test it on a different set of sentences, e.g., from a novel, it is possible that the two sets will have very different characteristics.
- The more complicated the method is, however, the more likely it will be that it will “overfit” the training data, learning patterns that are too specific to the training data itself and which may not appear in the testing data or which may be associated with the opposite class in the testing data.

Upper Bounds On Accuracy

- The 52% accuracy we've seen so far is our current lower bound. Now, what is the upper bound?
- Usually, human performance is used for the upper bound. For PP attachment, using the four features mentioned earlier, human accuracy is around 88%.
- So, a hypothetical algorithm that achieves an accuracy of 87% is in fact very close to the upper bound (on a scale from 52% to 88%, it is 97% of the way to the upper bound).

Using Linguistic Knowledge

- One way to beat the two baselines is to use linguistic information. For example, the preposition “of” is much more likely to be associated with low attachment than high attachment.
- In the training data set, this number is an astounding 98.7% ($5,534/5,607$)
- Therefore the feature *prep_of* is very valuable. What are the two main reasons?

Answer

- Reason 1 – it is very informative (98.7% of the time it is linked with the low attachment class)
- Reason 2 – it is very frequent (27.0% of the entire training set – 5607/20801).
- Reason 1 alone would not be sufficient!

Sidebar 1/3

- The PTB (Penn Treebank) data set has been used for competitive evaluation of pp attachment algorithms since 1994.
- Each new algorithm is allowed to look at the training and development sets and use any knowledge extracted from them.
- The test set data can never be looked at and can be used only once per algorithm for its evaluation.
- Doing the contrary (repeatedly tuning a new algorithm based on its performance on the designated test set) results in a performance level that is irreproducible on new data and such approaches are not allowed in NLP research.
- Note that the development set can be used in a well-specified way as part of the training process.

Sidebar 2/3

- Let's look at the training data then and see if we can learn some patterns that would help us improve over the silly "label everything as noun attachment" baseline and its 52% expected accuracy.
- For example, some prepositions tend to favor particular attachments.
- Let's start with "against". It appears 172 times in the training set, of which 82 (48%) are noun attached, the rest (52%) being high attached. This ratio (48:52) is very similar to the baseline (52:48), so clearly, knowing that the preposition is "against" gives us very little new information.

Sidebar 3/3

- Furthermore, the total occurrence of “against” in the training corpus is rather small (less than 1% of the prepositional phrases).
- In two special cases, however, the identity of the preposition gives a lot of information.
- At one extreme, “of” is associated with low attachment in 99% of the cases, whereas at the other end of the scale, “to” is associated with high attachment in 82% of its appearances.
- It is also important to note that both of these prepositions are fairly common in the training set (“to” representing 27% of all prepositions and “of” accounting for another 11% of them).
- With this knowledge, we can build a very simple decision list algorithm (Brill and Resnik) that looks like this (the rules are sorted by their expected accuracy on their majority class).

Are There Any Other Such Features?

- Yes, *prep_to*: $2,182/2,699 = 80.8\%$ in favor of high attachment.
- Which leads us to our next algorithm:

Algorithm 2

If the preposition is “of”, label the tuple as “low”.

Else

If the preposition is “to”, label the tuple as “high”.

Else

label the tuple as “low” (default)

Sidebar 1/2

- Let's compare the performance of this simple decision list algorithm with that of the baseline.
- On the training set, the first rule would fire in 5,577 cases, of which 5,527 will be correctly labeled as low attachment and another 50 will be incorrectly labeled as high attachment (the accuracy of this rule is expected to be 99% on the training set).
- The second rule (with an expected accuracy of 82% on the training set) would result in 2,672 decisions, of which 2,172 will be correctly processed as high attachment and 500 will be mislabeled as low attachment.
- Finally, the default rule will kick in, whenever the first two rules are not applicable. In this case, it will apply on all remaining phrases ($20,801 - 5,577 - 2,672 = 12,552$ cases).
- Specifically, among these, it will result in 4,837 phrases correctly labeled as low and 7,714 incorrectly labeled as high.

Sidebar 2/2

- Overall, we have $5,527 + 2,172 + 4,837 = 12,536$ correct decisions, or an accuracy of $12,536/20,801 = 60\%$.
- Clearly Algorithm 2 outperforms Algorithm 1 on the training data.
- This is not surprising, since its expected accuracy should be no less than the worst expected accuracy of its rules (that of rule 3) and it is likely to be higher than that lowest number because of the priority given to the easier to classify cases (rules 1 and 2).
- More complicated algorithms can look at additional features, e.g., the nouns or the verb, or some combination thereof. For example, the verb “dropped” appears 75/81 times (93%) along with high attachment and only 7% with low attachment.

NLP

NLP

Parsing

*Prepositional Phrase Attachment
(3)*

Algorithm 2a

If the preposition is “of”, label the tuple as “low”.

Else

If the preposition is “to”, label the tuple as “high”.

Else

label the tuple as “high”

Some Observations

- First, even though the expected performance of rule 3 was 52%, its actual performance on the training set dropped to 39% after rules 1 and 2 were applied.
- In other words, these rules used up some of the information hidden in the data ahead of rule 3 and left it less useful information to rely upon.
- Even more, one can see that a better decision would have been to replace rule 3 with its exact opposite, label everything left at this stage as “high”, which would have boosted the combined performance.
- Algorithm 2a would achieve $5,527 + 2,172 + 7,714 = 15,413$ correct decisions for an overall accuracy of 74% on the training set.
- Second, one cannot help but notice that Algorithms 2 and 2a each have only three rules. We can imagine a classifier with 20,801 rules, one per training example, each rule of the form “if the preposition is “of” and the nouns are such and such and the verbs are such and such, then classify the data point as the actual class observed in the training set”.

Some Observations

- Third, we so far reported performance on the training set.
- Can we project the performance on the training set to the test set?
- Let's start with Algorithms 1 and 3.
- Algorithm 1 labels everything as low attachment. It achieved 52% on the training set. We expect its performance on the test set to be similar. In fact it is 59% ($1,826/3,097$).
- This clearly demonstrates the variability of text across subsets of the data.
- In this case, this variability favors Algorithm 1 since its performance actually goes up when moving to the test set. In other cases (e.g., if we had swapped the training and test sets), its performance would have gone down.
- On average though, its performance on the test data is expected to vary around the performance on the training data.

Algorithm 3

If the preposition is “on” and the verb is “casting” and the first noun is “cloud” and the second noun is “economy”, label the phrase as “high”.

Else

If the preposition is “of” and the verb is “opened” and the first noun is “can” and the second noun is “worms”, label the phrase as “low”.

Else

... 20799 more rules ...

Some Observations 1/4

- Now, let's consider Algorithm 3. It achieved a very high performance on the training data (way above the “upper bound” achieved by humans).
- However, we will now see the meaning of the word *overfitting* in action.
- Algorithm 3 was so specific to the training data that most of the rules it learned don't apply at all in the test set.
- Only 117 combinations (out of 3032) of words in the test set match a combination previously seen in the training set.
- In other words, Algorithm 3 learned a lot of good rules, but it failed to learn many more. In fact, its accuracy on the test data is only around 4%.
- An alternative to Algorithm 3 would be to combine it with a default rule (just like rule 3 in Algorithm 2) that labels everything that Algorithm 3 missed as noun attachment.
- Unfortunately, even this algorithm (let's call it Algorithm 3a) would only achieve a performance slightly above the baseline (Algorithm 1) of 59% on the test data.
- The lesson to learn here is that, on unseen data, a simple algorithm (Algorithm 1) is much better than a really complicated one that overfits (Algorithm 3). Also, the combination of the two (overfitting + baseline) just barely outperforms the baseline itself and is nowhere close to competitive.

Some Observations 2/4

- Clearly this algorithm (Algorithm 3) would achieve close to 100% accuracy on the training set.
- Why “close to 100%” and not “100%”?
- It turns out that the training set there are mutually inconsistent labels for the same data point.
- For example, “won verdict in case” appears once as high and once as low attachment.
- There are a total of 56 such “discrepancies” in the training set.
- Some of them are caused by inconsistent annotators whereas others would require more context (e.g., the entire paragraph or document) to be correctly disambiguated.

Some Observations 3/4

- Next, let's see how algorithms 2 and 2a will fare on the test set.
- First, let's look at Algorithm 2.
- There are 3097 items to classify in the test set.
- Rule 1 correctly classifies 918 out of 926 instances of “of” (99% accuracy) while rule 2 gets 70% accuracy (234/332 correctly classified).
- Rule 3 achieves $810/1,839 = 44\%$.
- Overall the accuracy of Algorithm 2 on the test set is 63% ($1,962/3,097$).
- Again on the test data, Algorithm 2a outperforms Algorithm 2. Its Rule 3 gets $1,029/1,839 = 56\%$ accuracy and the overall accuracy of Algorithm 2a on the test set is 70% ($2,181/3,097$).

Some Observations 4/4

- Let's now summarize the performance of the five algorithms that we have looked at so far.

Algorithm	Number of rules	Training set accuracy	Test set accuracy
Algorithm 1: default	1	52%	59%
Algorithm 2: of/to + default	3	60%	63%
Algorithm 2a: of/to + better default	3	74%	70%
Algorithm 3: memorize everything	20801	near 100%	4%
Algorithm 3a: memorize + default	20802	near 100%	62%

What's Next?

- So far, so good. We have been able to go from 59% test set accuracy to 70% with two simple rules.
- What additional sources of information can we use to improve the algorithm?
- Here are some ideas:
 - use a few more *good* word features (e.g., more prepositions, perhaps some verb and nouns)
 - use clever ways to deal with missing information
 - use lexical semantic information (e.g., synonyms)
 - use additional context beyond the four feature types used so far.

PREP		
about	67	132
as	380	94
at	552	136
for	1136	1044
in	2251	1577
of	73	5534
on	666	550
to	2182	517
with	698	340
VERB		
bring	58	18
buy	217	126
cut	57	34
drop	84	15
follow	15	91
include	22	221
put	178	37

NOUN1		
company	61	33
director	6	51
increase	17	57
loan	23	9
rate	72	68
NOUN2		
asset	5	49
bank	33	31
board	22	23
client	23	9
company	68	204
day	57	14
year	427	106

What's Next?

- Let's first consider a combination of the first two ideas above: looking for ways to use all possible information that can be extracted from the training data.
- This is the approach that was used by Collins and Brooks (1995).
- Their method was based on a principle called backoff which is somewhat of a combination of all the algorithms used so far (e.g., Algorithms 1, 2, and 3).
- Backoff allows us to use the most specific evidence from the training data, when available but then make reasonable approximations for the missing evidence.

Collins and Brooks

- So what do we do? Collins and Brooks used the following algorithm.
- If a 4-tuple is available, use it.
- If not, combine the evidence from the triples that form the 4-tuple (looking only at the triples that include the preposition).
- If that is not available, look at the pairs, then the singletons, and finally use a default class.
- A 4-tuple is just a set of 4 features in a particular order, e.g., (verb, noun1, preposition, noun2).
- The matching term for 3 features is a triple; for 2 features it is a pair; and for 1 feature, the word singleton is used.

If the denominator of the next formula is 0, then use the classification for the 4-tuple

$$\hat{p}_4(H|v, n_1, p, n_2) = \frac{f(H, v, n_1, p, n_2)}{f(v, n_1, p, n_2)}$$

Else

If the denominator of the next formula > 0, then use the following estimate:

$$\hat{p}_3(H|v, n_1, p, n_2) = \frac{f(H, v, n_1, p) + f(H, v, p, n_2) + f(H, n_1, p, n_2)}{f(v, n_1, p) + f(v, p, n_2) + f(n_1, p, n_2)}$$

Else

If the denominator of the next formula > 0, then use the following estimate:

$$\hat{p}_2(H|v, n_1, p, n_2) = \frac{f(H, v, p) + f(H, p, n_2) + f(H, n_1, p)}{f(v, p) + f(p, n_2) + f(n_1, p)}$$

Else

If the denominator of the next formula > 0, then use the following estimate:

$$\hat{p}_1(H|v, n_1, p, n_2) = \frac{f(H, p)}{f(p)}$$

Else label as “low” (default):

$$\hat{p}_0(H|v, n_1, p, n_2) = 0$$

What's Next?

- The idea behind Algorithm 3 was quite reasonable – assume that if the same object appear again (as defined by the same set of four features), it will likely have the same tag.
- The problem with this approach is that there is not enough data in the training set to learn the likely classes of all possible combinations of features.
- Let's do the math. To cover all the data points in the test set, we'd need information in the training set for a total of 102,998,280,840 combinations (more than 100 Billion combinations)!
- How did we arrive at this number? It is simply the product of the numbers 1123, 1295, 52, and 1362, which are, respectively, the numbers of distinct verbs, noun1s, prepositions, and noun2s in the test set.
- It is impossible to label so much data and even if it could be done, there would be billions more combinations needed to cover a new test set.

Other Methods

- Zhao and Lin 2004 – nearest neighbors
- Find most similar examples – 86.5% best accuracy
- Similar to Zavrel, Daelemans, and Veenstra 1997 – memory-based learning
- Boosting – Abney et al. 1999
- Semantics – Stetina and Nagao 1997
- Graph-based method – Toutanova et al. 2004

Comparative Results

Algorithm	Test set accuracy
Algorithm 1: default	59%
Algorithm 2a: of/to + better default	70%
Best class per preposition	72%
Collins and Brooks	84%
K-nearest neighbors	80%
TUMBL	82%
Human average (4-tuples only)	88%
Human average (whole sentence)	93%

NLP

NLP

Introduction to NLP

Statistical Parsing

Need For Probabilistic Parsing

- Time flies like an arrow
 - Many parses
 - Some (clearly) more likely than others
 - Need for a probabilistic ranking method

Probabilistic Context-free Grammars

- Just like (deterministic) CFG, a 4-tuple (N, Σ, R, S)
 - N : non-terminal symbols
 - Σ : terminal symbols (disjoint from N)
 - R : rules $(A \rightarrow \beta) [p]$
 - β is a string from $(\Sigma \cup N)^*$
 - p is the probability $P(\beta | A)$
 - S : start symbol (from N)

Example

S → NP VP

NP → DT N | NP PP

PP → PRP NP

VP → V NP | VP PP

DT → 'a' | 'the'

N → 'child' | 'cake' | 'fork'

PRP → 'with' | 'to'

V → 'saw' | 'ate'

Example

S → NP VP
NP → DT N
NP → NP PP
PP → PRP NP
VP → V NP
VP → VP PP
DT → 'a'
DT → 'the'
N → 'child'
N → 'cake'
N → 'fork'
PRP → 'with'
PRP → 'to'
V → 'saw'
V → 'ate'

Example

S → NP VP
NP → DT N
NP → NP PP
PP → PRP NP
VP → V NP
VP → VP PP
DT → 'a'
DT → 'the'
N → 'child'
N → 'cake'
N → 'fork'
PRP → 'with'
PRP → 'to'
V → 'saw'
V → 'ate'

Example

S → NP VP [p₀=1]
NP → DT N [p₁]
NP → NP PP [p₂]
PP → PRP NP [p₃=1]
VP → V NP [p₄]
VP → VP PP [p₅]
DT → 'a' [p₆]
DT → 'the' [p₇]
N → 'child' [p₈]
N → 'cake' [p₉]
N → 'fork' [p₁₀]
PRP → 'with' [p₁₁]
PRP → 'to' [p₁₂]
V → 'saw' [p₁₃]
V → 'ate' [p₁₄]

Probability Of A Parse Tree

- The probability of a parse tree t given all n productions used to build it:

$$p(t) = \prod_{i=1}^n p(\alpha_i \rightarrow \beta_i)$$

- The most likely parse is determined as follows:

$$\arg \max_{t \in T(s)} p(t)$$

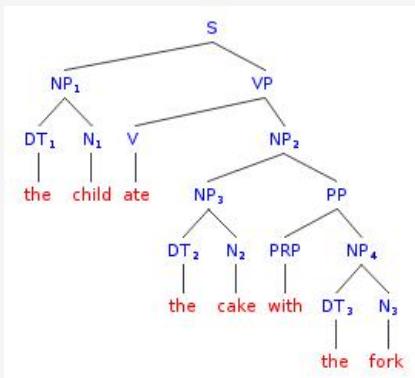
- The probability of a sentence is the sum of the probabilities of all of its parses

Example

S → NP VP [p0=1]
NP → DT N [p1]
NP → NP PP [p2]
PP → PRP NP [p3=1]
VP → V NP [p4]
VP → VP PP [p5]
DT → 'a' [p6]
DT → 'the' [p7]
N → 'child' [p8]
N → 'cake' [p9]
N → 'fork' [p10]
PRP → 'with' [p11]
PRP → 'to' [p12]
V → 'saw' [p13]
V → 'ate' [p14]

Example

S -> NP VP	[p0=1]
NP -> DT N	[p1]
NP -> NP PP	[p2]
PP -> PRP NP	[p3=1]
VP -> V NP	[p4]
VP -> VP PP	[p5]
DT -> 'a'	[p6]
DT -> 'the'	[p7]
N -> 'child'	[p8]
N -> 'cake'	[p9]
N -> 'fork'	[p10]
PRP -> 'with'	[p11]
PRP -> 'to'	[p12]
V -> 'saw'	[p13]
V -> 'ate'	[p14]



Example

S → NP VP [p0=1]

NP → DT N [p1]

NP → NP PP [p2]

PP → PRP NP [p3=1]

VP → V NP [p4]

VP → VP PP [p5]

DT → 'a' [p6]

DT → 'the' [p7]

N → 'child' [p8]

N → 'cake' [p9]

N → 'fork' [p10]

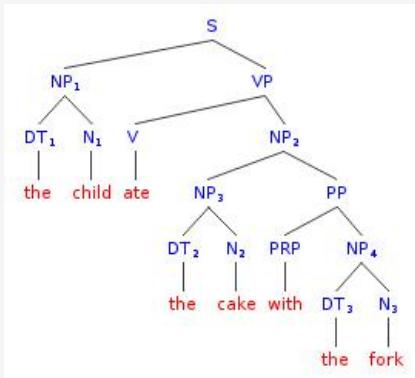
PRP → 'with' [p11]

PRP → 'to' [p12]

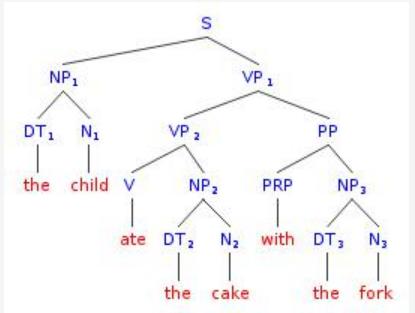
V → 'saw' [p13]

V → 'ate' [p14]

t_1



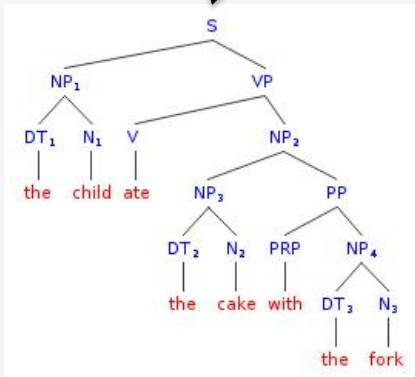
t_2



Example

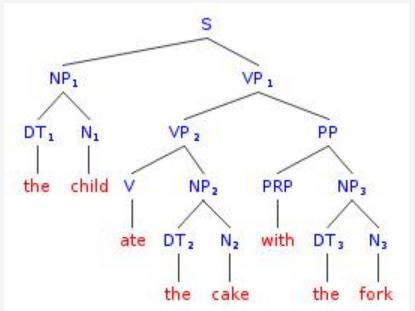
S -> NP VP	[p0=1]
NP -> DT N	[p1]
NP -> NP PP	[p2]
PP -> PRP NP	[p3=1]
VP -> V NP	[p4]
VP -> VP PP	[p5]
DT -> 'a'	[p6]
DT -> 'the'	[p7]
N -> 'child'	[p8]
N -> 'cake'	[p9]
N -> 'fork'	[p10]
PRP -> 'with'	[p11]
PRP -> 'to'	[p12]
V -> 'saw'	[p13]
V -> 'ate'	[p14]

t_1



$$p(t_1) = p_0 p_1 p_4 p_7 p_8 p_{14} p_2 p_1 p_3 p_7 p_9 p_{11} p_1 p_7 p_{10}$$

t_2



Example

S → NP VP [p₀=1]

NP → DT N [p₁]

NP → NP PP [p₂]

PP → PRP NP [p₃=1]

VP → V NP [p₄]

VP → VP PP [p₅]

DT → 'a' [p₆]

DT → 'the' [p₇]

N → 'child' [p₈]

N → 'cake' [p₉]

N → 'fork' [p₁₀]

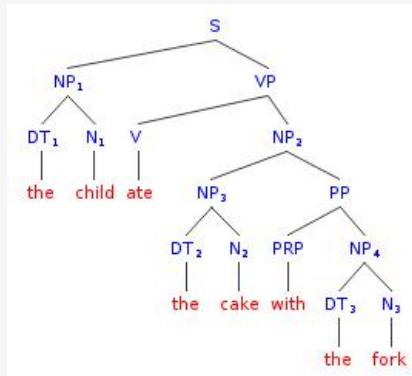
PRP → 'with' [p₁₁]

PRP → 'to' [p₁₂]

V → 'saw' [p₁₃]

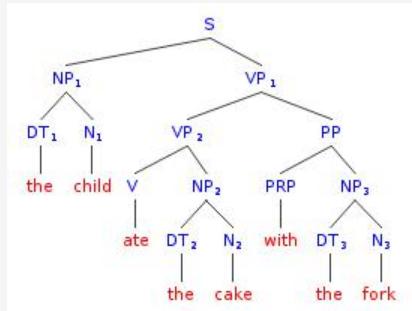
V → 'ate' [p₁₄]

t₁



$$p(t_1) = p_0 p_1 p_4 p_7 p_8 p_{14} p_2 p_1 p_3 p_7 p_9 p_{11} p_1 p_7 p_{10}$$

t₂



$$p(t_2) = p_0 p_1 p_5 p_7 p_8 p_4 p_3 p_{14} p_1 p_{11} p_1 p_7 p_9 p_7 p_{10}$$

Example

S → NP VP [p₀=1]

NP → DT N [p₁]

NP → NP PP [p₂]

PP → PRP NP [p₃=1]

VP → V NP [p₄]

VP → VP PP [p₅]

DT → 'a' [p₆]

DT → 'the' [p₇]

N → 'child' [p₈]

N → 'cake' [p₉]

N → 'fork' [p₁₀]

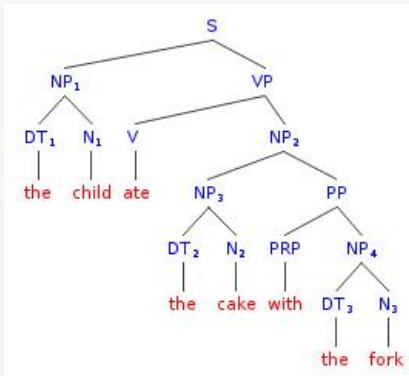
PRP → 'with' [p₁₁]

PRP → 'to' [p₁₂]

V → 'saw' [p₁₃]

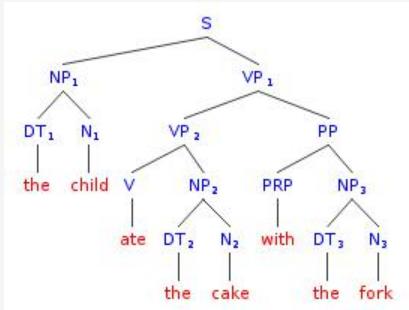
V → 'ate' [p₁₄]

t₁



$$p(t_1) = p_0 p_1 p_4 p_7 p_8 p_{14} p_2 p_1 p_3 p_7 p_9 p_{11} p_1 p_7 p_{10}$$

t₂



$$p(t_2) = p_0 p_1 p_5 p_7 p_8 p_4 p_3 p_{14} p_1 p_{11} p_1 p_7 p_9 p_7 p_{10}$$

Example

$S \rightarrow NP\ VP$ [p₀₌₁]

$NP \rightarrow DT\ N$ [p₁]

$NP \rightarrow NP\ PP$ [p₂]

$PP \rightarrow PRP\ NP$ [p₃₌₁]

$VP \rightarrow V\ NP$ [p₄]

$VP \rightarrow VP\ PP$ [p₅]

$DT \rightarrow 'a'$ [p₆]

$DT \rightarrow 'the'$ [p₇]

$N \rightarrow 'child'$ [p₈]

$N \rightarrow 'cake'$ [p₉]

$N \rightarrow 'fork'$ [p₁₀]

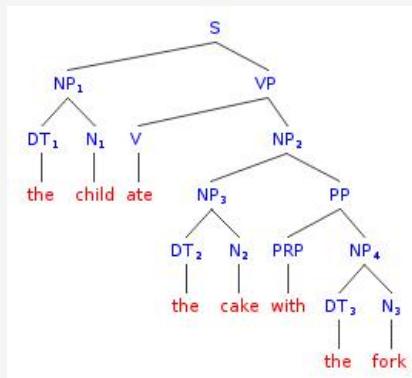
$PRP \rightarrow 'with'$ [p₁₁]

$PRP \rightarrow 'to'$ [p₁₂]

$V \rightarrow 'saw'$ [p₁₃]

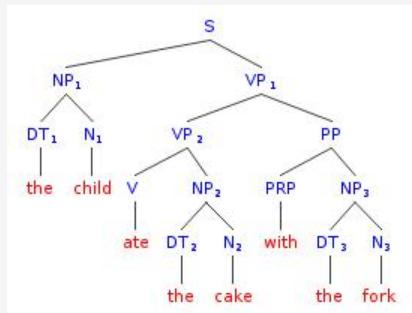
$V \rightarrow 'ate'$ [p₁₄]

t_1



$$p(t_1) = P_0 P_1 P_4 P_7 P_8 P_{14} P_2 P_1 P_3 P_7 P_9 P_{11} P_1 P_7 P_{10}$$

t_2



$$p(t_2) = P_0 P_1 P_5 P_7 P_8 P_4 P_3 P_4 P_1 P_{11} P_1 P_7 P_9 P_7 P_{10}$$

Main Tasks With Pcfgs

- Given a grammar G and a sentence s , let $T(s)$ be all parse trees that correspond to s
- Task 1
 - find which tree t among $T(s)$ maximizes the probability $p(t)$
- Task 2
 - find the probability of the sentence $p(s)$ as the sum of all possible tree probabilities $p(t)$

Probabilistic Parsing Methods

- Probabilistic Earley algorithm
 - Top-down parser with a dynamic programming table
- Probabilistic Cocke–Kasami–Younger (CKY) algorithm
 - Bottom-up parser with a dynamic programming table

Probabilistic Grammars

- Probabilities can be learned from a training corpus (Treebank)
- Intuitive meaning
 - Parse #1 is twice as probable as parse #2
- Possible to do reranking
- Possible to combine with other stages
 - E.g., speech recognition, translation

Maximum Likelihood Estimates

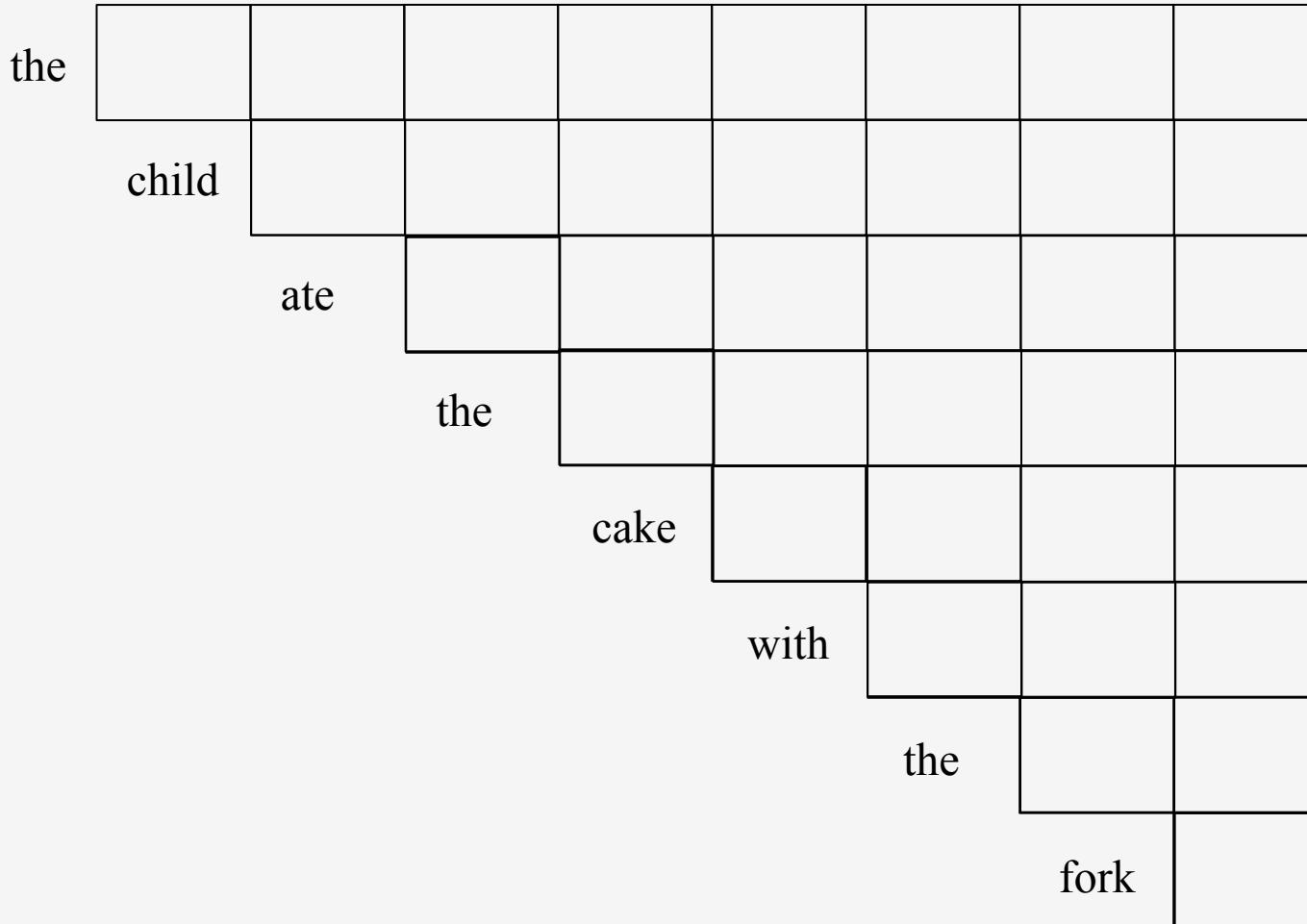
- Use the parsed training set for getting the counts
 - $P_{ML}(\alpha \rightarrow \beta) = \text{Count } (\alpha \rightarrow \beta) / \text{Count}(\alpha)$
- Example:
 - $P_{ML}(S \rightarrow \text{NP VP}) = \text{Count } (S \rightarrow \text{NP VP}) / \text{Count}(S)$

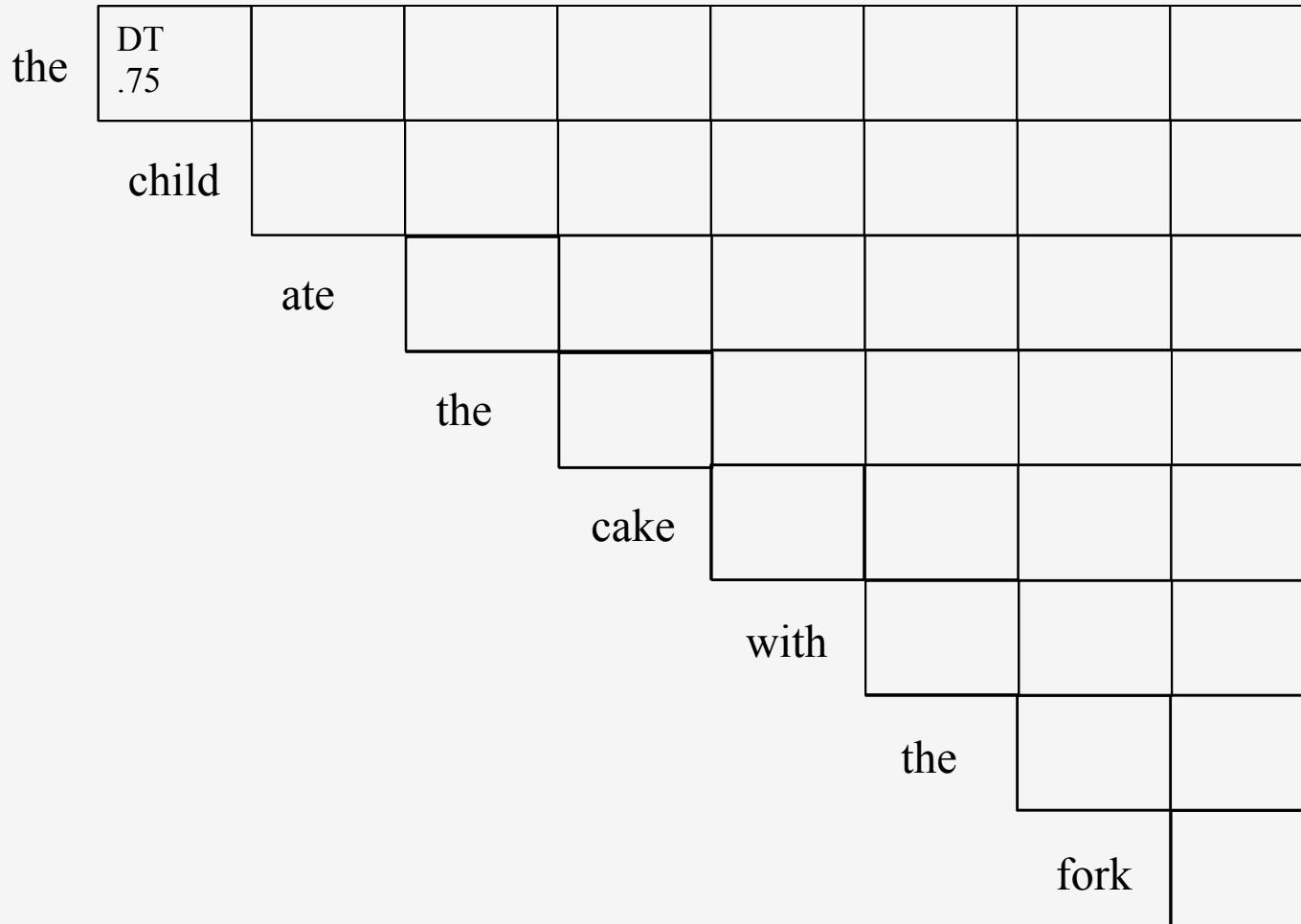
Grammar		Lexicon
$S \rightarrow NP VP$	[.80]	$Det \rightarrow that [.10] \mid a [.30] \mid the [.60]$
$S \rightarrow Aux NP VP$	[.15]	$Noun \rightarrow book [.10] \mid flight [.30]$
$S \rightarrow VP$	[.05]	$\mid meal [.15] \mid money [.05]$
$NP \rightarrow Pronoun$	[.35]	$\mid flights [.40] \mid dinner [.10]$
$NP \rightarrow Proper-Noun$	[.30]	$Verb \rightarrow book [.30] \mid include [.30]$
$NP \rightarrow Det Nominal$	[.20]	$\mid prefer; [.40]$
$NP \rightarrow Nominal$	[.15]	$Pronoun \rightarrow I [.40] \mid she [.05]$
$Nominal \rightarrow Noun$	[.75]	$\mid me [.15] \mid you [.40]$
$Nominal \rightarrow Nominal Noun$	[.20]	$Proper-Noun \rightarrow Houston [.60]$
$Nominal \rightarrow Nominal PP$	[.05]	$\mid NWA [.40]$
$VP \rightarrow Verb$	[.35]	$Aux \rightarrow does [.60] \mid can [.40]$
$VP \rightarrow Verb NP$	[.20]	$Preposition \rightarrow from [.30] \mid to [.30]$
$VP \rightarrow Verb NP PP$	[.10]	$\mid on [.20] \mid near [.15]$
$VP \rightarrow Verb PP$	[.15]	$\mid through [.05]$
$VP \rightarrow Verb NP NP$	[.05]	
$VP \rightarrow VP PP$	[.15]	
$PP \rightarrow Preposition NP$	[1.0]	

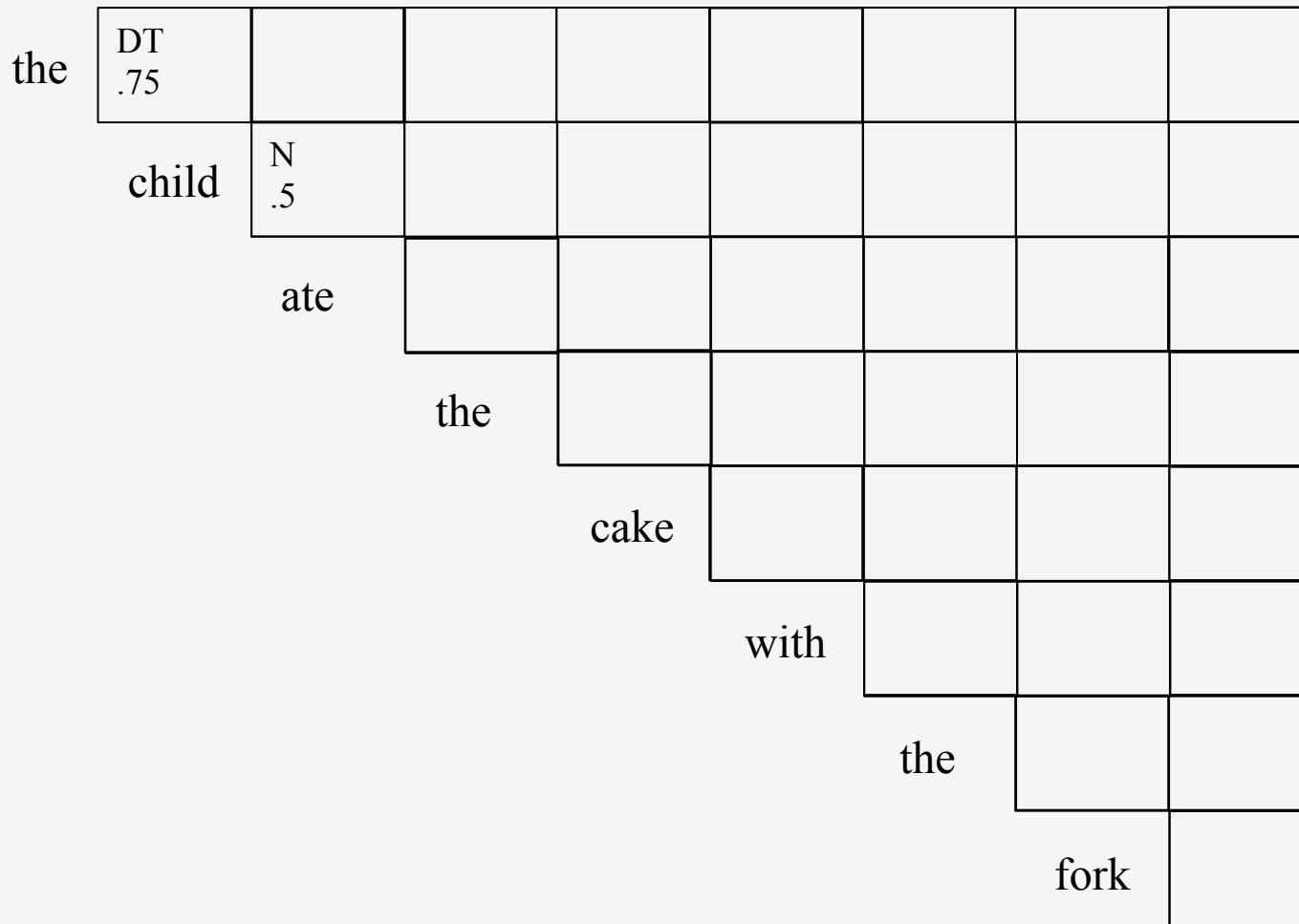
Example from Jurafsky and Martin

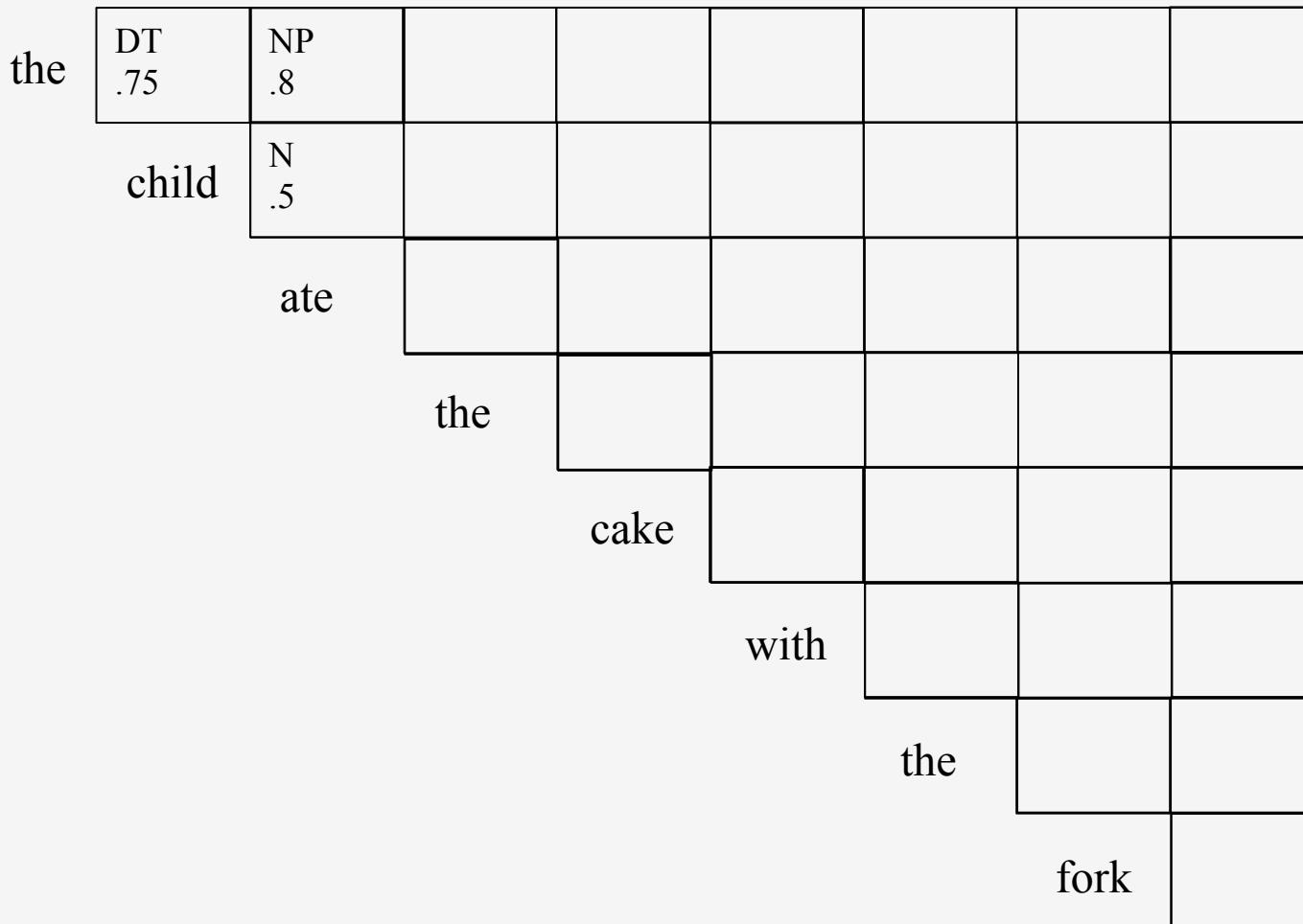
Example

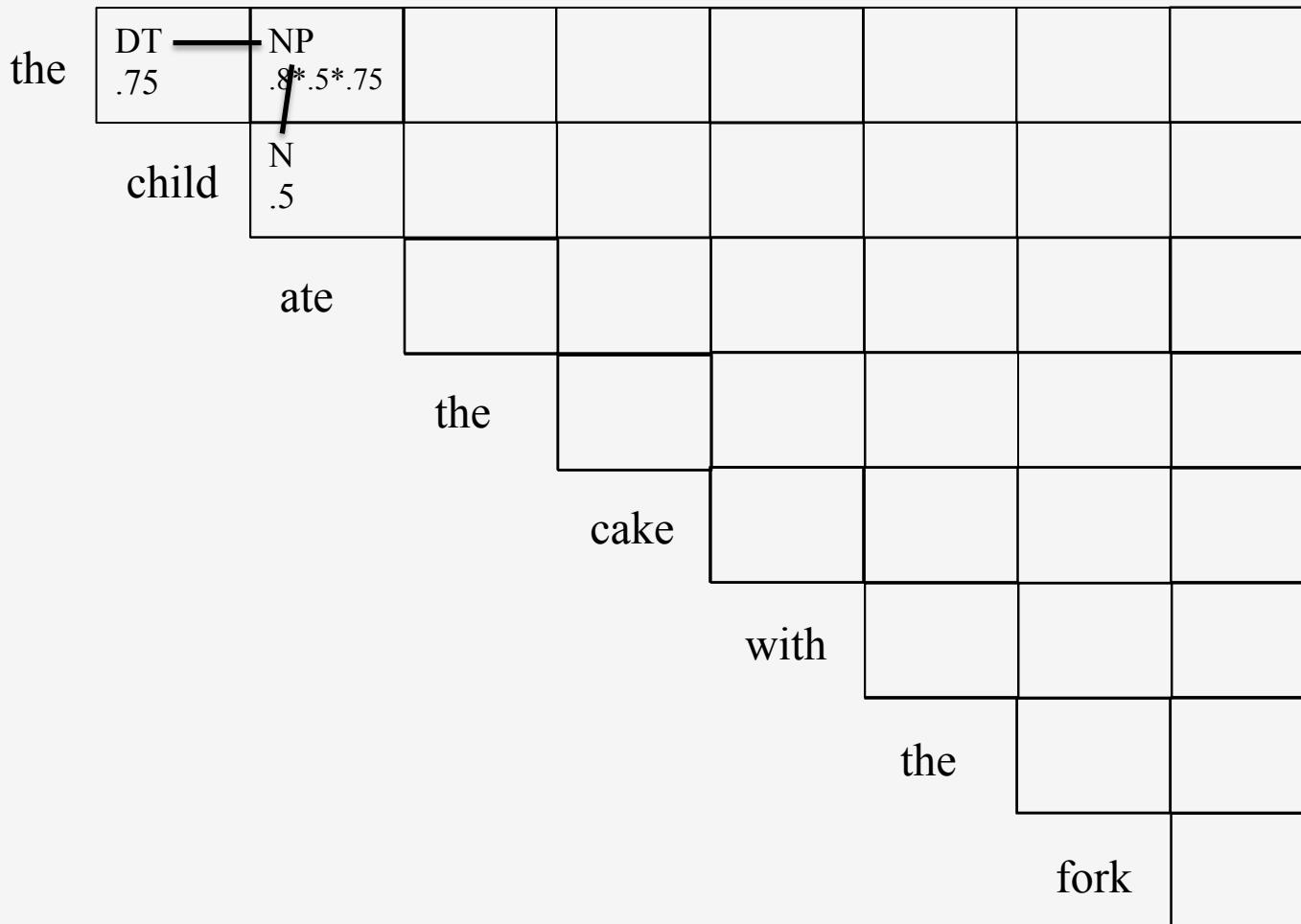
S -> NP VP	[p0=1]
NP -> DT N	[p1=.8]
NP -> NP PP	[p2=.2]
PP -> PRP NP	[p3=1]
VP -> V NP	[p4=.7]
VP -> VP PP	[p5=.3]
DT -> 'a'	[p6=.25]
DT -> 'the'	[p7=.75]
N -> 'child'	[p8=.5]
N -> 'cake'	[p9=.3]
N -> 'fork'	[p10=.2]
PRP -> 'with'	[p11=.1]
PRP -> 'to'	[p12=.9]
V -> 'saw'	[p13=.4]
V -> 'ate'	[p14=.6]











Question

- How, on your own, could you compute the probability of the entire sentence using Probabilistic CKY?
- Don't forget that there may be multiple parses, so you will need to add the corresponding probabilities.

NLP

NLP

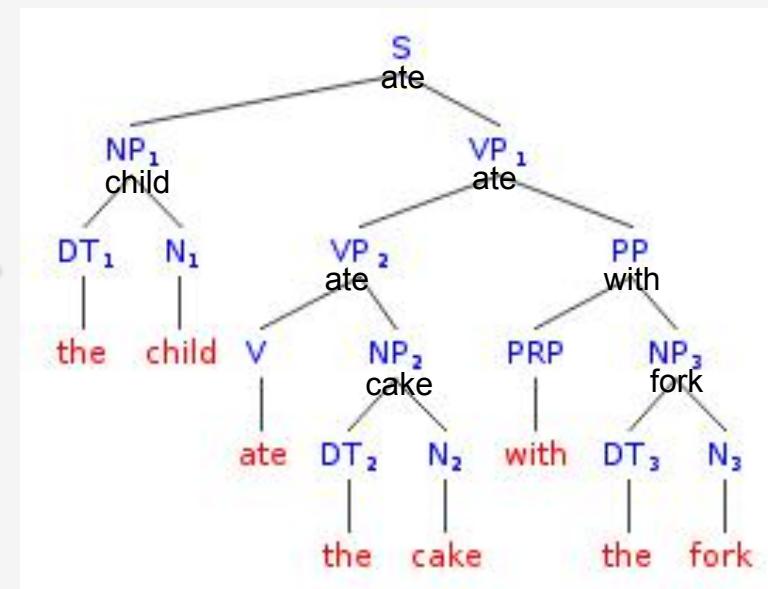
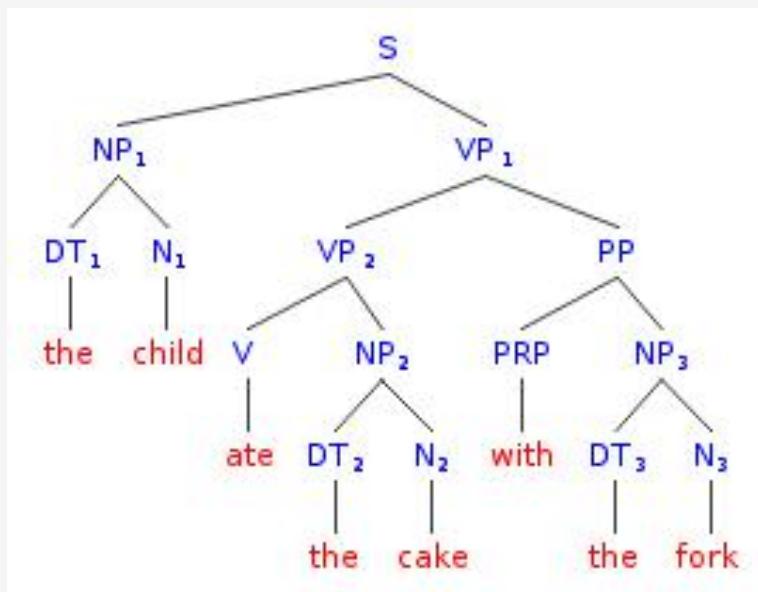
Introduction to NLP

Lexicalized parsing

Limitations of PCFGs

- The probabilities don't depend on the specific words
 - E.g., *give someone something* (2 arguments) vs. *see something* (1 argument)
- It is not possible to disambiguate sentences based on semantic information
 - E.g., *eat pizza with pepperoni* vs. *eat pizza with fork*
- Lexicalized grammars – idea
 - Use the head of a phrase as an additional source of information
 - $\text{VP}[\text{ate}] \rightarrow \text{V}[\text{ate}]$

Lexicalization Example



Collins Parser (1999) 1/2

- Generative, lexicalized model
- Types of rules
 - LHS $\rightarrow L_n L_{n-1} \dots L_1 H R_1 \dots R_{m-1} R_m$
 - H gets generated first
 - L gets generated next
 - R gets generated last

Collins Parser (1999) 2/2

- Maximum likelihood estimates

$$\begin{aligned} P_{ML}(PPof-IN \mid VPthink-VB) = \\ \text{Count (PPof-IN right of the head VPthink-VB)} / \\ \text{Count (symbols right of the head VPthink-VB)} \end{aligned}$$

- Smoothing

$$\begin{aligned} \text{smoothedP (PPof-IN | VPthink-VB)} = \lambda_1 P(PPof-IN \mid VPthink-VB) + \\ + \lambda_2 P(PPof-IN \mid VP-VB) + (1-\lambda_1-\lambda_2) P(PPof-IN \mid VP) \end{aligned}$$

Issues With Lexicalized Grammars

- Sparseness of training data
 - Many probabilities are difficult to estimate from the Penn Treebank
- Combinatorial explosion
 - Need for parameterization

Discriminative Reranking

- A parser may return many parses of a sentence, with small differences in probabilities
- The top returned parse may not necessarily be the best because the PCFG may be deficient
- Other considerations may need to be taken into account
 - parse tree depth
 - left attachment vs. right attachment
 - discourse structure
- Can you think of other features that may affect the reranking?

Answer

- Considerations that may affect the reranking
 - parse tree depth
 - left attachment vs. right attachment
 - discourse structure
- Can you think of others?
 - consistency across sentences
 - or other stages of the NLU pipeline

Discriminative Reranking

- n-best list
 - Get the parser to produce a list of n-best parses (where n can be in the thousands)
- reranking
 - Train a discriminative classifier to rerank these parses based on external information such as a bigram probability score or the amount of right branching in the tree

Sample Performances

- F1 (sentences \leq 40 words)
 - Charniak (2000) – 90.1%
 - Charniak and Johnson (2005) – 92%
(discriminative reranking)

NLP

NLP

Introduction to NLP

Dependency Parsing

Dependency Structure



blue house

- **blue**
 - modifier, dependent, child, subordinate
- **house**
 - head, governor, parent, regent

Dependency Structure

Unionized workers are usually better paid than their non-union counterparts.

1 2 3 4 5 6 7 8 9 10

Dependency Structure



Unionized workers are usually better paid than their non-union counterparts.

1 2 3 4 5 6 7 8 9 10

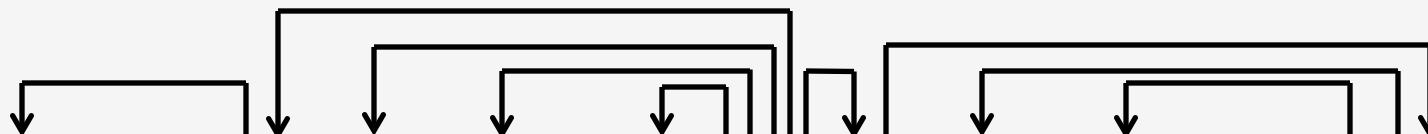
Dependency Structure



Unionized workers are usually better paid than their non-union counterparts.

1 2 3 4 5 6 7 8 9 10

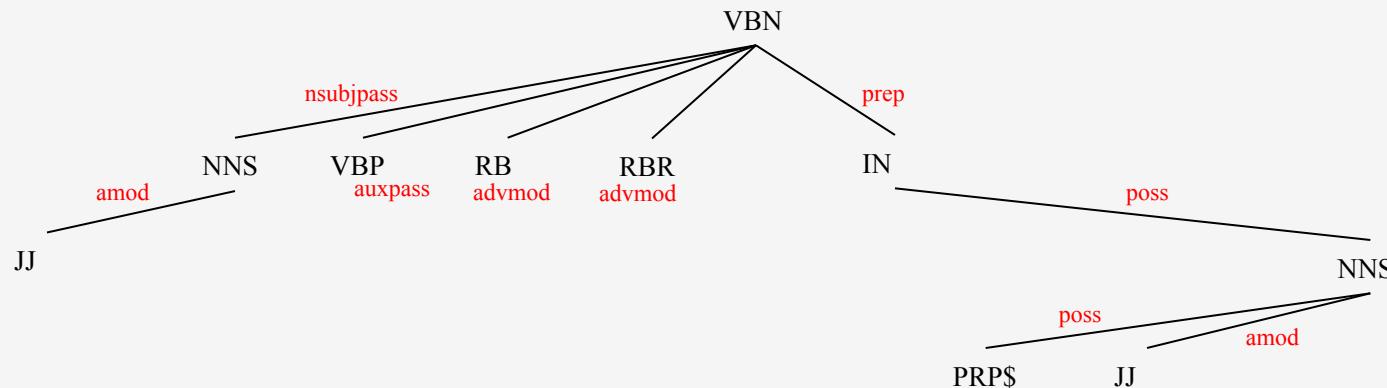
Dependency Structure



Unionized workers are usually better paid than their non-union counterparts.

1 2 3 4 5 6 7 8 9 10

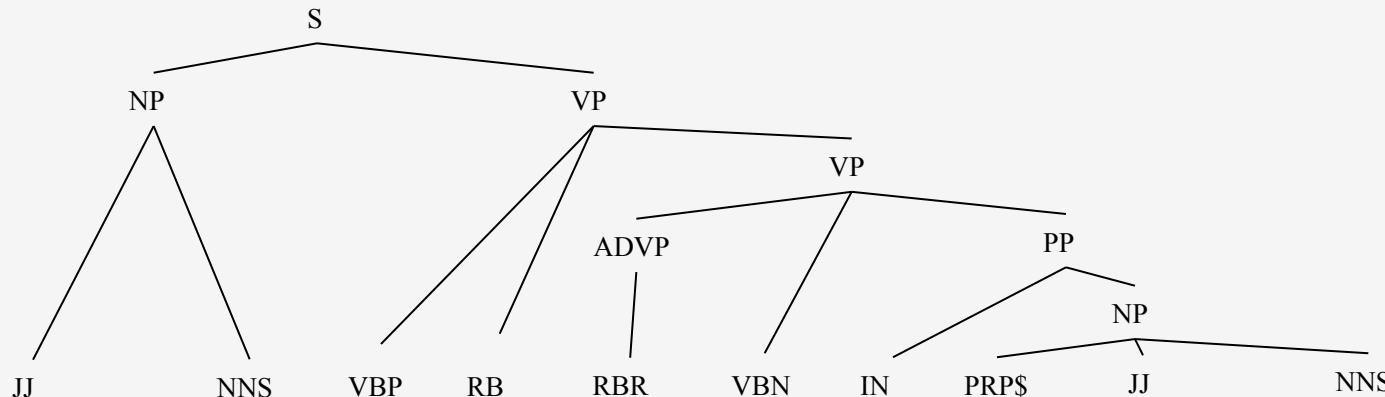
Other Notations



Unionized workers are usually better paid than their non-union counterparts.

1 2 3 4 5 6 7 8 9 10

Phrase Structure



Unionized workers are usually better paid than their non-union counterparts.

1 2 3 4 5 6 7 8 9 10

Dependency Grammars

- Characteristics
 - Lexical/syntactic dependencies between words
 - The top-level predicate of a sentence is the root
 - Simpler to parse than context-free grammars
 - Particularly useful for free word order languages

How To Identify The Heads

- H=head, M=modifier
 - H determines the syntactic category of the construct
 - H determines the semantic category of the construct
 - H is required; M may be skipped
 - Fixed linear position of M with respect to H

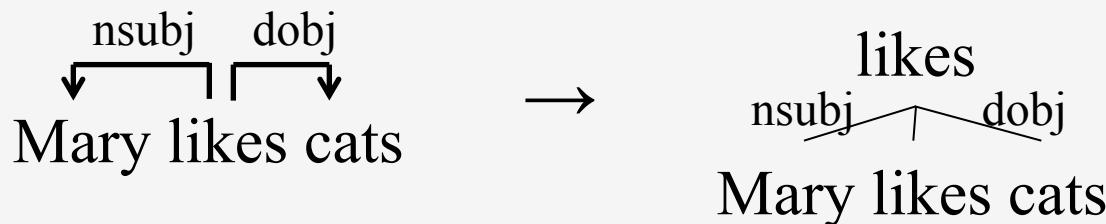
Head Rules From Collins

Parent Non-terminal	Direction	Priority List
ADJP	Left	NNS QP NN \$ ADVP JJ VBN VBG ADJP JJR NP JJS DT FW RBR RBS SBAR RB
ADVP	Right	RB RBR RBS FW ADVP TO CD JJR JJ IN NP JJS NN
CONJP	Right	CC RB IN
FRAG	Right	
INTJ	Left	
LST	Right	LS :
NAC	Left	NN NNS NNP NNPS NP NAC EX \$ CD QP PRP VBG JJ JJS JJR ADJP FW
PP	Right	IN TO VBG VBN RP FW
PRN	Left	
PRT	Right	RP
QP	Left	\$ IN NNS NN JJ RB DT CD NCD QP JJR JJS
RRC	Right	VP NP ADVP ADJP PP
S	Left	TO IN VP S SBAR ADJP UCP NP
SBAR	Left	WHNP WHPP WHADVP WHADJP IN DT S SQ SINV SBAR FRAG
SBARQ	Left	SQ S SINV SBARQ FRAG
SINV	Left	VBZ VBD VBP VB MD VP S SINV ADJP NP
SQ	Left	VBZ VBD VBP VB MD VP SQ
UCP	Right	
VP	Left	TO VBD VBN MD VBZ VB VBG VBP VP ADJP NN NNS NP
WHADJP	Left	CC WRB JJ ADJP
WHADVP	Right	CC WRB
WHNP	Left	WDT WP WP\$ WHADJP WHPP WHNP
WHPP	Right	IN TO FW

Table A.1: The head-rules used by the parser. *Parent* is the non-terminal on the left-hand-side of a rule. *Direction* specifies whether search starts from the left or right end of the rule. *Priority* gives a priority ranking, with priority decreasing when moving down the list.

Techniques (1)

- Dynamic programming
 - CKY – similar to lexicalized PCFG, cubic complexity (Eisner 96)



Techniques (2)

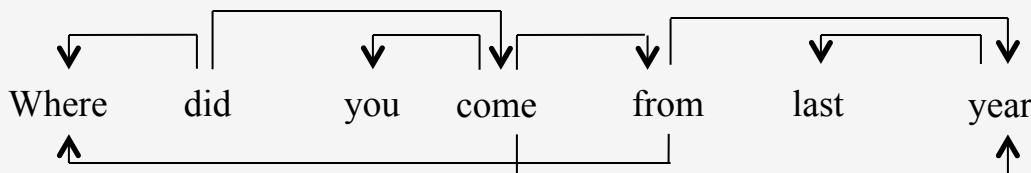
- Constraint-based methods
 - Maruyama 1990, Karlsson 1990
 - Example
 - $\text{word}(\text{pos}(x)) = \text{DET} \Rightarrow (\text{label}(X) = \text{NMOD}, \text{word}(\text{mod}(x)) = \text{NN}, \text{pos}(x) < \text{mod}(x))$
 - A determiner (DET) modifies a noun (NN) on the right with the label NMOD.
 - NP complete problem; heuristics needed
- Constraint graph
 - For initial constraint graph using a core grammar: nodes, domains, constraints
 - Find an assignment that doesn't contradict any constraints. If more than one assignment exists, add more constraints.

Techniques (3)

- Deterministic parsing
 - Covington 2001
 - MaltParser by Nivre
 - shift/reduce as in a shift/reduce parser
 - reduce creates dependencies with the head on either the left or the right
- Graph-based methods
 - Maximum spanning trees (MST)
 - MST Parser by McDonald et al.

Non-projectivity

Output of (the non-projective) MSTParser



1	Where	Where	WRB	WRB	-	2
2	did	did	VBD	VBD	-	0
3	you	you	PRP	PRP	-	4
4	come	come	VBP	VBP	-	2
5	from	from	IN	IN	-	4
6	last	last	JJ	JJ	-	7
7	year	year	NN	NN	-	5
8	?	?	.	.	-	2

Output of Stanford parser

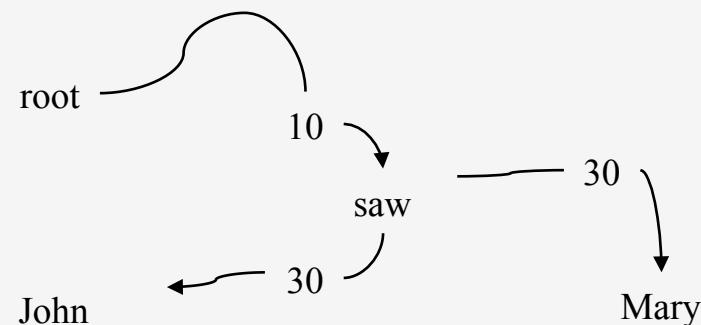
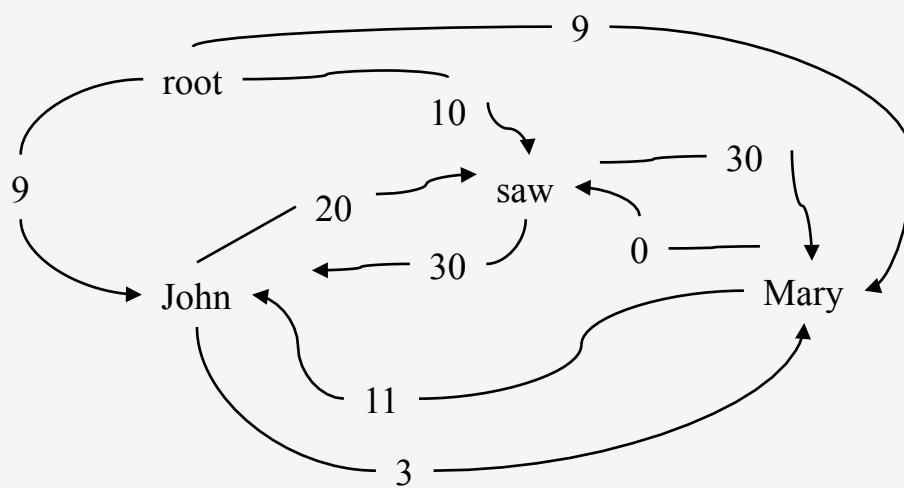
admod(come-4, Where-1)
aux(come-4, did-2)
nsubj(come-4, you-3)
root(ROOT-0, come-4)
prep(come-4, from-5)
amod(year-7, last-6)
pobj(from-5, year-7)

Dependency Parsing

- Background
 - McDonald et al. 2005
- Projectivity
 - English dependency trees are mostly projective (can be drawn without crossing dependencies).
 - Other languages are not.
- Idea
 - Dependency parsing is equivalent to search for a maximum spanning tree in a directed graph.
 - Chu and Liu (1965) and Edmonds (1967) give an efficient algorithm for finding MST for directed graphs.

MST Parser Example

- Consider the sentence “John saw Mary”
- The Chu-Liu-Edmonds algorithm gives the MST on the right hand side (right). This is in general a non-projective tree.



MaltParser (Nivre 2008)

- Very similar to shift-reduce parsing.
- It includes the following components
 - A stack
 - A buffer
 - Set of dependencies (arcs)
- The reduce operations combine an element from the stack and one from the buffer
- Arc-eager parser
 - The actions are shift, reduce, left-arc, right-arc

MaltParser Actions

$$\text{Shift} \quad \frac{[\dots]_S \quad [w_i, \dots]_Q}{[\dots, w_i]_S \quad [\dots]_Q}$$

$$\text{Reduce} \quad \frac{[\dots, w_i]_S \quad [\dots]_Q \quad \exists w_k : w_k \rightarrow w_i}{[\dots]_S \quad [\dots]_Q}$$

$$\text{Left-Arc}_r \quad \frac{[\dots, w_i]_S \quad [w_j, \dots]_Q \quad \neg \exists w_k : w_k \rightarrow w_i}{[\dots]_S \quad [w_j, \dots]_Q \quad w_i \xleftarrow{r} w_j}$$

$$\text{Right-Arc}_r \quad \frac{[\dots, w_i]_S \quad [w_j, \dots]_Q \quad \neg \exists w_k : w_k \rightarrow w_j}{[\dots, w_i, w_j]_S \quad [\dots]_Q \quad w_i \xrightarrow{r} w_j}$$

[Example from Nivre and Kuebler]

Example

- People want to be free
 - [ROOT] [People, want, to, be, free] \emptyset
 - Shift [ROOT, People] [want, to, be, free]
 - LA_{nsubj} [ROOT] [want, to, be, free] $A_1 = \{\text{nsubj(want, people)}\}$
 - RA_{root} [ROOT, want] [to, be, free] $A_2 = A_1 \cup \{\text{root(ROOT, want)}\}$
- The next action is chosen using a classifier
- There is no search
- The final list of arcs is returned as the dependency tree
- Very fast method

Evaluation Metric

- Labeled dependency accuracy
- # correct deps/# deps

1	Unionized	Unionized	VBN	VBN	-	2	NMOD	-	-
2	workers	workers	NNS	NNS	-	3	SBJ	-	-
3	are	are	VBP	VBP	-	0	ROOT	-	-
4	usually	usually	RB	RB	-	3	TMP	-	-
5	better	better	RBR	RBR	-	4	ADV	-	-
6	paid	paid	VBN	VBN	-	5	AMOD	-	-
7	than	than	IN	IN	-	5	AMOD	-	-
8	their	their	PRP\$	PRP\$	-	10	NMOD	-	-
9	non-union	non-union	JJ	JJ	-	10	NMOD	-	-
10	counterparts	counterparts	NNS	NNS	-	7	PMOD	-	-

Complexity

- Projective (CKY) $O(n^5)$
- Projective (Eisner) $O(n^3)$
- Non-projective (MST – Chu-Liu-Edmonds)
 $O(n^2)$
- Projective (Malt) $O(n)$

Use In Information Extraction

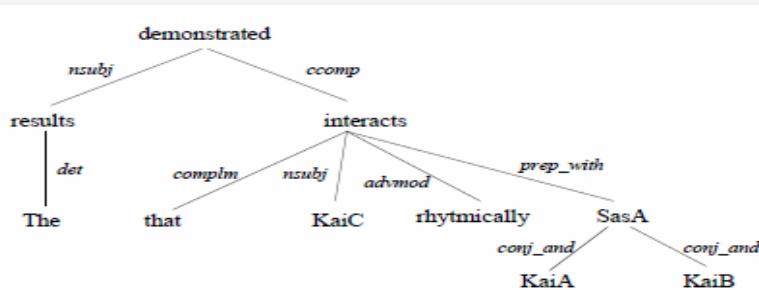


Figure 1: The dependency tree of the sentence “*The results demonstrated that KaiC interacts rhythmically with KaiA, KaiB, and SasA.*”

1. KaiC - nsubj - interacts - prep_with - SasA
 2. KaiC - nsubj - interacts - prep_with - SasA - conj_and - KaiA
 3. KaiC - nsubj - interacts - prep_with - SasA - conj_and - KaiB
 4. SasA - conj_and - KaiA
 5. SasA - conj_and - KaiB
 6. KaiA - conj_and - SasA - conj_and - KaiB
1. *PROTX1* - nsubj - interacts - prep_with - *PROTX2*
 2. *PROTX1* - nsubj - interacts - prep_with - *PROTX0* - conj_and - *PROTX2*
 3. *PROTX1* - nsubj - interacts - prep_with - *PROTX0* - conj_and - *PROTX2*
 4. *PROTX1* - conj_and - *PROTX2*
 5. *PROTX1* - conj_and - *PROTX2*
 6. *PROTX1* - conj_and - *PROTX0* - conj_and - *PROTX2*

Dependency Kernels

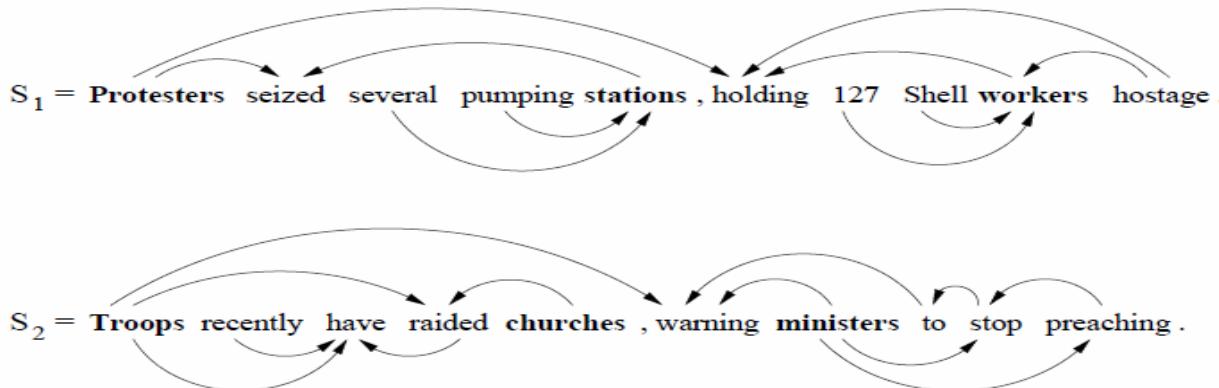


Figure 1: Sentences as dependency graphs.

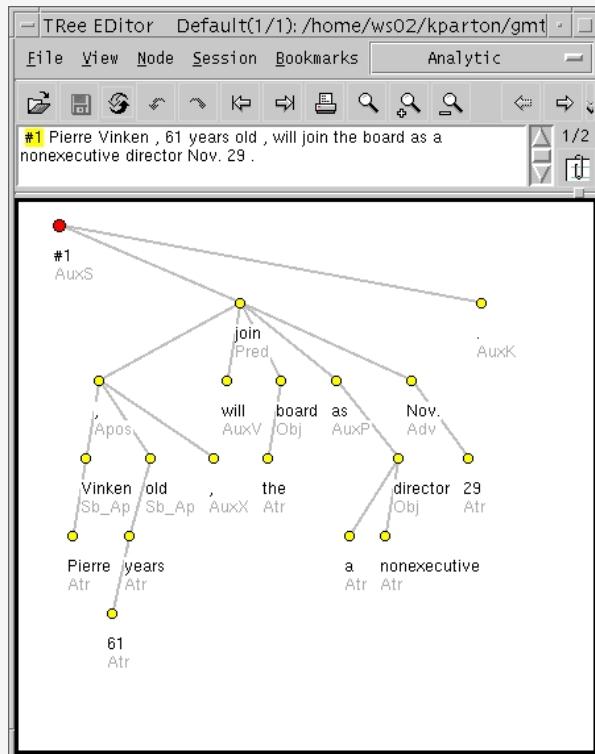
Relation Instance	Shortest Path in Undirected Dependency Graph
S_1 : protesters AT stations	protesters → seized ← stations
S_1 : workers AT stations	workers → holding ← protesters → seized ← stations
S_2 : troops AT churches	troops → raided ← churches
S_2 : ministers AT churches	ministers → warning ← troops → raided ← churches

Table 1: Shortest Path representation of relations.

External Links

- <http://ilk.uvt.nl/conll/>
 - CONLL-X Shared task
- <http://ufal.mff.cuni.cz/pdt2.0/>
 - Prague Dependency Treebank
- <http://nextens.uvt.nl/depparse-wiki/SharedTaskWebsite>
- <http://nextens.uvt.nl/depparse-wiki/DataOverview>
- <http://maltparser.org/>
 - Joakim Nivre's Maltparser
- <http://www.cs.ualberta.ca/~lindek/minipar.htm>
 - Dekang Lin's Minipar
- <http://www.link.cs.cmu.edu/link/>
 - Daniel Sleator and Davy Temperley's Link parser

Prague Dependency Treebank Example



NLP

NLP

Introduction to NLP

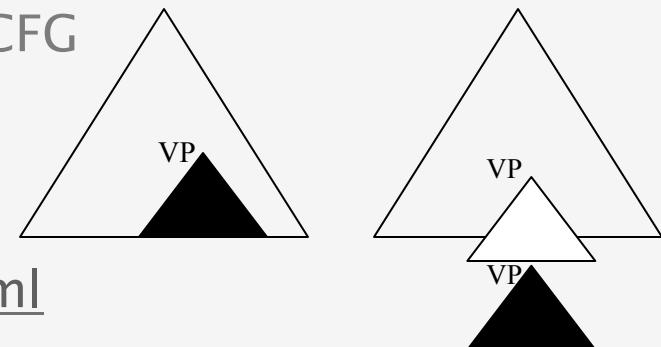
Alternative Syntactic Formalisms

Mildly Context-Sensitive Grammars

- Tree Substitution Grammar (TSG)
 - Terminals generate entire tree fragments
 - TSG and CFG are formally equivalent
- Tree Adjoining Grammar (TAG)
- Combinatory Categorial Grammar (CCG)

Tree Adjoining Grammar (TAG)

- Like TSG but allow adjunction
- It can generate languages like $a^n b^n c^n$ or ww (cross-serial dependencies):
 - e.g., Mary gave a book and a magazine to Chen and Mike, respectively.
- Expressive power
 - TAG is formally more powerful than CFG
 - TAG is less powerful than CSG
- Card game online!
 - <http://www.ltaggame.com/>
 - <http://www.ltaggame.com/family.html>



Combinatory Categorial Grammar (CCG)

- **Complex types**

- E.g., X/Y and $X|Y$
- These take an argument of type Y and return an object of type X .
- X/Y – means that Y should appear on the right
- $X\backslash Y$ – means that Y should appear on the left

I	NP
books	NP
sleep	$S \backslash NP$
enjoy	$(S \backslash NP) / NP$



- **Expressive power**

- CCGs can generate the language $a^n b^n c^n d^n$, $n > 0$

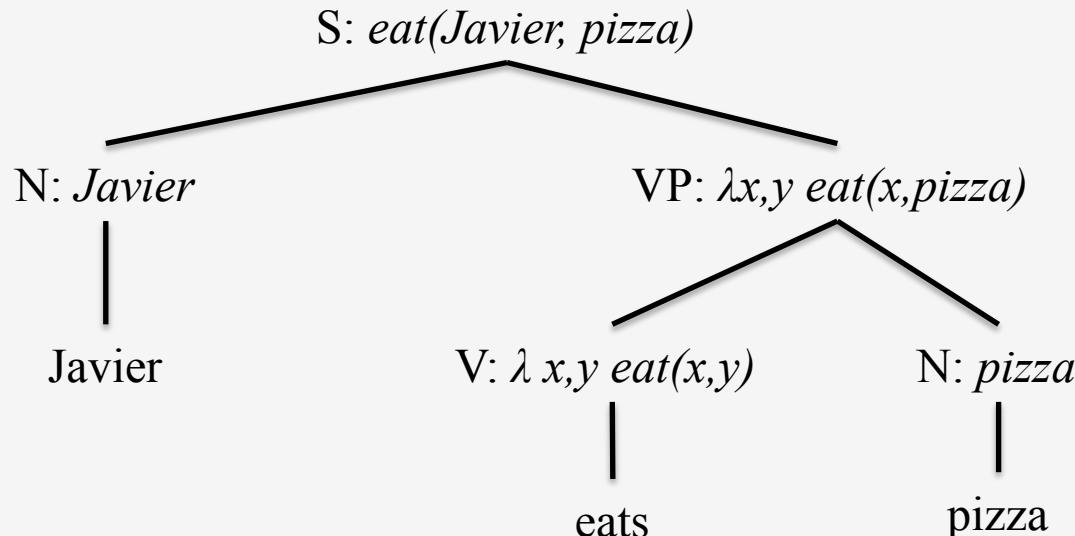
Example from Jonathan Kummerfeld, Aleka Blackwell, and Patrick Littell

Introduction to NLP

Semantic parsing

Semantic Parsing

- Associate a semantic expression with each node



Introduction to NLP

NACLO problems on parsing

External Links

- Twodee (by Jason Eisner)
 - <http://nacloweb.org/resources/problems/2013/N2013-H.pdf>
- One, Two, Tree (by Noah Smith, Kevin Gimbel, and Jason Eisner)
 - <http://www.nacloweb.org/resources/problems/2012/N2012-R.pdf>
- CCG (by Jonathan Kummerfeld, Aleka Blackwell, and Patrick Littell)
 - <http://www.nacloweb.org/resources/problems/2014/N2014-O.pdf>
- Combining categories in Tok Pisin (same authors)
 - <http://www.nacloweb.org/resources/problems/2014/N2014-P.pdf>
- Grammar Rules (Andrea Schalley and Pat Littell)
 - <http://www.nacloweb.org/resources/problems/2013/N2013-F.pdf>
- Sk8 Parser (Pat Littell)
 - <http://www.nacloweb.org/resources/problems/2009/N2009-G.pdf>

NLP

NLP

Introduction to NLP

Probabilities

Probabilistic Reasoning

- Very important for language processing
- Example in speech recognition:
 - “recognize speech” vs “wreck a nice beach”
- Example in machine translation:
 - “l'avocat général”: “the attorney general” vs. “the general avocado”
- Probabilities make it possible to combine evidence from multiple sources in a systematic way.

Probabilities

- Probability theory
 - predicting how likely it is that something will happen
- Experiment (trial)
 - e.g., throwing a coin
- Possible outcomes
 - heads or tails
- Sample spaces
 - discrete or continuous
- Events
 - Ω is the certain event
 - \emptyset is the impossible event
 - event space – all possible events

Probabilities

- **Probabilities**
 - numbers between 0 and 1
- **Probability distribution**
 - distributes a probability mass of 1 throughout the sample space Ω .
- **Example:**
 - A fair coin is tossed three times.
 - What is the probability of 3 heads?
 - What is the probability of 2 heads?

Meaning of Probabilities

- Frequentist
 - I threw the coin 10 times and it turned up heads 5 times
- Subjective
 - I am willing to bet 50 cents on heads

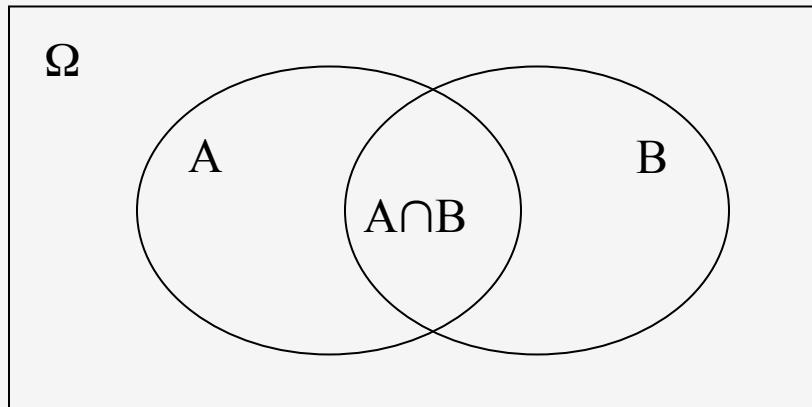
Properties of Probabilities

- $p(\emptyset) = 0$
- $P(\text{certain event})=1$
- $p(X) \leq p(Y)$, if $X \subseteq Y$
- $p(X \cup Y) = p(X) + p(Y)$, if $X \cap Y = \emptyset$

Conditional Probability

- Prior and posterior probability
- Conditional probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$



Conditional Probability

- Six-sided fair die
 - $P(D \text{ even})=?$
 - $P(D >= 4) = ?$
 - $P(D \text{ even} | D >= 4) = ?$
 - $P(D \text{ odd} | D >= 4) = ?$
- Multiple conditions
 - $P(D \text{ odd} | D >= 4, D <= 5) = ?$

Conditional Probability

- Six-sided fair die
 - $P(D \text{ even}) = 3/6 = 1/2$
 - $P(D >= 4) = 3/6 = 1/2$
 - $P(D \text{ even} | D >= 4) = 2/3$
 - $P(D \text{ odd} | D >= 4) = 1/3$
- Multiple conditions
 - $P(D \text{ odd} | D >= 4, D <= 5) = 1/2$

The Chain Rule

- $P(w_1, w_2, w_3 \dots w_n) = ?$
- Using the chain rule:
 - $P(w_1, w_2, w_3 \dots w_n) = P(w_1) P(w_2 | w_1) P(w_3 | w_1, w_2) \dots P(w_n | w_1, w_2 \dots w_{n-1})$
- This rule is used in many ways in statistical NLP, more specifically in Markov Models

Independence

- Two events are independent when
 - $P(A \cap B) = P(A)P(B)$
- Unless $P(B)=0$ this is equivalent to saying that $P(A) = P(A|B)$
- If two events are not independent, they are considered dependent

Adding vs. Removing Constraints

- Adding constraints
 - $P(\text{walk}=\text{yes} \mid \text{weather}=\text{nice})$
 - $P(\text{walk}=\text{yes} \mid \text{weather}=\text{nice}, \text{freetime}=\text{yes}, \text{crowded}=\text{yes})$
 - More accurate
 - But more difficult to estimate
- Removing constraints (Backoff)
 - $P(\text{walk}=\text{yes} \mid \text{weather}=\text{nice}, \text{freetime}=\text{yes}, \text{crowded}=\text{yes})$
 - $P(\text{walk}=\text{yes} \mid \text{weather}=\text{nice}, \text{freetime}=\text{yes})$
 - $P(\text{walk}=\text{yes} \mid \text{weather}=\text{nice})$
 - Note that it is *not* possible to do backoff on the left hand side of the conditional

Random Variables

- Simply a function: $X: \Omega \rightarrow \mathbb{R}^n$
- The numbers are generated by a *stochastic process* with a certain probability distribution
- Example
 - the discrete random variable X that is the sum of the faces of two randomly thrown fair dice
- Probability mass function (pmf) which gives the probability that the random variable has different numeric values: $P(x) = P(X = x) = P(A_x)$ where $A_x = \{\omega \in \Omega : X(\omega) = x\}$

Random Variables

- If a random variable X is distributed according to the pmf $p(x)$, then we write $X \sim p(x)$
- For a discrete random variable, we have

$$\sum p(x_i) = P(\Omega) = 1$$

Example

- $p(1) = 1/6$
- $p(2) = 1/6$
- etc.
- $P(D)=?$
- $P(D) = \{1/6, 1/6, 1/6, 1/6, 1/6, 1/6\}$
- $P(D|odd) = \{1/3, 0, 1/3, 0, 1/3, 0\}$

NLP

NLP

Introduction to NLP

Bayes' Theorem

Bayes' Theorem

- Formula for joint probability
 - $p(A,B) = p(B|A)p(A)$
 - $p(A,B) = p(A|B)p(B)$
- Therefore
 - $p(B|A)=p(A|B)p(B)/p(A)$
- Bayes' theorem is used to calculate $P(A|B)$ given $P(B|A)$

Example

- Diagnostic test
- Test accuracy
 - $P(\text{positive} \mid \neg \text{disease}) = 0.05$ – false positive
 - $P(\text{negative} \mid \text{disease}) = 0.05$ – false negative
 - so $p(\text{positive} \mid \text{disease}) = 1 - 0.05 = 0.95$

Example

- Diagnostic test with errors

$P(A B)$		A=TEST	
		Positive	Negative
B=DISEASE	Yes	0.95	0.05
	No	0.05	0.95

Example

- What is $p(\text{disease} \mid \text{positive})$?
 - $P(\text{disease}|\text{positive}) = P(\text{positive}|\text{disease}) * P(\text{disease}) / P(\text{positive})$
 - $P(\neg\text{disease}|\text{positive}) = P(\text{positive}|\neg\text{disease}) * P(\neg\text{disease}) / P(\text{positive})$
 - $P(\text{disease}|\text{positive}) / P(\neg\text{disease}|\text{positive}) = ?$
- We don't really care about $p(\text{positive})$
 - as long as it is not zero, we can divide by it on both sides

Example

- $P(\text{disease}|\text{positive})/P(\neg\text{disease}|\text{positive}) = (P(\text{positive}|\text{disease}) \times P(\text{disease}))/ (P(\text{positive}|\neg\text{disease}) \times P(\neg\text{disease}))$
- Suppose $P(\text{disease}) = 0.001$
 - so $P(\neg\text{disease}) = 0.999$
- $P(\text{disease}|\text{positive})/P(\neg\text{disease}|\text{positive}) = (0.95 \times 0.001)/(0.05 \times 0.999) = 0.019$
- $P(\text{disease}|\text{positive}) + P(\neg\text{disease}|\text{positive}) = 1$
- $P(\text{disease}|\text{positive}) \approx 0.02$
- $P(\text{disease})$ is called the prior probability
- $P(\text{disease}|\text{positive})$ is called the posterior probability
- In this example the posterior is 20 times larger than the prior

NLP

NLP

Introduction to NLP

Language models (Part 1)

Probabilistic Language Models

- Assign a probability to a sentence
 - $P(S) = P(w_1, w_2, w_3, \dots, w_n)$
- Different from deterministic methods using CFG
- The sum of the probabilities of all possible sentences must add up to 1

Predicting The Next Word

- Example
 - Let's meet in Times ...
 - General Electric has lost some market ...
- Formula
 - $P(w_n | w_1, w_2, \dots, w_{n-1})$

Predicting the Next Word

- What word follows “your”?

- [http://norvig.com/ngrams/
count_2w.txt](http://norvig.com/ngrams/count_2w.txt)

• your abilities	160848
your ability	1116122
your ablum	112926
your academic	274761
your acceptance	783544
your access	492555
your accommodation	320408
your account	8149940
your accounting	128409
your accounts	257118
your action	121057

your actions	492448
your activation	459379
your active	140797
your activities	226183
your activity	156213
your actual	302488
your ad	1450485
your address	1611337
your admin	117943
your ads	264771
your advantage	242238
your adventure	109658
your advert	101178
your advertisement	172783

Uses of Language Models

- Speech recognition
 - $P(\text{"recognize speech"}) > P(\text{"wreck a nice beach"})$
- Text generation
 - $P(\text{"three houses"}) > P(\text{"three house"})$
- Spelling correction
 - $P(\text{"my cat eats fish"}) > P(\text{"my xat eats fish"})$
- Machine translation
 - $P(\text{"the blue house"}) > P(\text{"the house blue"})$
- Other uses
 - OCR
 - Summarization
 - Document classification
- Usually coupled with a translation model (later)

Probability Of A Sentence

- How to compute the probability of a sentence?
- What if the sentence is novel?
- What we need to estimate:
 - $P(S) = P(w_1, w_2, w_3 \dots w_n)$
- Using the chain rule:
 - $P(S) = P(w_1) P(w_2 | w_1) P(w_3 | w_1, w_2) \dots P(w_n | w_1, w_2 \dots w_{n-1})$
- Example:
 - $P(\text{"I would like the pepperoni and spinach pizza"}) = ?$

N-gram Models

- Predict the probability of a word based on the words before:
 - $P(\text{square} | \text{Let's meet in Times})$
- Markov assumption
 - Only look at limited history
- N-gram models
 - Unigram – no context: $P(\text{square})$
 - Bigram: $P(\text{square} | \text{Times})$
 - Trigram: $P(\text{square} | \text{in Times})$

Random Text (Brown Corpus)

- 2-grams:

The 53-year-old Shea was no acceptable formula to help the abuse of events were a wall in 1908 , called upon his hand in Southern New Orleans , Miss Garson was named Maurice Couve De Havilland signed a privilege resolution had had happened on a tax applied to the Chisholm , the thriving systems of the `` Pride and musician , and Moscow made good team spirit of the culmination of the metal tube through the amateur , but rather than a special prosecutor . This knowledge of each member of these savings of golf course can see the 13 straight 69 . Since 1927 by Harry Truman Cleveland of railroad retirement age groups . No Vacancy " . `` I have to congressmen . The remainder of the rear bumper and on a benefit in U.S. amateur , as far as a thrill a \$100 U.S. if not indicted . The state's occupation tax dollars over the newest product of the address he attended Arlington State University will pay half years .

Random Text (Brown Corpus)

- 3-grams:

The Fulton County Jail and `` a very strong central government of Laos that the presence of picket lines and featuring a flared skirt and lace jacket with bateau neckline and princesse skirt accented by lace appliques . Her acting began with the members of the government -- such control is necessary to build in a final exchange between Moscow and Washington last week . Of course , since the views of another one . It urged that the games are not essential to provide federal contributions to the 85-student North Carolina group to play , was addressing a meeting in the manufacture of a tax bill since most of his uncle and aunt , also was particularly struck by the reams came in from shareholders of these co-operative systems , the 9th precinct of the guiding spirits of the Armed Services Committee . Davis received 1,119 votes in Saturday's election , the executive organs of participation can hardly escape the impression that he made no attempt to get agreement among the conference's top four in rushing , he was awarded the top but that presently we're not acting as we head that way .

Random Text (Brown Corpus)

- 4-grams:

The broadcast said Anderson , a Seattle ex-marine and Havana businessman , and McNair , of Miami , were condemned on charges of smuggling arms to Cuban rebels . Anderson operated three Havana automobile service stations and was commander of the Havana American Legion post before it disbanded since the start of August have shown gains averaging nearly 10% above last year . That , too , in improving motorists' access to many turnpikes . The Kansas Turnpike offers an illustration . Net earnings of that road rose from 62 per cent of the prices that the avid buyers bid it up to . Dallas and North Texas is known world-wide as the manufacturing and distribution center of cotton gin machinery and equipment . The firm is design-conscious , sales-conscious , advertising-conscious . `` Hodges predicted : ' I think we should certainly follow through on it " . Rep. Henry C. Grover , who teaches history in the Houston public schools , would reduce from 24 to 12 semester hours the so-called `` blue law " controversy came in the form of a letter to Mayor Grady that plowing and salting crews should be dispatched earlier in storms and should be kept on the job .

Higher Order n-grams

- It is possible to go to 3,4,5-grams
- Longer n-grams suffer from sparseness

N-grams

- Shakespeare unigrams
 - 29,524 types, approx. 900K tokens
- Bigrams
 - 346,097 types, approx. 900K tokens
- Notice – very sparse data!

Estimation

- Can we compute the conditional probabilities directly?
 - No, because the data is sparse
- Markov assumption
 - $P(\text{"musical} \mid \text{I would like two tickets for the"}) = P(\text{"musical} \mid \text{the"})$
 - or
 - $P(\text{"musical} \mid \text{I would like two tickets for the"}) = P(\text{"musical} \mid \text{for the"})$

Maximum Likelihood Estimates

- Use training data
- Count how many times a given context appears in it.
- Unigram example:
 - The word “pizza” appears 700 times in a corpus of 10,000,000 words.
 - Therefore the MLE for its probability is $P'(\text{“pizza”}) = 700/10,000,000 = 0.00007$
- Bigram example:
 - The word “with” appears 1,000 times in the corpus.
 - The phrase “with spinach” appears 6 times
 - Therefore the MLE for $P'(\text{spinach}|\text{with}) = 6/1,000 = 0.006$
- These estimates may not be good for corpora from other genres

Example

- $P(" <S> \text{ I will see you on Monday } </S>") =$
 $P(\text{I}|<S>)$
 $\times P(\text{will}|\text{I})$
 $\times P(\text{see}|\text{will})$
 $\times P(\text{you}|\text{see})$
 $\times P(\text{on}|\text{you})$
 $\times P(\text{Monday}|\text{on})$
 $\times P(</S>|\text{Monday})$

Example from Jane Austen

- $P(\text{"Elizabeth looked at Darcy"})$
- Use maximum likelihood estimates for the n-gram probabilities
 - unigram: $P(w_i) = c(w_i)/V$
 - bigram: $P(w_i|w_{i-1}) = c(w_{i-1}, w_i)/c(w_{i-1})$
- Values
 - $P(\text{"Elizabeth"}) = 474/617091 = .000768120$
 - $P(\text{"looked|Elizabeth"}) = 5/474 = .010548523$
 - $P(\text{"at|looked"}) = 74/337 = .219584569$
 - $P(\text{"Darcy|at"}) = 3/4055 = .000739827$
- Bigram probability
 - $P(\text{"Elizabeth looked at Darcy"}) = .000000001316 = 1.3 \times 10^{-9}$
- Unigram probability
 - $P(\text{"Elizabeth looked at Darcy"}) = 474/617091 * 337/617091 * 4055/617091 * 304/617091 = .00000000001357 = 1.3 \times 10^{-12}$
- $P(\text{"looked Darcy Elizabeth at"}) = ?$

NLP

NLP

N-grams And Regular Languages

- N-grams are just one way to represent weighted regular languages
- More about this in the lecture on regular languages

Generative Models

- Unigram: generate a word, then generate the next one, until you generate </S>.



- Bigram: generate <S>, generate a word, then generate the next one based on the previous one, etc., until you generate </S>.



Engineering Trick

- The MLE values are often on the order of 10^{-6} or less
 - Multiplying 20 such values gives a number on the order of 10^{-120}
 - This leads to underflow
- Use (base 10) logarithms instead
 - 10^{-6} becomes -6
 - Use sums instead of products

NLP

NLP

Introduction to NLP

Language models (Part 2)

Smoothing

- If the vocabulary size is $|V|=1M$
 - Too many parameters to estimate even a unigram model
 - MLE assigns values of 0 to unseen (yet not impossible) data
 - Let alone bigram or trigram models
- Smoothing (regularization)
 - Reassigning some probability mass to unseen data

Smoothing

- How to model novel words?
 - Or novel bigrams?
- Distributing some of the probability mass to allow for novel events
- Add-one (Laplace) smoothing:
 - Bigrams: $P(w_i|w_{i-1}) = (c(w_{i-1}, w_i) + 1)/(c(w_{i-1}) + V)$
 - This method reassigned too much probability mass to unseen events
- Possible to do add-k instead of add-one
- Both of these don't work well in practice

Advanced Smoothing

- Good-Turing
 - Try to predict the probabilities of unseen events based on the probabilities of seen events
- Kneser-Ney
- Class-based n-grams

Example

- Corpus:
 - cat dog cat rabbit mouse fish fish mouse hamster hamster fish turtle tiger cat rabbit cat dog dog fox lion
- What is the probability the next item is “mouse”?
 - $P_{MLE}(\text{mouse}) = 2/20$
- What is the probability the next item is “elephant” or some other previously unseen animal?
 - Trickier
 - Is it 0/20?
 - Note that $P(\text{that the next animal is unseen}) > 0$
 - Therefore we need to discount the probabilities of the animals that have already been seen
 - $P_{MLE}(\text{mouse}) < 2/20$

Good Turing

- Actual counts c
- $N_r =$ number of n-grams that occur exactly c times in the corpus
- $N_0 =$ total number of n-grams in the corpus
- Revised counts c^*
 - $c^* = (c+1) N_{c+1} / N_c$

Example

- Corpus:
 - cat dog cat rabbit mouse fish fish mouse hamster hamster fish turtle tiger cat rabbit cat dog dog fox lion
- Counts
 - $C(\text{cat}) = 4$
 - $C(\text{dog}) = 3$
 - $C(\text{fish}) = 3$
 - $C(\text{mouse}) = 2$
 - $C(\text{rabbit}) = 2$
 - $C(\text{hamster}) = 2$
 - $C(\text{fox}) = 1$
 - $C(\text{turtle}) = 1$
 - $C(\text{tiger}) = 1$
 - $C(\text{lion}) = 1$
- N1=4, N2=3, N3=2, N4=1

Example (cont'd)

- $N_1=4, N_2=3, N_3=2, N_4=1$
- Revised counts $c^* = (c+1) N_{c+1} / N_c$
 - $C^*(\text{cat}) = 4$
 - $C^*(\text{dog}) = (3+1) \times 1/2 = 2$
 - $C^*(\text{mouse}) = (2+1) \times 2/3 = 2$
 - $C^*(\text{rabbit}) = (2+1) \times 2/3 = 2$
 - $C^*(\text{hamster}) = (2+1) \times 2/3 = 2$
 - $C^*(\text{fox}) = (1+1) \times 3/4 = 6/4$
 - $C^*(\text{turtle}) = (1+1) \times 3/4 = 6/4$
 - $C^*(\text{tiger}) = (1+1) \times 3/4 = 6/4$
 - $C^*(\text{lion}) = (1+1) \times 3/4 = 6/4$
 - $C^*(\text{elephant}) = N_1/N = 4/20$
- Note that these counts don't necessarily add to 1, so they still need to be normalized.
 - $P^*(\text{lion}) = 6/4 / 20 = 6/80$

Dealing with Sparse Data

- Two main techniques used
 - Backoff
 - Interpolation

Backoff

- Going back to the lower-order n-gram model if the higher-order model is sparse (e.g., frequency ≥ 1)
- Learning the parameters
 - From a development data set

Interpolation

- If $P'(w_i | w_{i-1}, w_{i-2})$ is sparse:
 - Use $\lambda_1 P'(w_i | w_{i-1}, w_{i-2}) + \lambda_2 P'(w_i | w_{i-1}) + \lambda_3 P'(w_i)$
- Better than backoff
- See [Chen and Goodman 1998] for more details

NLP

Introduction to NLP

Language models (Part 3)

Evaluation of LM

- Extrinsic
 - Use in an application
- Intrinsic
 - Cheaper
- Correlate the two for validation purposes

Perplexity

- Does the model fit the data?
 - A good model will give a high probability to a real sentence
- Perplexity
 - Average branching factor in predicting the next word
 - Lower is better (lower perplexity -> higher probability)
 - N = number of words

$$Per = \sqrt[N]{\frac{1}{P(w_1 w_2 \dots w_N)}}$$

Perplexity

- Example:
 - A sentence consisting of N equiprobable words: $p(w_i) = 1/k$
$$Per = \sqrt[N]{\frac{1}{P(w_1 w_2 \dots w_N)}}$$
 - $Per = (1/k)^N \wedge (-1/N) = k$
- Perplexity is like a branching factor
- Logarithmic version:
$$Per = 2^{-(1/N) \sum \log_2 P(w_i)}$$

The Shannon Game

- Consider the Shannon game:
 - New York governor Andrew Cuomo said ...
- What is the perplexity of guessing a digit if all digits are equally likely?
 - 10
- How about a letter?
 - 26
- How about guessing A (“operator”) with a probability of 1/4, B (“sales”) with a probability of 1/4 and 10,000 other cases with a probability of 1/2 total (example modified from Joshua Goodman).

Perplexity Across Distributions

- What if the actual distribution is very different from the expected one?
- Example:
 - All of the 10,000 other cases are equally likely but $P(A) = P(B) = 0$.
- Cross-entropy = \log (perplexity), measured in bits

$$H(p, q) = - \sum_x p(x) \log q(x).$$

Sample Values for Perplexity

- Wall Street Journal (WSJ) corpus
 - 38 M words (tokens)
 - 20 K types
- Perplexity
 - Evaluated on a separate 1.5M sample of WSJ documents
 - Unigram 962
 - Bigram 170
 - Trigram 109

Word Error Rate

- Another evaluation metric
 - Number of insertions, deletions, and substitutions
 - Normalized by sentence length
 - Same as Levenshtein Edit Distance
- Example:
 - governor Andrew Cuomo met with the mayor
 - the governor met the senator
 - 3 deletions + 1 insertion + 1 substitution = WER of 5

Issues

- Out of vocabulary words (OOV)
 - Split the training set into two parts
 - Label all words in part 2 that were not in part 1 as <UNK>
- Clustering
 - e.g., dates, monetary amounts, organizations, years

Long Distance Dependencies

- This is where n-gram language models fail by definition
- Missing syntactic information
 - The students who participated in the game are tired
 - The student who participated in the game is tired
- Missing semantic information
 - The pizza that I had last night was tasty
 - The class that I had last night was interesting

Other Ideas in LM

- Syntactic models
 - Condition words on other words that appear in a specific syntactic relation with them
- Caching models
 - Take advantage of the fact that words appear in bursts

External Resources

- SRI-LM
 - <http://www.speech.sri.com/projects/srilm/>
- CMU-LM
 - <http://www.speech.cs.cmu.edu/SLM/toolkit.html>
- Google n-gram corpus
 - <http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html>
- Google book n-grams
 - <http://ngrams.googlelabs.com/>

Example Google n-grams

house a	302435	house hotel	139282
house after	118894	house in	3553052
house all	105970	house is	1962473
house and	3880495	house music	199346
house are	136475	house near	131889
house arrest	254629	house now	127043
house as	339590	house of	3164591
house at	694739	house on	1077835
house before	102663	house or	1172783
house built	189451	house party	162668
house but	137151	house plan	172765
house by	249118	house plans	434398
house can	133187	house price	158422
house cleaning	125206	house prices	643669
house design	120500	house rental	209614
house down	109663	house rules	108025
house fire	112325	house share	101238
house for	1635280	house so	133405
house former	112559	house that	687925
house from	249091	house the	478204
house had	154848	house to	1452996
house has	440396	house training	163056
house he	115434	house value	135820

N-gram External Links

- <http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html>
- <http://norvig.com/mayzner.html>
- <http://storage.googleapis.com/books/ngrams/books/datasetsv2.html>
- <https://books.google.com/ngrams/>
- <http://www.visi.com/~sgrantz/plot.html>
- <http://www.elsewhere.org/pomo/>
- <http://pdos.csail.mit.edu/scigen/>
- <http://www.magliery.com/Band/>
- <http://www.magliery.com/Country/>
- <http://johno.jsmf.net/knowhow/ngrams/index.php>
- <http://coastalweb.ca/building-sites/content-generation-with-n-grams.html>
- <http://www.decontextualize.com/teaching/rwet/n-grams-and-markov-chains/>
- <http://gregstevens.name/2012/08/16/simulating-h-p-lovecraft>
- <http://kingjamesprogramming.tumblr.com/>

NLP

Introduction to NLP

Word Sense Disambiguation

Introduction

- Polysemy
 - Words have multiple senses
- Example
 - Let's have a drink in the bar
 - I have to study for the bar
 - Bring me a chocolate bar
- Homonymy
 - May I come in?
 - Let's meet again in May
- Part of speech ambiguity
 - Joe won the first round
 - Joe has a round toy

Senses Of The Word “Bar”

- S: (n) barroom, **bar**, saloon, ginmill, taproom (a room or establishment where alcoholic drinks are served over a counter) "*he drowned his sorrows in whiskey at the bar*"
- S: (n) **bar** (a counter where you can obtain food or drink) "*he bought a hot dog and a coke at the bar*"
- S: (n) **bar** (a rigid piece of metal or wood; usually used as a fastening or obstruction or weapon) "*there were bars in the windows to prevent escape*"
- S: (n) measure, **bar** (musical notation for a repeating pattern of musical beats) "*the orchestra omitted the last twelve bars of the song*"
- S: (n) **bar** (an obstruction (usually metal) placed at the top of a goal) "*it was an excellent kick but the ball hit the bar*"
- S: (n) prevention, **bar** (the act of preventing) "*there was no bar against leaving*"; "*money was allocated to study the cause and prevention of influenza*"
- S: (n) **bar** ((meteorology) a unit of pressure equal to a million dynes per square centimeter) "*unfortunately some writers have used bar for one dyne per square centimeter*"
- S: (n) **bar** (a submerged (or partly submerged) ridge in a river or along a shore) "*the boat ran aground on a submerged bar in the river*"
- S: (n) legal profession, **bar**, legal community (the body of individuals qualified to practice law in a particular jurisdiction) "*he was admitted to the bar in New Jersey*"
- S: (n) stripe, streak, **bar** (a narrow marking of a different color or texture from the background) "*a green toad with small black stripes or bars*"; "*may the Stars and Stripes forever wave*"
- S: (n) cake, **bar** (a block of solid substance (such as soap or wax)) "*a bar of chocolate*"
- S: (n) Browning automatic rifle, **BAR** (a portable .30 caliber automatic rifle operated by gas pressure and fed by cartridges from a magazine; used by United States troops in World War I and in World War II and in the Korean War)
- S: (n) **bar** (a horizontal rod that serves as a support for gymnasts as they perform exercises)
- S: (n) **bar** (a heating element in an electric fire) "*an electric fire with three bars*"
- S: (n) **bar** ((law) a railing that encloses the part of the courtroom where the judges and lawyers sit and the case is tried) "*spectators were not allowed past the bar*"

Word Sense Disambiguation

- Task
 - given a word
 - and its context
 - determine which sense it is
- Use for Machine Translation
 - e.g., translate “play” into Spanish
 - play the violin = tocar el violín
 - play tennis = jugar al tenis
- Other uses
 - Accent restoration (cote)
 - Text to speech generation (lead)
 - Spelling correction (aid/aide)
 - Capitalization restoration (Turkey)

Dictionary Method (Lesk)

- Match sentences to dictionary definitions
- Examples of plant (m-w.com):
 - plant₁ = a living thing that grows in the ground, usually has leaves or flowers, and needs sun and water to survive
 - plant₂ = a building or factory where something is made
- Examples of leaf
 - leaf₁ = a lateral outgrowth from a plant stem that is typically a flattened expanded variably shaped greenish organ, constitutes a unit of the foliage, and functions primarily in food manufacture by photosynthesis
 - leaf₂ = a part of a book or folded sheet containing a page on each side
- Find the pair of meanings that have the most overlapping definitions
 - “The *leaf* is the food making factory of green *plants*.”

Decision Lists (Yarowsky)

- Method introduced by Yarowsky (1994)
- Two senses per word
- Ordered rules: collocation -> sense
- Formula

Decision Lists (Yarowsky)

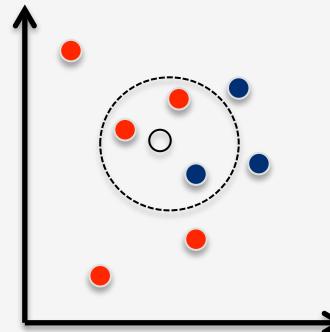
- *fish* within window -> bass1
- *striped bass* -> bass1
- *guitar* within window -> bass2
- *bass player* -> bass2
- Play/V *bass* -> bass2

Classification Features

- Adjacent words (collocations)
 - e.g., chocolate bar, bar exam, bar stool, bar fight, foreign aid, presidential aide
- Position
 - e.g., plant pesticide vs. pesticide plant
- Adjacent parts of speech
- Nearby words
 - e.g., within 10 words
- Syntactic information
 - e.g., object of the verb “play”
- Topic of the text

Classification Methods

- K-nearest neighbor (memory-based)
- Using Euclidean distance
- Find the k most similar examples and return the majority class for them



Bootstrapping

- Start with two senses and seeds for each sense
 - e.g., plant1:leaf, plant2:factory
- Use these seeds to label the data using a supervised classifier (decision list)
- Add some of the newly labeled examples to the training data
- Repeat until no more examples can be labeled

Bootstrapping

- Two principles:
 - one sense per collocation
 - one sense per discourse (e.g., document)

Training Data for WSD

- Senseval/Semcor
 - <http://www.senseval.org/senseval3>
 - Lexical Sample
 - All words
 - Available for many languages
- Pseudo-words
 - E.g., banana/door
- Multilingual corpora
 - Aligned at the sentence level
 - Use the translations as an indication of sense

Senseval-1 Evaluation

- Metric
 - A = number of assigned senses
 - C = number of words assigned correct senses
 - T = total number of test words
 - Precision = C/A; Recall = C/T
- Results
 - best recall around 77P/77R
 - human lexicographer 97P/96R
 - most common sense 57P/50R (decent but depends on domain)

NLP

NLP

Introduction to NLP

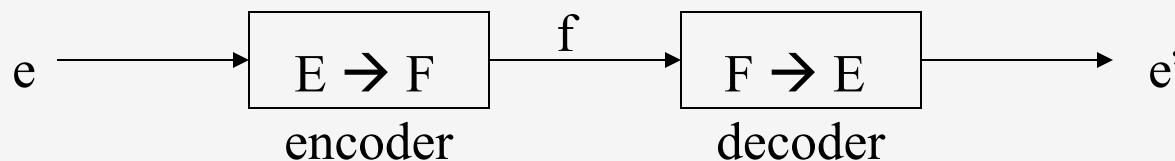
Noisy channel models

The Noisy Channel Model

- Example:
 - Input: Written English (X)
 - Encoder: garbles the input (X->Y)
 - Output: Spoken English (Y)
- More examples:
 - Grammatical English to English with mistakes
 - English to bitmaps (characters)
- $P(X,Y) = P(X)P(Y|X)$

Encoding and Decoding

- Given f , guess e



$$e' = \operatorname{argmax}_e P(e|f) = \operatorname{argmax}_e P(f|e) P(e)$$

translation model language model

Example

- Translate “la maison blanche”

	P(f e)	P(e)
cat plays piano	-	-
house white the	+	-
the house white	+	-
the red house	-	+
the small cat	-	+
the white house	+	+

Uses of the Noisy Channel Model

- Handwriting recognition
- Text generation
- Text summarization
- Machine translation
- Spelling correction
 - See separate lecture on text similarity and edit distance

Spelling Correction

w	c	w c	P(w c)	P(c)	$10^9 P(w c) P(c)$
thew	the	ew e	.000007	.02	144.
thew	thew		.95	.00000009	90.
thew	thaw	e a	.001	.0000007	0.7
thew	threw	h hr	.000008	.000004	0.03
thew	thwe	ew we	.000003	.00000004	0.0001

From Peter Norvig: <http://norvig.com/ngrams/ch14.pdf>

NLP

NLP

Introduction to NLP

Part of speech tagging

The POS Task

- Example
 - Bahrainis vote in second round of parliamentary election
- Jabberwocky (by Lewis Carroll, 1872)

‘Twas brillig, and the slithy toves
Did gyre and gimble in the wabe:
All mimsy were the borogoves,
And the mome raths outgrabe.

Parts of Speech

- Open class:
 - nouns, non-modal verbs, adjectives, adverbs
- Closed class:
 - prepositions, modal verbs, conjunctions, particles, determiners, pronouns

Penn Treebank tagset (1/2)

Tag	Description	Example
CC	coordinating conjunction	and
CD	cardinal number	1
DT	determiner	the
EX	existential there	<i>there</i> is
FW	foreign word	d'oeuvre
IN	preposition/subordinating conjunction	in, of, like
JJ	adjective	green
JJR	adjective, comparative	greener
JJS	adjective, superlative	greenest
LS	list marker	1)
MD	modal	could, will
NN	noun, singular or mass	table
NNS	noun plural	tables
NNP	proper noun, singular	John
NNPS	proper noun, plural	Vikings
PDT	predeterminer	<i>both</i> the boys
POS	possessive ending	friend's

Penn Treebank tagset (2/2)

Tag	Description	Example
PRP	personal pronoun	I, he, it
PRP\$	possessive pronoun	my, his
RB	adverb	however, usually, naturally, here, good
RBR	adverb, comparative	better
RBS	adverb, superlative	best
RP	particle	give <i>up</i>
TO	to	<i>to go, to him</i>
UH	interjection	uhhuhhuhh
VB	verb, base form	take
VBD	verb, past tense	took
VBG	verb, gerund/present participle	taking
VBN	verb, past participle	taken
VBP	verb, sing. present, non-3d	take
VBZ	verb, 3rd person sing. present	takes
WDT	wh-determiner	which
WP	wh-pronoun	who, what
WP\$	possessive wh-pronoun	whose
WRB	wh-abverb	where, when

Some Observations

- Ambiguity
 - count (noun) vs. count (verb)
 - 11% of all types but 40% of all tokens in the Brown corpus are ambiguous.
 - Examples
 - *like* can be tagged as ADP VERB ADJ ADV NOUN
 - *present* can be tagged as ADJ NOUN VERB ADV

Some Observations

- More examples:
 - transport, object, discount, address
 - content
- French pronunciation:
 - est, président, fils
- Three main techniques:
 - rule-based
 - machine learning (e.g., conditional random fields, maximum entropy Markov models)
 - transformation-based
- Useful for parsing, translation, text to speech, word sense disambiguation, etc

Example

- Bethlehem/NNP Steel/NNP Corp./NNP ,/, hammered/VBN by/IN higher/JJR **costs/NNS**
- Bethlehem/NNP Steel/NNP Corp./NNP ,/, hammered/VBN by/IN higher/JJR **costs/VBZ**

Sources of Information

- Bethlehem/NNP Steel/NNP Corp./NNP ,/, hammered/
VBN by/IN higher/JJR **costs/NNS**
- Bethlehem/NNP Steel/NNP Corp./NNP ,/, hammered/
VBN by/IN higher/JJR **costs/VBZ**
- Knowledge about individual words
 - lexical information
 - spelling (-or)
 - capitalization (IBM)
- Knowledge about neighboring words

Evaluation

- Baseline
 - tag each word with its most likely tag
 - tag each OOV word as a noun.
 - around 90%
- Current accuracy
 - around 97% for English
 - compared to 98% human performance

Rule-based POS tagging

- Use dictionary or finite-state transducers to find all possible parts of speech
- Use disambiguation rules
 - e.g., ART+V
- Hundreds of constraints can be designed manually

Example in French

<S>	^	beginning of sentence
La	rf b nms u	article
teneur	nfs nms	noun feminine singular
moyenne	jfs nfs v1s v2s v3s	adjective feminine singular
en	p a b	preposition
uranium	nms	noun masculine singular
des	p r	preposition
rivières	nfp	noun feminine plural
,	x	punctuation
bien_que	cs	subordinating conjunction
délicate	jfs	adjective feminine singular
à	p	preposition
calculer	v	verb

Sample Rules

- **BS3 BI1**
 - A BS3 (3rd person subject personal pronoun) cannot be followed by a BI1 (1st person indirect personal pronoun).
 - In the example: “il nous faut” (= “we need”) – “il” has the tag BS3MS and “nous” has the tags [BD1P BI1P BJ1P BR1P BS1P].
 - The negative constraint “BS3 BI1” rules out “BI1P”, and thus leaves only 4 alternatives for the word “nous”.
- **N K**
 - The tag N (noun) cannot be followed by a tag K (interrogative pronoun); an example in the test corpus would be: “... fleuve qui ...” (...river, that...).
 - Since “qui” can be tagged both as an ‘`E’’ (relative pronoun) and a “K” (interrogative pronoun), the “E” will be chosen by the tagger since an interrogative pronoun cannot follow a noun (“N”).
- **R V**
 - A word tagged with R (article) cannot be followed by a word tagged with V (verb): for example “l’ appelle” (calls him/her).
 - The word “appelle” can only be a verb, but “l’” can be either an article or a personal pronoun.
 - Thus, the rule will eliminate the article tag, giving preference to the pronoun.

NLP

NLP

Introduction to NLP

Hidden Markov Models

Markov Models

- Sequence of random variables that aren't independent
- Examples
 - weather reports
 - text

Properties

- Limited horizon:

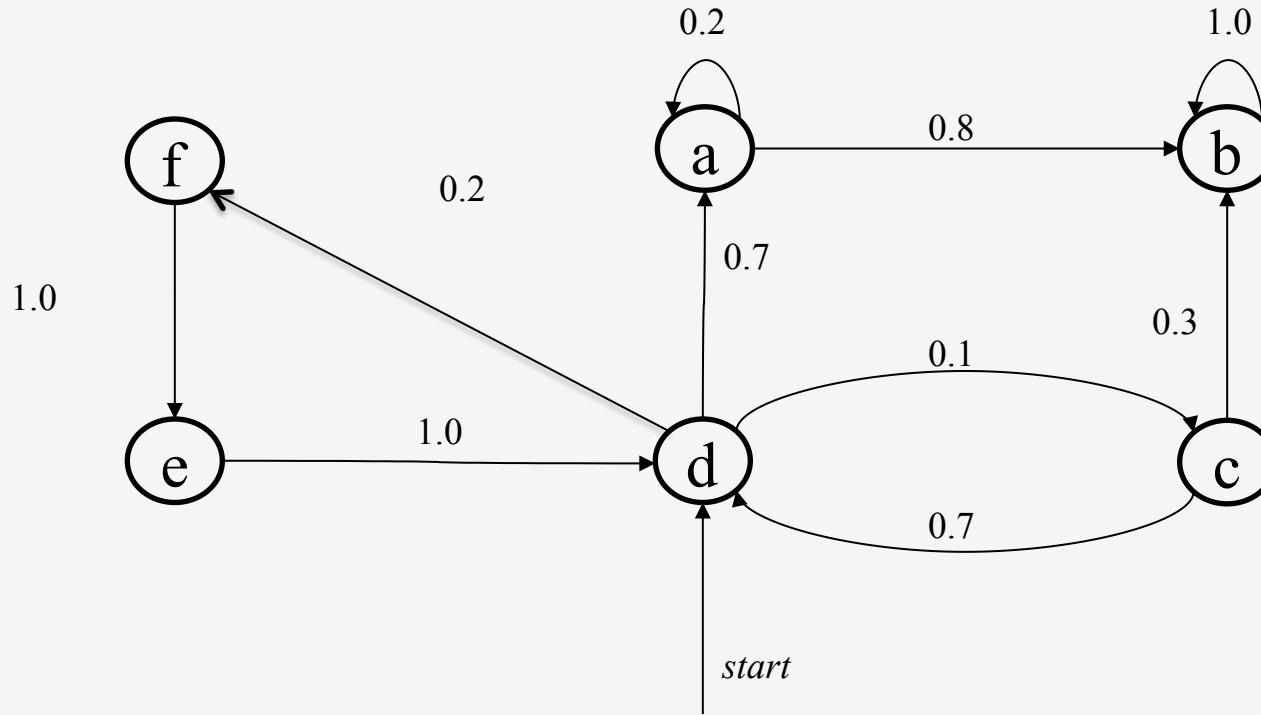
$$P(X_{t+1} = s_k | X_1, \dots, X_t) = P(X_{t+1} = s_k | X_t)$$

- Time invariant (stationary)

$$= P(X_2 = s_k | X_1)$$

- Definition: in terms of a transition matrix A and initial state probabilities Π .

Example



Visible MM

$$P(X_1, \dots, X_T) = P(X_1) P(X_2 | X_1) P(X_3 | X_1, X_2) \dots P(X_T | X_1, \dots, X_{T-1})$$

$$= P(X_1) P(X_2 | X_1) P(X_3 | X_2) \dots P(X_T | X_{T-1})$$

$$= \pi_{X_1} \prod_{t=1}^{T-1} a_{X_t, X_{t+1}}$$

$$P(d, a, b) = P(X_1=d) P(X_2=a|X_1=d) P(X_3=b|X_2=a)$$

$$= 1.0 \times 0.7 \times 0.8$$

$$= 0.56$$

Hidden MM

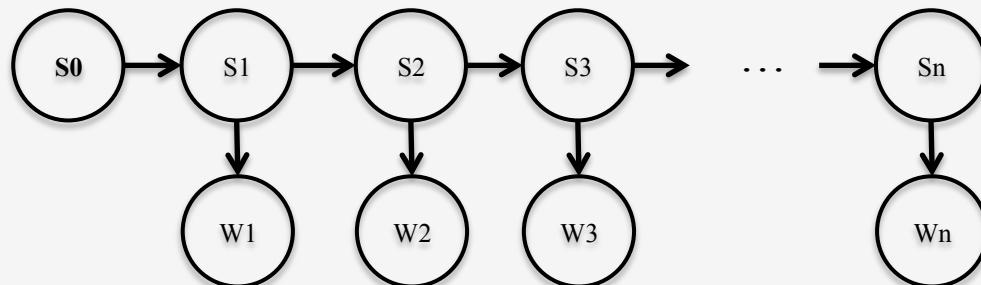
- Motivation
 - Observing a sequence of symbols
 - The sequence of states that led to the generation of the symbols is hidden
- Definition
 - Q = sequence of states
 - O = sequence of observations, drawn from a vocabulary
 - q_0, q_f = special (start, final) states
 - A = state transition probabilities
 - B = symbol emission probabilities
 - Π = initial state probabilities
 - $\mu = (A, B, \Pi)$ = complete probabilistic model

Hidden MM

- **Uses**
 - part of speech tagging
 - speech recognition
 - gene sequencing

Hidden Markov Model (HMM)

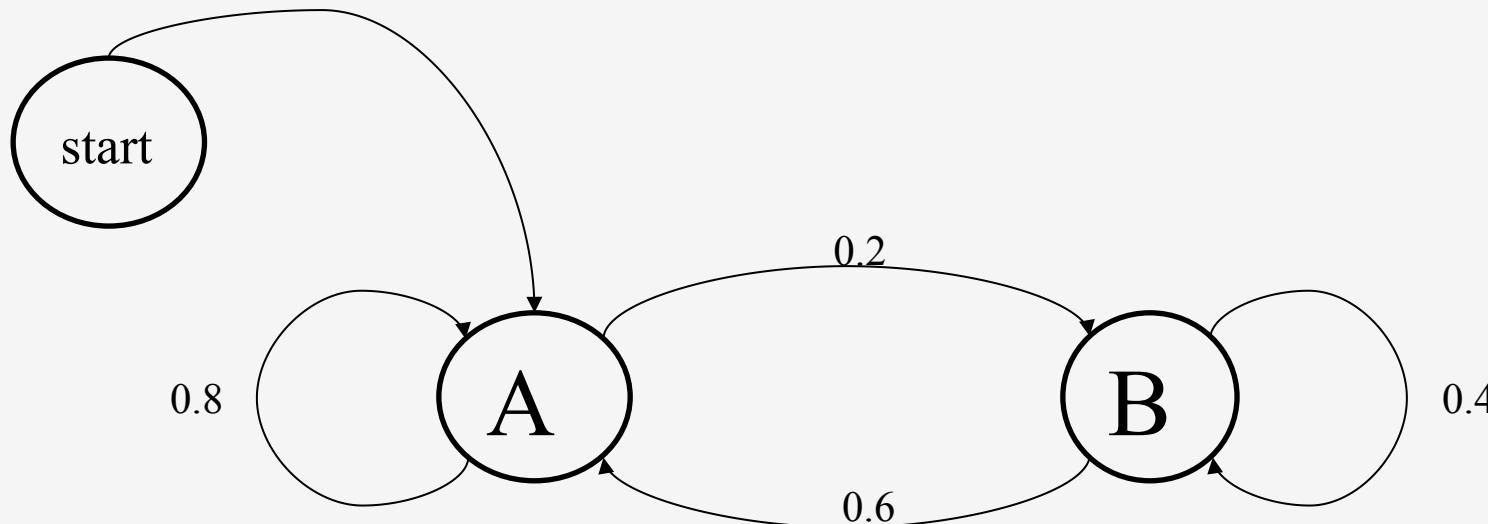
- Can be used to model state sequences and observation sequences
- Example:
 - $P(s, w) = \prod_i P(s_i | s_{i-1})P(w_i | s_i)$



Generative Algorithm

- Pick start state from Π
- For $t = 1..T$
 - Move to another state based on A
 - Emit an observation based on B

State Transition Probabilities



Emission Probabilities

- $P(O_t=k|X_t=s_i, X_{t+1}=s_j) = b_{ijk}$

	x	y	z
A	0.7	0.2	0.1
B	0.3	0.5	0.2

All Parameters of the Model

- Initial
 - $P(A|\text{start}) = 1.0, P(B|\text{start}) = 0.0$
- Transition
 - $P(A|A) = 0.8, P(A|B) = 0.6, P(B|A) = 0.2, P(B|B) = 0.4$
- Emission
 - $P(x|A) = 0.7, P(y|A) = 0.2, P(z|A) = 0.1$
 - $P(x|B) = 0.3, P(y|B) = 0.5, P(z|B) = 0.2$

Observation Sequence “yz”

- Starting in state A, $P(yz) = ?$
- Possible sequences of states:
 - AA
 - AB
 - BA
 - BB
- $P(yz) = P(yz|AA) + P(yz|AB) + P(yz|BA) + P(yz|BB) =$
 $= .8 \times .2 \times .8 \times .1$
 $+ .8 \times .2 \times .2 \times .2$
 $+ .2 \times .5 \times .4 \times .2$
 $+ .2 \times .5 \times .6 \times .1$
 $= .0128 + .0064 + .0080 + .0060 = .0332$

States and Transitions

- The states encode the most recent history
- The transitions encode likely sequences of states
 - e.g., Adj–Noun or Noun–Verb
 - or perhaps Art–Adj–Noun
- Use MLE to estimate the transition probabilities

Emissions

- Estimating the emission probabilities
 - Harder than transition probabilities
 - There may be novel uses of Word/POS combinations
- Suggestions
 - It is possible to use standard smoothing
 - As well as heuristics (e.g., based on the spelling of the words)

Sequence of Observations

- The observer can only see the emitted symbols
- Observation likelihood
 - Given the observation sequence S and the model $\mu = (A, B, \Pi)$, what is the probability $P(S|\mu)$ that the sequence was generated by that model.
- Being able to compute the probability of the observations sequence turns the HMM into a language model

Tasks With HMM

- Tasks
 - Given $\mu = (A, B, \Pi)$, find $P(O|\mu)$
 - Given O , μ , what is (X_1, \dots, X_{T+1})
 - Given O and a space of all possible μ , find model that best describes the observations
- Decoding – most likely sequence
 - tag each token with a label
- Observation likelihood
 - classify sequences
- Learning
 - train models to fit empirical data

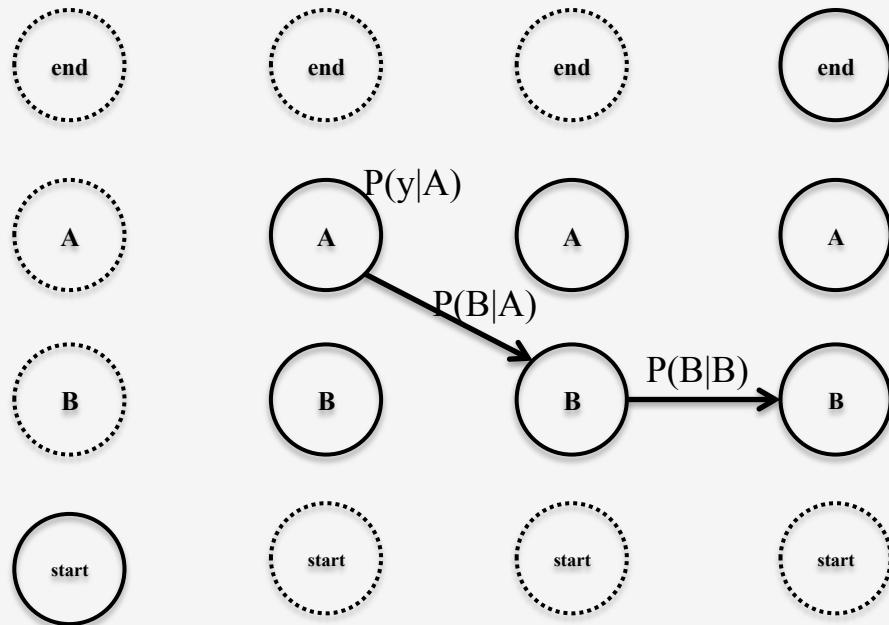
Inference

- Find the most likely sequence of tags, given the sequence of words
 - $t^* = \operatorname{argmax}_t P(t|w)$
- Given the model μ , it is possible to compute $P(t|w)$ for all values of t
- In practice, there are way too many combinations
- Possible solution:
 - Use beam search (partial hypotheses)
 - At each state, only keep the k best hypotheses so far
 - May not work

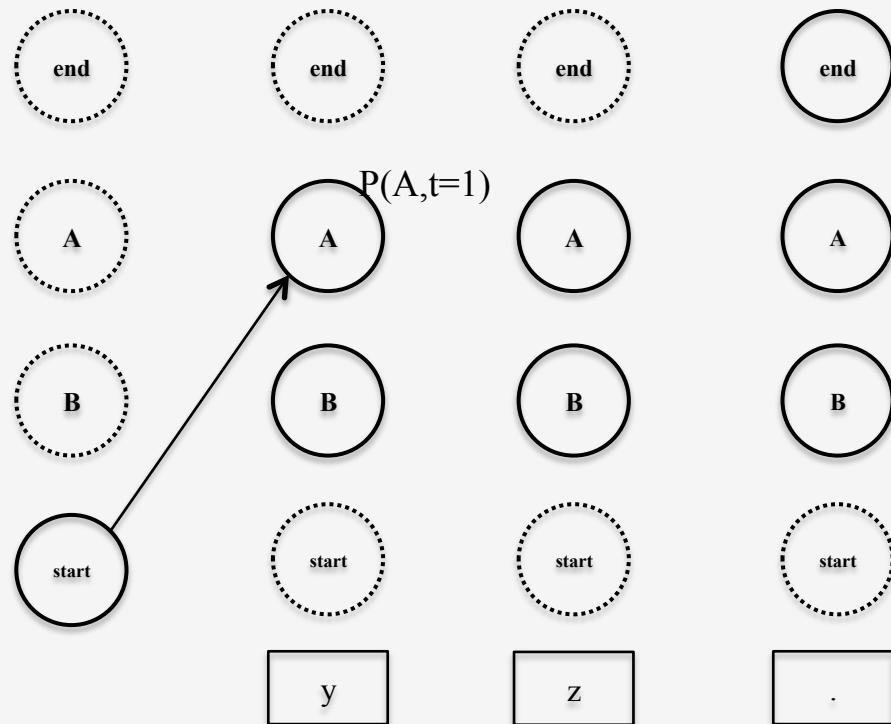
Viterbi Algorithm

- Find the best path up to observation i and state s
- Characteristics
 - Uses dynamic programming
 - Memoization
 - Backpointers

HMM Trellis

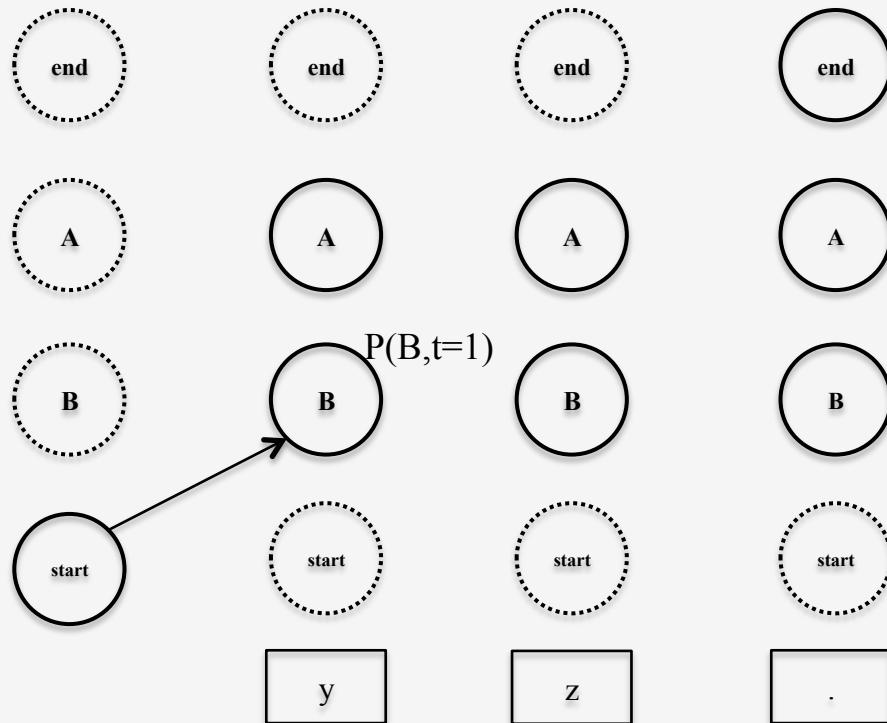


HMM Trellis



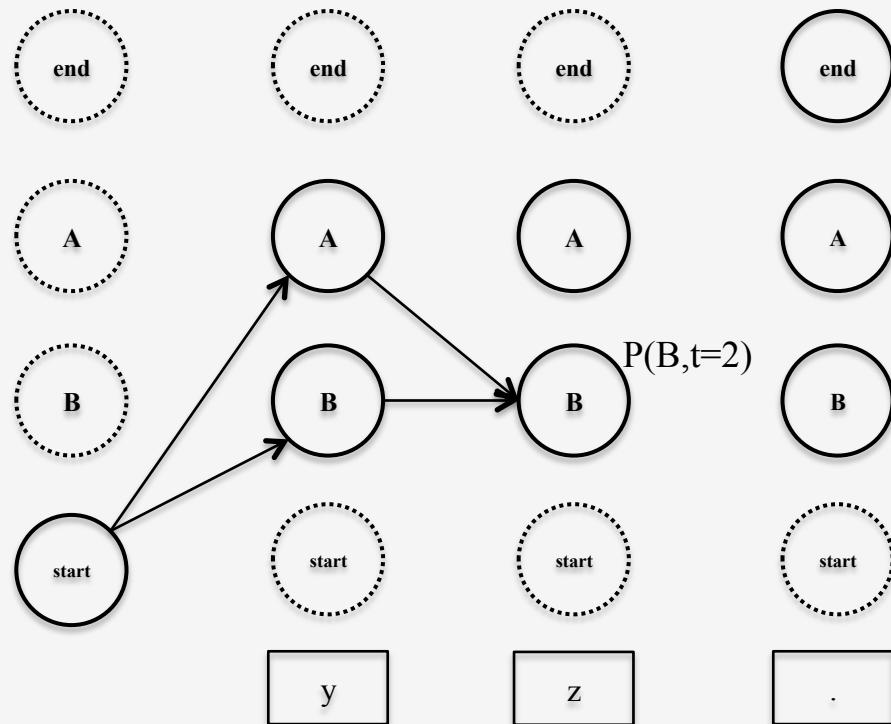
$$\begin{aligned} P(A, t=1) = \\ P(\text{start}) \times P(A|\text{start}) \times P(y|A) \end{aligned}$$

HMM Trellis



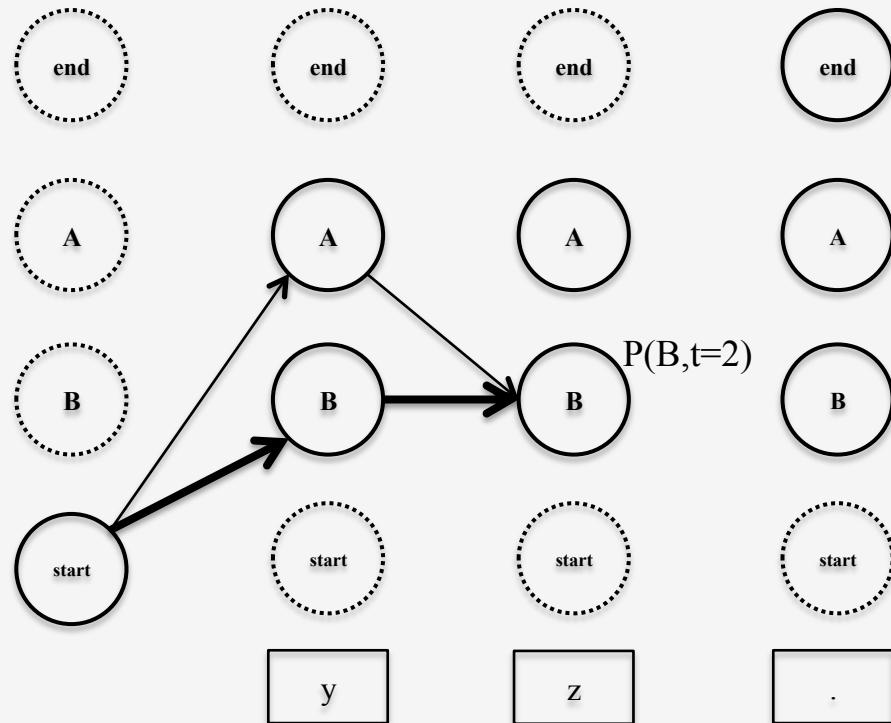
$$\begin{aligned} P(B, t=1) = \\ P(\text{start}) \times P(B|\text{start}) \times P(y|B) \end{aligned}$$

HMM Trellis

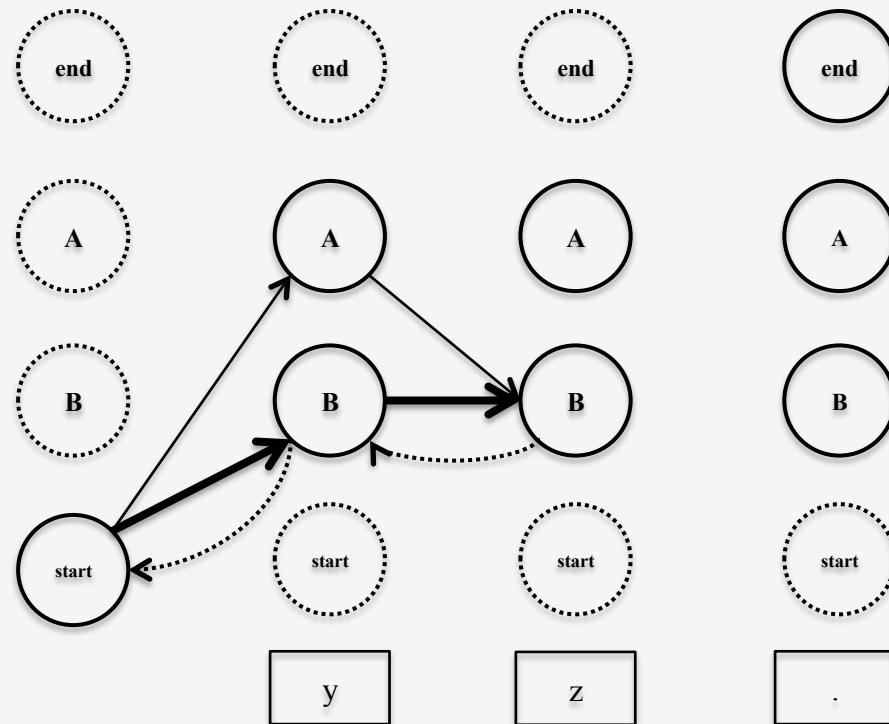


$$\begin{aligned} P(B, t=2) = \\ \max (P(A, t=1) \times P(B|A) \times P(z|b), \\ P(B, t=1) \times P(B|B) \times P(z|b)) \end{aligned}$$

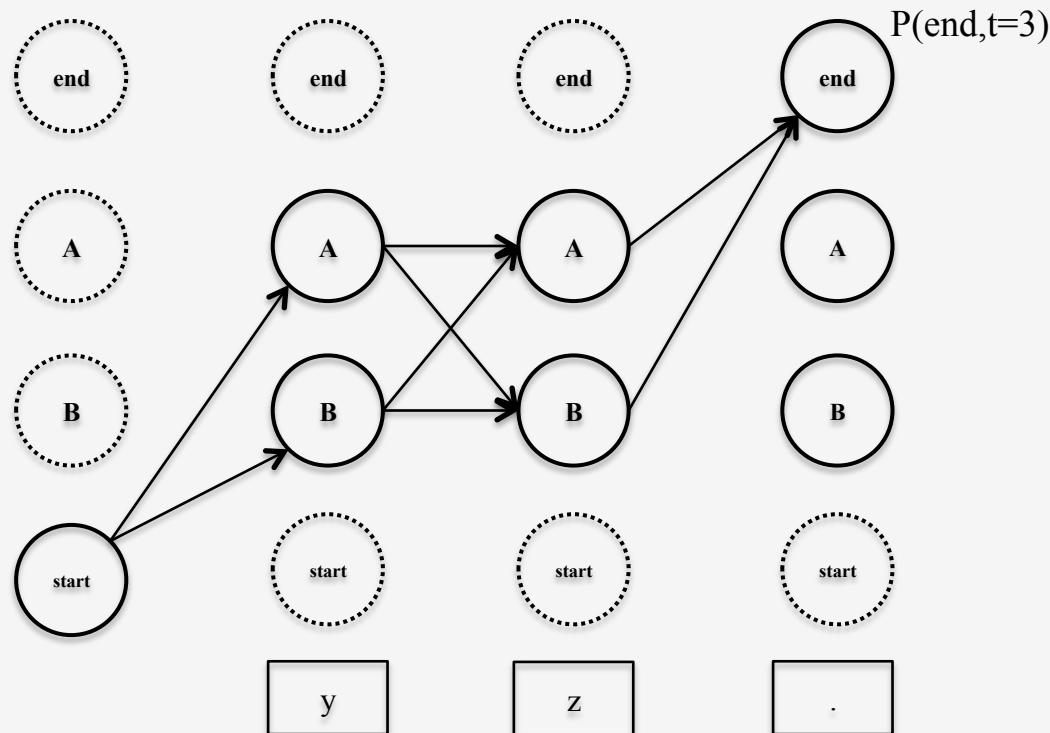
HMM Trellis



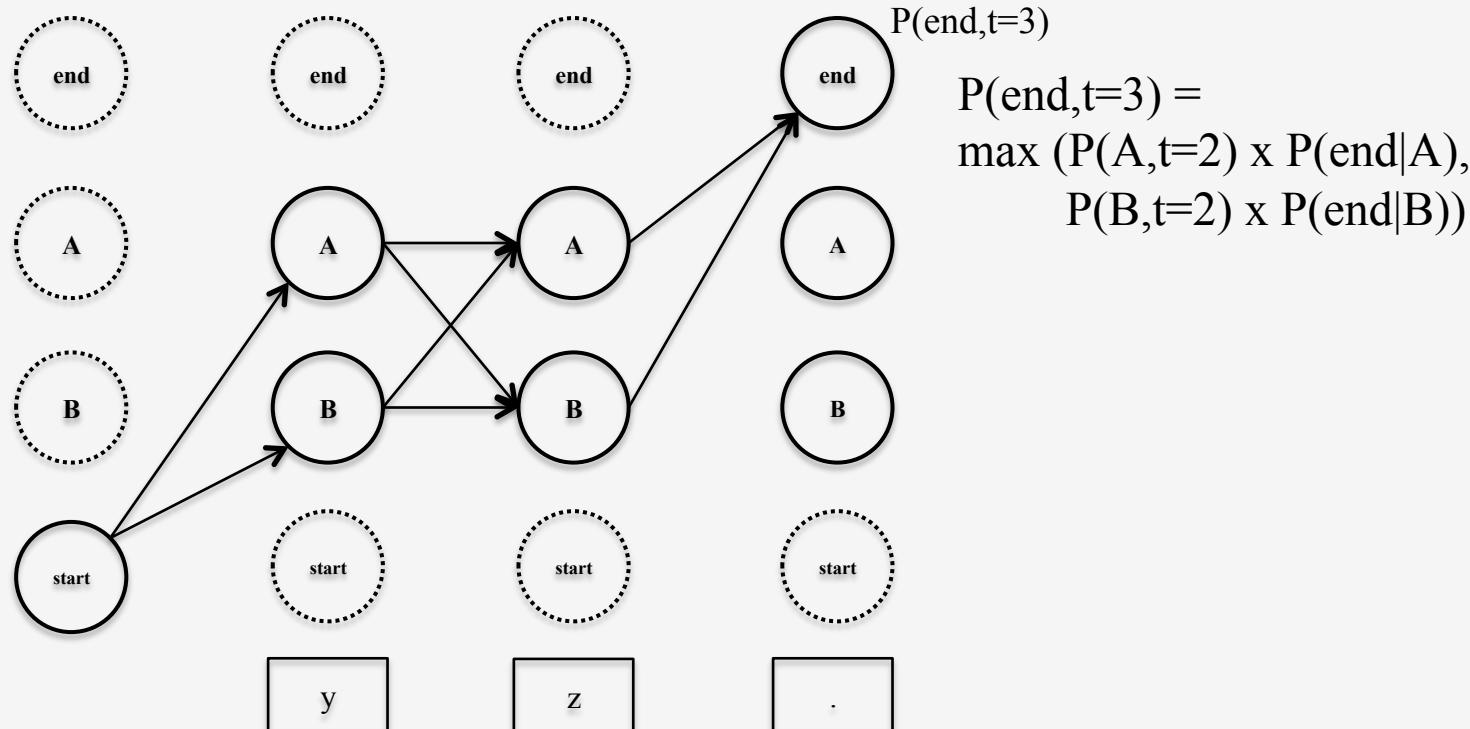
HMM Trellis



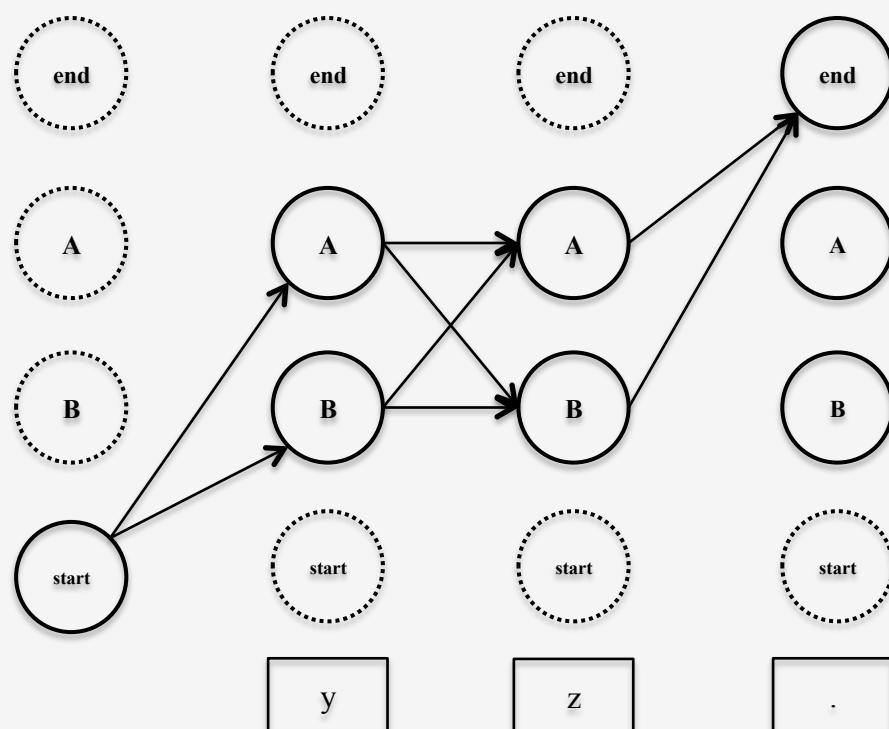
HMM Trellis



HMM Trellis



HMM Trellis



$P(\text{end}, t=3)$

$P(\text{end}, t=3) = \max (P(A, t=2) \times P(\text{end}|A), P(B, t=2) \times P(\text{end}|B))$

$P(\text{end}, t=3)$ = best score for the sequence

Use the backpointers to find the sequence of states.

NLP

NLP

Introduction to NLP

Hidden Markov Models (cont'd)

Observation Likelihood

- Given multiple HMMs
 - e.g., for different languages
 - Which one is the most likely to have generated the observation sequence
- Naïve solution
 - try all possible state sequences

Forward Algorithm

- Dynamic programming method
 - Computing a forward trellis that encodes all possible state paths.
 - Based on the Markov assumption that the probability of being in any state at a given time point only depends on the probabilities of being in all states at the previous time point

HMM Learning

- **Supervised**
 - Training sequences are labeled
- **Unsupervised**
 - Training sequences are unlabeled
 - Known number of states
- **Semi-supervised**
 - Some training sequences are labeled

Supervised HMM Learning

- Estimate the static transition probabilities using MLE

$$a_{ij} = \frac{\text{Count}(q_t = s_i, q_{t+1} = s_j)}{\text{Count}(q_t = s_i)}$$

- Estimate the observation probabilities using MLE

$$b_j(k) = \frac{\text{Count}(q_i = s_j, o_i = v_k)}{\text{Count}(q_i = s_j)}$$

- Use smoothing

Unsupervised HMM Training

- Given:
 - observation sequences
- Goal:
 - build the HMM
- Use EM (Expectation Maximization) methods
 - forward-backward (Baum-Welch) algorithm
 - Baum-Welch finds an approximate solution for $P(O|\mu)$

Outline of Baum-Welch

- **Algorithm**
 - Randomly set the parameters of the HMM
 - Until the parameters converge repeat:
 - E step – determine the probability of the various state sequences for generating the observations
 - M step – reestimate the parameters based on these probabilities
- **Notes**
 - the algorithm guarantees that at each iteration the likelihood of the data $P(O|\mu)$ increases
 - it can be stopped at any point and give a partial solution
 - it converges to a local maximum

NLP

NLP

Introduction to NLP

Statistical POS Tagging

Part of Speech Tagging Methods

- Rule-based
- Stochastic
 - HMM (generative)
 - Maximum Entropy MM (discriminative)
- Transformation-based

HMM Tagging

- $T = \operatorname{argmax} P(T|W)$
 - where $T=t_1, t_2, \dots, t_n$
- By Bayes' theorem
 - $P(T|W) = P(T)P(W|T)/P(W)$
- Thus we are attempting to choose the sequence of tags that maximizes the right hand side of the equation
 - $P(W)$ can be ignored
 - $P(T)$ is called the prior, $P(W|T)$ is called the likelihood.

HMM Tagging

- Complete formula
 - $P(T)P(W|T) = \prod P(w_i|w_1t_1\dots w_{i-1}t_{i-1}t_i)P(t_i|t_1\dots t_{i-2}t_{i-1})$
- Simplification 1:
 - $P(W|T) = \prod P(w_i|t_i)$
- Simplification 2:
 - $P(T) = \prod P(t_i|t_{i-1})$
- Bigram approximation
 - $T = \text{argmax } P(T|W) = \text{argmax } \prod P(w_i|t_i) P(t_i|t_{i-1})$

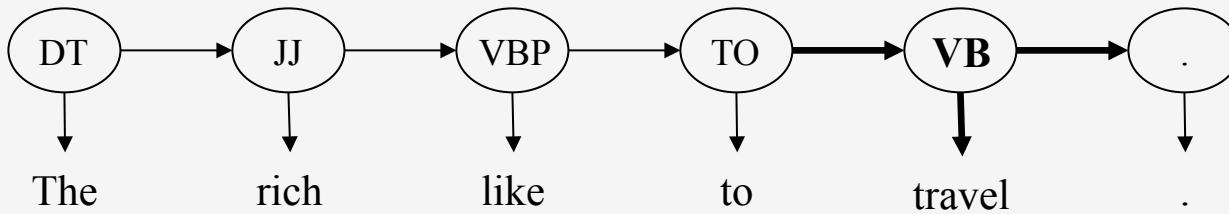
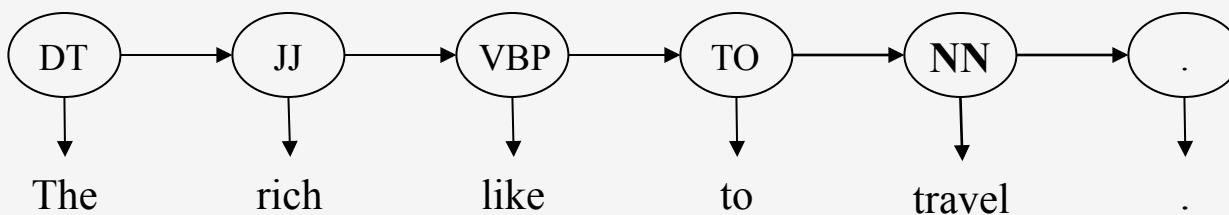
Maximum Likelihood Estimates

- $P(NN|JJ) = C(JJ, NN)/C(JJ) = 22301/89401 = .249$
- $P(this|DT) = C(DT, this)/C(DT) = 7037/103687 = .068$

Example

- The/DT rich/JJ like/VBP to/TO travel/VB ./.

Example



Evaluating Taggers

- Data set
 - Training set
 - Development set
 - Test set
- Tagging accuracy
 - how many tags right
- Results
 - Accuracy around 97% on PTB trained on 800,000 words
 - (50–85% on unknown words; 50% for trigrams)
 - Upper bound 98% – noise (e.g., errors and inconsistencies in the data, e.g., NN vs JJ)

Transformation-Based Learning

- [Brill 1995]
- Example
 - $P(NN|sleep) = .9$
 - $P(VB|sleep) = .1$
 - Change NN to VB when the previous tag is TO
- Types of rules:
 - The preceding (following) word is tagged z
 - The word two before (after) is tagged z
 - One of the two preceding (following) words is tagged z
 - One of the three preceding (following) words is tagged z
 - The preceding word is tagged z and the following word is tagged w

Transformation Based Tagger

#	From	To	Change Tag	Condition
1	NN	VB		Previous tag is <i>TO</i>
2	VBP	VB		One of the previous three tags is <i>MD</i>
3	NN	VB		One of the previous two tags is <i>MD</i>
4	VB	NN		One of the previous two tags is <i>DT</i>
5	VBD	VBN		One of the previous three tags is <i>VBZ</i>
6	VBN	VBD		Previous tag is <i>PRP</i>
7	VBN	VBD		Previous tag is <i>NNP</i>
8	VBD	VBN		Previous tag is <i>VBD</i>
9	VBP	VB		Previous tag is <i>TO</i>
10	POS	VBZ		Previous tag is <i>PRP</i>
11	VB	VBP		Previous tag is <i>NNS</i>
12	VBD	VBN		One of previous three tags is <i>VBP</i>
13	IN	WDT		One of next two tags is <i>VB</i>
14	VBD	VBN		One of previous two tags is <i>VB</i>
15	VB	VBP		Previous tag is <i>PRP</i>
16	IN	WDT		Next tag is <i>VBZ</i>
17	IN	DT		Next tag is <i>NN</i>
18	JJ	NNP		Next tag is <i>NNP</i>
19	IN	WDT		Next tag is <i>VBD</i>
20	JJR	RBR		Next tag is <i>JJ</i>

Figure 4

The first 20 nonlexicalized transformations.

Transformation Based Tagger

*Change tag **a** to tag **b** when:*

1. The preceding (following) word is w .
2. The word two before (after) is w .
3. One of the two preceding (following) words is w .
4. The current word is w and the preceding (following) word is x .
5. The current word is w and the preceding (following) word is tagged z .
6. The current word is w .
7. The preceding (following) word is w and the preceding (following) tag is t .
8. The current word is w , the preceding (following) word is w_2 and the preceding (following) tag is t .

where w and x are variables over all words in the training corpus, and z and t are variables over all parts of speech.

Transformation Based Tagger

#	From	To	Condition
1	NN	NNS	Has suffix -s
2	NN	CD	Has character .
3	NN	JJ	Has character -
4	NN	VBN	Has suffix -ed
5	NN	VBG	Has suffix -ing
6	??	RB	Has suffix -ly
7	??	JJ	Adding suffix -ly results in a word.
8	NN	CD	The word \$ can appear to the left.
9	NN	JJ	Has suffix -al
10	NN	VB	The word would can appear to the left.
11	NN	CD	Has character 0
12	NN	JJ	The word be can appear to the left.
13	NNS	JJ	Has suffix -us
14	NNS	VBZ	The word it can appear to the left.
15	NN	JJ	Has suffix -ble
16	NN	JJ	Has suffix -ic
17	NN	CD	Has character 1
18	NNS	NN	Has suffix -ss
19	??	JJ	Deleting the prefix un- results in a word
20	NN	JJ	Has suffix -ive

Figure 6

The first 20 transformations for unknown words.

Thoughts About POS Taggers

- New domains
 - Lower performance
- Distributional clustering
 - Combine statistics about semantically related words
 - Example: names of companies
 - Example: days of the week
 - Example: animals

External Links

- Jason Eisner's awesome interactive spreadsheet about learning HMMs
 - <http://cs.jhu.edu/~jason/papers/#eisner-2002-tnlp>
 - <http://cs.jhu.edu/~jason/papers/eisner.hmm.xls>

NLP

NLP

Introduction to NLP

Information extraction

Information Extraction

- Usually from unstructured or semi-structured data
- Examples
 - News stories
 - Scientific papers
 - Resumes
- Entities
 - Who did what, when, where, why
- Build knowledge base

Named Entities

- **Types:**
 - People
 - Locations
 - Organizations
 - Teams
 - Newspapers
 - Companies
 - Geo-political entities
- **Ambiguity:**
 - London can be a person, city, country (by metonymy) etc.
- **Useful for interfaces to databases, question answering, etc.**

Times and Events

- Times
 - Absolute expressions
 - Relative expressions (e.g., “last night”)
- Events

Sequence Labeling

- Many NLP problems can be cast as sequence labeling problems
 - POS – part of speech tagging
 - NER – named entity recognition
 - SRL – semantic role labeling
- Input
 - Sequence $w_1 w_2 w_3$
- Output
 - Labeled words
- Classification methods
 - Can use the categories of the previous tokens as features in classifying the next one
 - Direction matters

Named Entity Recognition (NER)

- Segmentation
 - Which words belong to a named entity?
 - Brazilian football legend Pele's condition has improved, according to a Thursday evening statement from a Sao Paulo hospital.
- Classification
 - What type of named entity is it?
 - Use gazetteers, spelling, adjacent words, etc.
 - Brazilian football legend [_{PERSON} Pele]'s condition has improved, according to a [_{TIME} Thursday evening] statement from a [_{LOCATION} Sao Paulo] hospital.

NER, Time, and Event Extraction

- Brazilian football legend [PERSON Pele]'s condition has improved, according to a [TIME Thursday evening] statement from a [LOCATION Sao Paulo] hospital.
- There had been earlier concerns about Pele's health after [ORG Albert Einstein Hospital] issued a release that said his condition was "unstable."
- [TIME Thursday night]'s release said [EVENT Pele was relocated] to the intensive care unit because a kidney dialysis machine he needed was in ICU.

Biomedical Example

- Gene labeling
- Sentence:
 - [_{GENE} BRCA1] and [_{GENE} BRCA2] are human genes that produce tumor suppressor proteins

NLP

NLP

Introduction to NLP

Relation extraction

Relation Extraction

- Links between entities
 - Works-for
 - Manufactures
 - Located-at

MUC

- Annual competition
 - DARPA, 1990s
- Events in news stories
 - Terrorist events
 - Joint ventures
 - Management changes
- Evaluation metrics
 - Precision
 - Recall
 - F-measure

MUC Example

```
<DOCNO> 0592 </DOCNO>
<DD> NOVEMBER 24, 1989, FRIDAY </DD>
<SO> Copyright (c) 1989 Jiji Press Ltd.; </SO>
<TXT>
BRIDGESTONE SPORTS CO. SAID FRIDAY IT HAS SET UP A JOINT VENTURE IN TAIWAN WITH
A LOCAL CONCERN AND A JAPANESE TRADING HOUSE TO PRODUCE GOLF CLUBS TO BE
SHIPPED 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. THE MONTHLY OUTPUT WILL BE
LATER RAISED TO 50,000 UNITS, BRIDGESTON SPORTS OFFICIALS SAID.
THE NEW COMPANY, BASED IN KAOHSIUNG, SOUTHERN TAIWAN, IS OWNED 75 PCT BY
BRIDGESTONE SPORTS, 15 PCT BY UNION PRECISION CASTING CO. OF TAIWAN AND THE
REMAINDER BY TAGA CO., A COMPANY ACTIVE IN TRADING WITH TAIWAN, THE OFFICIALS
SAID.
BRIDGESTONE SPORTS HAS SO FAR BEEN ENTRUSTING PRODUCTION OF GOLF CLUB PARTS
WITH UNION PRECISION CASTING AND OTHER TAIWAN COMPANIES.
WITH THE ESTABLISHMENT OF THE TAIWAN UNIT, THE JAPANESE SPORTS GOODS MAKER
PLANS TO INCREASE PRODUCTION OF LUXURY CLUBS IN JAPAN.
</TXT>
</DOC>
```

Figure 2: A sample article from the MUC-5 English joint ventures task.

```
<TEMPLATE-0592-1> :=
  DOC NR: 0592
  DOC DATE: 241189
  DOCUMENT SOURCE: "Jiji Press Ltd."
  CONTENT: <TIE_UP_RELATIONSHIP-0592-1>
  <TIE_UP_RELATIONSHIP-0592-1> :=
    TIE-UP STATUS: EXISTING
    ENTITY: <ENTITY-0592-1>
      <ENTITY-0592-2>
        <ENTITY-0592-3>
        JOINT VENTURE CO: <ENTITY-0592-4>
        OWNERSHIP: <OWNERSHIP-0592-1>
        ACTIVITY: <ACTIVITY-0592-1>
    ENTITY-0592-1 :=
      NAME: BRIDGESTONE SPORTS CO
      ALIASES: "BRIDGESTONE SPORTS"
        "BRIDGESTON SPORTS"
      NATIONALITY: Japan (COUNTRY)
      TYPE: COMPANY
      ENTITY RELATIONSHIP: <ENTITY_RELATIONSHIP-0592-1>
    ENTITY-0592-2 :=
      NAME: UNION PRECISION CASTING CO
      ALIASES: "UNION PRECISION CASTING"
      LOCATION: Taiwan (COUNTRY)
      NATIONALITY: Taiwan (COUNTRY)
      TYPE: COMPANY
      ENTITY RELATIONSHIP: <ENTITY_RELATIONSHIP-0592-1>
    ENTITY-0592-3 :=
      NAME: TAGA CO
      NATIONALITY: Japan (COUNTRY)
      TYPE: COMPANY
      ENTITY RELATIONSHIP: <ENTITY_RELATIONSHIP-0592-1>
    ENTITY-0592-4 :=
      NAME: BRIDGESTONE SPORTS TAIWAN CO
      LOCATION: "KAOHSIUNG" (UNKNOWN) Taiwan (COUNTRY)
      TYPE: COMPANY
      ENTITY RELATIONSHIP: <ENTITY_RELATIONSHIP-0592-1>
  <INDUSTRY-0592-1> :=
    INDUSTRY-TYPE: PRODUCTION
    PRODUCT/SERVICE: (38 "20,000 IRON AND 'METAL WOOD' [CLUBS]")
  <ENTITY_RELATIONSHIP-0592-1> :=
    ENTITY1: <ENTITY-0592-1>
      <ENTITY-0592-2>
        <ENTITY-0592-3>
    ENTITY2: <ENTITY-0592-4>
    REL OF ENTITY2 TO ENTITY1: CHILD
    STATUS: CURRENT
  <ACTIVITY-0592-1> :=
    INDUSTRY: <INDUSTRY-0592-1>
    ACTIVITY-SITE: (Taiwan (COUNTRY) <ENTITY-0592-4>
    START TIME: <TIME-0592-1>
  <TIME-0592-1> :=
    DURING: 0190
  <OWNERSHIP-0592-1> :=
    OWNED: <ENTITY-0592-4>
    TOTAL-CAPITALIZATION: 20000000 TWD
    OWNERSHIP-%: (<ENTITY-0592-3> 10)
      (<ENTITY-0592-2> 15)
      (<ENTITY-0592-1> 75)
```

Figure 3: A sample filled template from the MUC-5 English joint ventures task.

Example from
Grishman and Sundheim 1996

Other Examples

- Job announcements
 - Location, title, starting date, qualifications, salary
- Seminar announcements
 - Time, title, location, speaker
- Medical papers
 - Drug, disease, gene/protein, cell line, species, substance

Filling the Templates

- Some fields get filled by text from the document
 - E.g., the names of people
- Others can be pre-defined values
 - E.g., successful/unsuccessful merger
- Some fields allow for multiple values

Approaches

- View IE as a sequence labeling problem
 - Use HMM
- Use patterns
 - E.g., regular expressions
- Features
 - Capitalization (initial, allcaps), contains digits, spelling (e.g., suffixes), punctuation

Perl Regular Expressions

^	beginning of string; complement inside []
\$	end of string
.	any character except newline
*	match 0 or more times
+	match 1 or more times
?	match 0 or 1 times
	alternatives
()	grouping and memory
[]	set of characters
{ }	repetition modifier
\	special symbol

Perl Regular Expressions

a^*	zero or more
a^+	one or more
$a^?$	zero or one
$a\{m\}$	exactly m
$a\{m,\}$	at least m
$a\{m,n\}$	at least m but at most n
<i>repetition?</i>	shortest match

Perl Regular Expressions

\t	tab
\n	newline
\r	carriage return (CR)
*	asterisk
\?	question mark
\.	period
\xhh	hexadecimal character
\w	Matches one alphanumeric (or ‘_’) character
\W	matches the complement of \w
\s	space, tab, newline
\S	complement of \s
\d	same as [0-9]
\D	complement of \d
\b	“word” boundary
\B	complement of \b
[x-y]	inclusive range from x to y

Sample Patterns

- Price (e.g., \$14,000.00)
 - `\$[0-9,]+(\.[0-9]{2})?`
- Date (e.g., 2015-02-01)
 - `^(19|20)\d\d[- /.](0[1-9]|1[012])[- /.](0[1-9]|1[2][0-9]|3[01])$`
- Email
 - `^[_a-z0-9-]+(\.[_a-z0-9-]+)*@[a-z0-9-]+(\.[a-z0-9-]+)*(\.[a-z]{2,4})$`
- Person
- May include HTML code
- May include POS information
- May include Wordnet information

Sample Input for NER

```
( (S
  (NP-SBJ-1
    (NP (NNP Rudolph) (NNP Agnew) )
    (, ,)
    (UCP
      (ADJP
        (NP (CD 55) (NNS years) )
        (JJ old) )
      (CC and)
      (NP
        (NP (JJ former) (NN chairman) )
        (PP (IN of)
          (NP (NNP Consolidated) (NNP Gold) (NNP Fields) (NNP PLC) ))))
      (, ,) )
    (VP (VBD was)
      (VP (VBN named)
        (S
          (NP-SBJ (-NONE- *-1) )
          (NP-PRD
            (NP (DT a) (JJ nonexecutive) (NN director) )
            (PP (IN of)
              (NP (DT this) (JJ British) (JJ industrial) (NN conglomerate) )))))
        (..) ))
```

Sample Output for NER (IOB format)

file_id	sent_id	word_id	ioe_inner	pos	word
0002	1	0	B-PER	NNP	Rudolph
0002	1	1	I-PER	NNP	Agnew
0002	1	2	O	COMMA	COMMA
0002	1	3	B-NP	CD	55
0002	1	4	I-NP	NNS	years
0002	1	5	B-ADJP	JJ	old
0002	1	6	O	CC	and
0002	1	7	B-NP	JJ	former
0002	1	8	I-NP	NN	chairman
0002	1	9	B-PP	IN	of
0002	1	10	B-ORG	NNP	Consolidated
0002	1	11	I-ORG	NNP	Gold
0002	1	12	I-ORG	NNP	Fields
0002	1	13	I-ORG	NNP	PLC
0002	1	14	O	COMMA	COMMA
0002	1	15	B-VP	VBD	was
0002	1	16	I-VP	VBN	named
0002	1	17	B-NP	DT	a
0002	1	18	I-NP	JJ	nonexecutive
0002	1	19	I-NP	NN	director
0002	1	20	B-PP	IN	of
0002	1	21	B-NP	DT	this
0002	1	22	I-NP	JJ	British
0002	1	23	I-NP	JJ	industrial
0002	1	24	I-NP	NN	conglomerate
0002	1	25	O	.	.

Evaluating Template-based IE

- For each test document
 - Number of correct template extractions
 - Number of slot/value pairs extracted
 - Number of extracted slot/value pairs that are correct

Relation Extraction

- Person–person
 - ParentOf, MarriedTo, Manages
- Person–organization
 - WorksFor
- Organization–organization
 - IsPartOf
- Organization–location
 - IsHeadquarteredAt

ACE Evaluation

- 2002 newspaper data
- Entities:
 - Person, Organization, Facility, Location, Geopolitical Entity
- Relations:
 - Role, Part, Located, Near, Social

Relation Extraction

- Core NLP task
 - Used for building knowledge bases, question answering
- Input
 - **Mazda North American Operations** *is headquartered in Irvine, Calif.*, and oversees the sales, marketing, parts and customer service support of Mazda vehicles in the United States and Mexico through nearly 700 dealers.
- Output
 - IsHeadquarteredIn (Mazda North American Operations, Irvine)

Relation Extraction

- Using patterns
 - Regular expressions
 - Gazetteers
- Supervised learning
- Semi-supervised learning
 - Using seeds

Extracting IS-A Relations

- Hearst's patterns
 - X and other Y
 - X or other Y
 - Y such as X
 - Y, including X
 - Y, especially X
- Example
 - Evolutionary relationships between the platypus and other mammals

Supervised Relation Extraction

- Look for sentences that have two entities that we know are part of the target relation
- Look at the other words in the sentence, especially the ones between the two entities
- Use a classifier to determine whether the relation exists

Example

- English
 - Beethoven *was born* in December 1770 in Bonn
 - *Born* in Bonn in 1770, Beethoven ...
 - After his *birth* on December 16, 1770,
Beethoven grew up in a musical family
 - Ludwig van Beethoven (1770–1827)
 - While this evidence supports the case for 16
December 1770 as Beethoven's *date of birth*

Example (non-English)

- German
 - **Ludwig van Beethoven** wurde am 17. Dezember 1770 in Bonn getauft
 - **Ludwig van Beethoven** wurde in Bonn, 15. Dezember 1770, eine Familie ursprünglich aus Brabant in Belgien geboren
 - Der *Geburtstag* von **Ludwig van Beethoven** wurde im Winter 1770 in Bonn nicht genau dokumentiert
- Spanish
 - **Ludwig van Beethoven** nació en Bonn el 17 de diciembre de 1770
 - *Nacido en Bonn 1770, Beethoven ...*
 - **Ludwig van Beethoven**, *nace* en diciembre de 1770

Semi-supervised Relation Extraction

- Start with some seeds, e.g.,
 - Beethoven *was born* in December 1770 in Bonn
- Look for other sentences with the same words
- Look for expressions that appear nearby
- Look for other sentences with the same expressions

Evaluating Relation Extraction

- Precision P
 - correctly extracted relations/all extracted relations
- Recall R
 - correctly extracted relations/all existing relations
- F1 measure
 - $F1 = 2PR/(P+R)$
- If there is no annotated data
 - only measure precision

Conclusion

- Probabilistic NLP
- Part of Speech Tagging
- Hidden Markov Models
- Information Extraction

NLP

NLP

Introduction to NLP

Question Answering

Siri



Ask Jeeves (ask.com)

Ask.com - What's Your Qu... + ↻ www.ask.com/web?ad=dirN&qo=contentPageSearchBox&qs=1&o=2545&oo=2545&l=dir&q=Who+played+Linus W Wikipedia (en) Advanced Search Sign In

Answers Q&A Community

Ask Who played Linus in Lost?

Find Answers

Popular Q&A

Q: Who plays Ben Linus on Lost?
A: The character of Ben Linus is played by Michael Emerson born on the 7th of September 1954, ...
[Read More »](#)
Source: www.chacha.com

Q: Who plays ben linus in lost?
A: Michael Emerson plays Ben Linus on the tv show [Read More »](#)
Source: www.chacha.com

Q: Who plays young Benjamin Linus on lost?
A: Sterling Beaumon has played young Benjamin Linus on 5 episodes of Lost. Want his filmography? [Read More »](#)
Source: www.kgbanswers.com

Q: Who played Linus on True Jackson VP?
A: Tyler Peterson. [Read More »](#)
Source: wiki.answers.com

Q: Who played Linus in Charlie Brown Christmas Special?
A: Chris Shea. [Read More »](#)
Source: wiki.answers.com

More Answers

Ben Linus

Benjamin Ben Linus is a fictional character portrayed by Michael Emerson on the ABC television series *Lost*. Ben was the leader of a group of island natives called the Others and was initially known as Henry Gale to the survivors of Oceanic Flight 815. He began as the antagonist during the second and third seasons, but in subsequent seasons, becomes something of an uneasy ally to the main characters; even so, however, throughout the series, he is persistently characterized by spontaneous coldbloodedness.

[Read More »](#)

Go to: Ask Encyclopedia · Images · Videos

Browse Article: Arc · Personality · Development · Reception · References ·

Source: Wikipedia

Related Questions:

- Spiritually speaking, who is better looking: Ben Linus, Henry Gale, or Michael Emerson?
- Order of importance of these Lost Characters? in my opinion it's 1. Jack Sheppard 2. Kate 3. Ben Linus 4. John Locke 5. James Sawyer Ford 6. Hurley 7. Desmond 8. Juliet 9. Sun 10. Claire

[View more Q&A »](#)

Michael Emerson - IMDb
www.imdb.com/name/nm0256237/
Michael Emerson, Actor: Lost, Michael ... He is an actor, known for Lost (2004), Person of Interest (2011) and Saw (2004). He has ... Narrator / Benjamin Linus.

Roger Linus - Lostpedia - The Lost Encyclopedia - Wikipedia
lostpedia.wikia.com/wiki/Roger_Linus
Roger Linus was the father of Benjamin Linus, whom he regularly abused, both verbally and.... Portrayed by Outside references to Lost/Video games.

WolframAlpha

The screenshot shows a web browser window displaying the WolframAlpha website. The URL in the address bar is <https://www.wolframalpha.com/input/?i=What+is+the+area+of+a+circle+with+radius+5+ft?>. The main content area shows the query "What is the area of a circle with radius 5 ft?" and the resulting computation: $25\pi \text{ ft}^2$ (square feet) $\approx 78.5398 \text{ ft}^2$ (square feet). Below this, there is a visual representation of a circle with a radius labeled "5 ft". To the right of the main content, there are two promotional boxes: "New to Wolfram|Alpha?" and "New! Wolfram Problem Generator". The "New to Wolfram|Alpha?" box includes a "Take the Tour >" button. The "New! Wolfram Problem Generator" box includes a "Need a hint?" button.

IBM's Watson



<http://www.geekwire.com/2013/ibm-takes-watson-cloud/>

Watson on Jeopardy

- Sample questions

- On December 8, 2008 this national newspaper raised its newsstand price by 25 cents to \$1
 - *USA Today*
- In 2010 this former first lady published the memoir "Spoken From the Heart"
 - *Laura Bush*
- This person is appointed by a testator to carry out the directions & requests in his will
 - *Executor*
- Familiarity is said to breed this, from the Latin for "Despise"
 - *Contempt*
- As of 2010, Croatia & Macedonia are candidates but this is the only former Yugoslav republic in the EU
 - *Slovenia*
- The ancient "Lion of Nimrud" went missing from this city's national museum in 2003 (along with a lot of other stuff)
 - *Baghdad*
- It's just a bloody nose! You don't have this hereditary disorder once endemic to European royalty
 - *Haemophilia*
- It's Michelangelo's fresco on the wall of the Sistine Chapel, depicting the saved and the damned
 - *The Last Judgement*
- She "Died in the church and was buried along with her name. Nobody came"
 - *Eleanor Rigby*
- A camel is a horse designed by this
 - *Committee*

Watson Videos

- IBM's Watson Supercomputer Destroys Humans in Jeopardy | Engadget
 - https://www.youtube.com/watch?v=WFR3lOm_xhE
- "What is IBM Watson?" 7 Videos from the Jeopardy! Era
 - <http://mentalfloss.com/article/51543/what-ibm-watson-7-videos-jeopardy-era>

People Ask Questions Online

- **Excite corpus**
 - 2,477,283 queries (one day's worth)
- **8.4% of them are questions**
 - 43.9% factual (what is the country code for Belgium)
 - 56.1% procedural (how do I get out of debt) or other

The Excite Corpus

- In what year did baseball become an offical sport?
- How do I get out of debt?
- Where can I found out how to pass a drug test?
- When is the Super Bowl?
- Who is California's District State Senator?
- Where can I buy extra nibs for a foutain pen?
- What time is it in west samoa?
- Where can I buy a little kitty cat?
- What are the symptoms of attention deficit disorder?
- Where can I get some information on Michael Jordan?
- How does the character Seyavash in Ferdowsi's Shahnameh exhibit characteristics of a hero?
- When did the Neanderthal man live?
- Which Frenchman declined the Nobel Prize for Literature for ideological reasons?

Murax Questions

- What U.S. city is at the junction of the Allegheny and Monongahela rivers?
- Who wrote “Across the River and into the Trees” ?
- Who married actress Nancy Davis?
- What’s the capital of the Netherlands?
- Who was the last of the Apache warrior chiefs?
- What chief justice headed the commission that declared: “Lee Harvey Oswald . . . acted alone.”?
- What famed falls are split in two by Goat Island?
- What is November’s birthstone?
- Who’s won the most Oscars for costume design?
- What is the state flower of Alaska?

AOL Corpus Questions

- what does cerebral cortical atrophy mean
- what fraction is closest to pi
- what is the highest calories consumed by a person in a 24 hor period
- what happens to limewater in the presence of carbon dioxide
- what language is Numa Numa in
- who discovered Saturn
- who discovered the gene for Huntington's disease
- who is the daughter of indian sitar maestro Ravi Shankar
- who invented marshmallow sandwich cookies
- who invented lithograph
- who was the first Michigan football player to put on the cover of a video game
- who invented smallpox vaccine
- when was the berlin wall torn down
- when was yahoo launched
- when was the wedding march by Mendelssohn composed
- when was the declaration of independence signed
- when did France become a republic
- when was Alexander the Great born
- which fish contains low mercury
- when is the youngest a minor can move out of their parents house in Texas
- when should take a child to see an psychologist
- what hormone imbalance could mimic pregnancy symptoms

Question Type Taxonomy

- Question Types
 - Yes/no vs. wh-
 - Factual vs. procedural
 - Single answer vs. multiple answers
 - Objective vs. subjective
 - Context-specific (personalized) vs. generic
 - Known answer in the collection, y/n

State of the Art

- Question types
 - Mostly factual, short-answer questions
- System architecture
 - IR component
 - Statistical approaches using lots of data
 - Relatively little knowledge

Older Systems (pre-TREC)

- BASEBALL (Green et al., 1961)
 - Baseball
- ELIZA (Weizenbaum 1966)
 - Dialogue, Therapist
- SHRDLU (Winograd 1972)
 - Blocks world
- LUNAR (Woods, 1973)
 - Lunar rocks
- Murax (Kupiec 1993)
 - Open-domain, encyclopedia-based
- START (Katz 1997)
 - Open-domain, web-based
- Deep Red (Hirschman & al. 1999)
 - Reading comprehension
- Jupiter (Zue et al., 2000)
 - Spoken questions

Eliza (Weizenbaum 1966)

- Eliza plays a therapist
 - and doesn't really answer questions ☺
- Simple pattern matching
- Sample dialogue

E: I am the psychotherapist. Please, describe your problems.

P: I am always tired.

E: Why do you say you are always tired?

P: Because I cannot sleep.

E: Is it because you cannot sleep that you came to me?

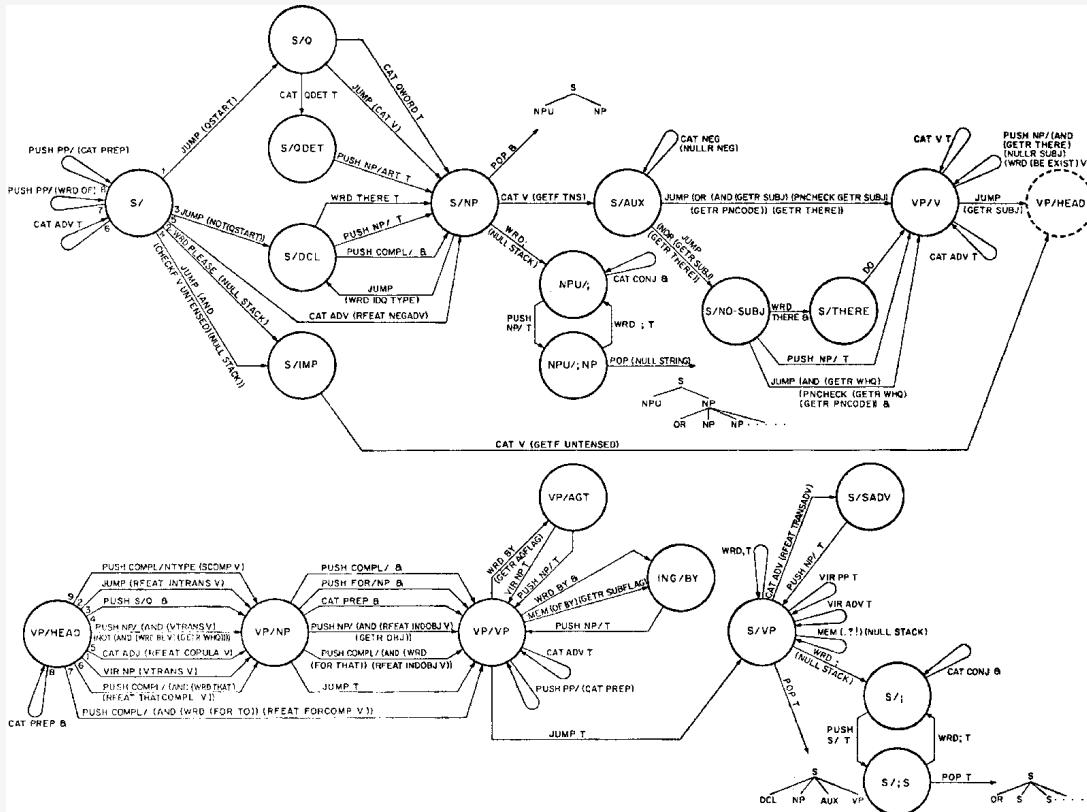
P: Yes, help me please.

E: Earlier you said you cannot sleep?

Lunar (Woods et al., 1972)

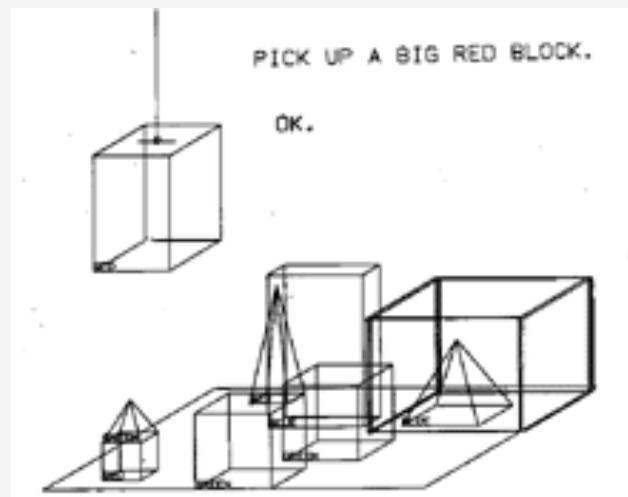
- Lunar Sciences Natural Language Information System
- Apollo 11 moon rocks
- Questions in English
- Components
 - a general-purpose ATN (augmented transition network) grammar for English
 - a procedural-semantics framework for semantic interpretation
 - mapping from the syntactic grammar to a logic representation
 - 3,500 word dictionary
 - 13,000 entries in the knowledge base
- Lunar answered 78% of sentences asked by geologists at 1971 Lunar Rocks conference
 - How many breccias contain olivine?
 - What are they?

Lunar Grammar Fragment



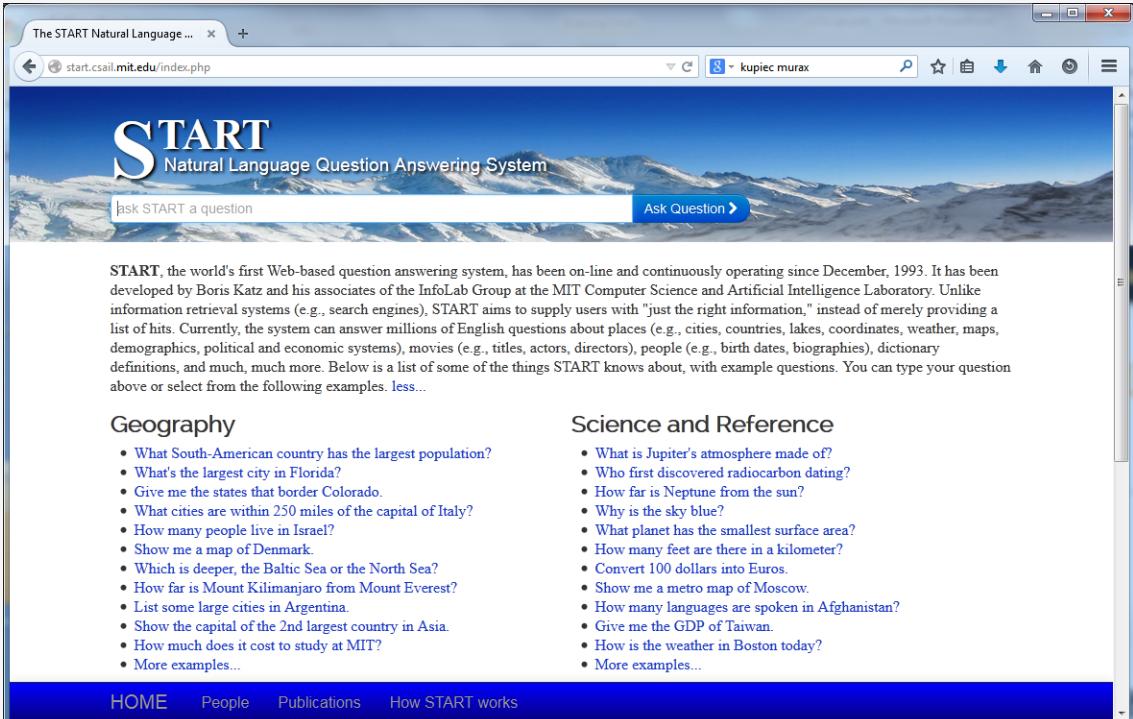
SHRDLU (Winograd 1972)

- Simple questions about the blocks world
 - Person: Pick up a big red block.
 - Computer: OK.
 - Person: Grasp the pyramid.
 - Computer: I don't understand which pyramid you mean.
 - Person (changing their mind): Find a block which is taller than the one you are holding and put it into the box.
 - Computer: By "it", I assume you mean the block which is taller than the one I am holding.
 - Computer: OK.
 - Person: What does the box contain?
 - Computer: The blue pyramid and the blue block.
 - Person: What is the pyramid supported by?
 - Computer: The box.
 - Person: How many blocks are not in the box?
 - Computer: Four of them.
 - Person: Is at least one of them narrower than the one which I told you to pick up?
 - Computer: Yes, the red cube.



Start (Katz 1997)

- Open-domain
- Uses the Web
- Online since 1993



The screenshot shows a web browser window for the START system at start.csail.mit.edu/index.php. The page features a large blue header with the word "START" in white. Below the header is a search bar with the placeholder "ask START a question" and a blue "Ask Question" button. The main content area contains a paragraph about the system's history and capabilities, followed by two sections: "Geography" and "Science and Reference", each listing several questions.

START, the world's first Web-based question answering system, has been on-line and continuously operating since December, 1993. It has been developed by Boris Katz and his associates of the InfoLab Group at the MIT Computer Science and Artificial Intelligence Laboratory. Unlike information retrieval systems (e.g., search engines), START aims to supply users with "just the right information," instead of merely providing a list of hits. Currently, the system can answer millions of English questions about places (e.g., cities, countries, lakes, coordinates, weather, maps, demographics, political and economic systems), movies (e.g., titles, actors, directors), people (e.g., birth dates, biographies), dictionary definitions, and much, much more. Below is a list of some of the things START knows about, with example questions. You can type your question above or select from the following examples. [less...](#)

Geography

- What South-American country has the largest population?
- What's the largest city in Florida?
- Give me the states that border Colorado.
- What cities are within 250 miles of the capital of Italy?
- How many people live in Israel?
- Show me a map of Denmark.
- Which is deeper, the Baltic Sea or the North Sea?
- How far is Mount Kilimanjaro from Mount Everest?
- List some large cities in Argentina.
- Show the capital of the 2nd largest country in Asia.
- How much does it cost to study at MIT?
- More examples...

Science and Reference

- What is Jupiter's atmosphere made of?
- Who first discovered radiocarbon dating?
- How far is Neptune from the sun?
- Why is the sky blue?
- What planet has the smallest surface area?
- How many feet are there in a kilometer?
- Convert 100 dollars into Euros.
- Show me a metro map of Moscow.
- How many languages are spoken in Afghanistan?
- Give me the GDP of Taiwan.
- How is the weather in Boston today?
- More examples...

HOME People Publications How START works

NLP

NLP

Introduction to NLP

Evaluation of QA

The TREC Q&A evaluation

- Run by NIST
 - Voorhees and Tice 1999, 2000
 - 2GB of input
 - AQUAINT corpus
 - 200 questions in 1999
 - 693 questions in 2000
- Essentially fact extraction
 - Who was Lincoln's secretary of state?
 - What does the Peugeot company manufacture?
- Assumptions
 - Questions are based on text
 - Answers are assumed to be present
 - Systems return five passages of 50 or 250 bytes
 - After 2002, only a single passage of 50 bytes + confidence score + NIL
 - No inference needed

TREC 1999

- What date in 1989 did East Germany open the Berlin Wall?
 - Nov 9
- Who was Johnny Mathis' high school track coach?
 - Lou Vasquez
- What is the shape of a porpoises' tooth?
 - spade-shaped
- What is the number of buffaloes thought to have been living in North America when Columbus landed in 1492?
 - 60 million
- The Faroes are a part of what northern European country?
 - Denmark
- The symptoms of Parkinson's disease are linked to the demise of cells in what area of the brain?
 - substantia nigra
- What was the date of the Chelsea flower show in 1992?
 - May 18
- What hotel was used for a setting of the Agatha Christie novel, "And Then There Were None"?
 - Burgh Island Hotel
- What year was the Magna Carta signed?
 - 1215
- Who was Lincoln's Secretary of State?
 - William Seward
- How long does it take to fly from Paris to New York in a Concorde?
 - 3 1/2 hours

TREC 1999 – Test Questions

- Who is the author of the book, "The Iron Lady: A Biography of Margaret Thatcher"?
- What was the monetary value of the Nobel Peace Prize in 1989?
- What does the Peugeot company manufacture?
- How much did Mercury spend on advertising in 1993?
- What is the name of the managing director of Apricot Computer?
- Why did David Koresh ask the FBI for a word processor?
- What debts did Qintex group leave?
- What is the name of the rare neurological disease with symptoms such as: involuntary movements (tics), swearing, and incoherent vocalizations (grunts, shouts, etc.)?
- How far is Yaroslavl from Moscow?
- Name the designer of the shoe that spawned millions of plastic imitations, known as "jellies".
- Who was President Cleveland's wife?
- How much did Manchester United spend on players in 1993?
- How much could you rent a Volkswagen bug for in 1966?
- What country is the biggest producer of tungsten?
- When was London's Docklands Light Railway constructed?
- What two US biochemists won the Nobel Prize in medicine in 1992?
- How long did the Charles Manson murder trial last?
- Who was the first Taiwanese President?
- Who was the leader of the Branch Davidian Cult confronted by the FBI in Waco, Texas in 1993?
- Where is Inoco based?
- Who was the first American in space?
- When did the Jurassic Period end?
- When did Spain and Korea start ambassadorial relations?
- When did Nixon visit China?
- Who was the lead actress in the movie "Sleepless in Seattle"?
- What is the name of the "female" counterpart to El Nino, which results in cooling temperatures and very dry weather?
- What is the brightest star visible from Earth?
- What are the Valdez Principles?

Evaluation

- MRR
 - Mean Reciprocal Rank
 - Introduced by TREC in 1999
- Example
 - What is the capital of Canada?
 - 1.Toronto, 2.Ottawa, 3.Albany, 4.Philadelphia, 5.Ottawa
 - Correct answer ranks = 2, 5 (highest rank = 2)
 - MRR = $1/2 = 0.5$
 - TRR = $1/2 + 1/5 = 0.7 = \text{total reciprocal rank}$
- Later years
 - Confidence-weighted score

$$\text{MRR} = \frac{1}{n} \left(\sum_{i=1}^n \frac{1}{rank_i} \right)$$

Performance

- 1999
 - Cymphony, SMY, AT&T, IBM, XRCE, Umd
- 2000
 - SMU, ISI, Waterloo, IBM, CUNY-QC
- 2001
 - InsightSoft, LCC, Oracle, ISI
- 2002
 - LCC, InsightSoft, NUS
- 2003
 - LCC, NUS, LexiClone, ISI
- 2004
 - LCC, NUS, Uwales, IBM
- 2005
 - LCC, NUS, IBM, Albany

Other Types of Questions

- **Definitional**
 - What is a boll weevil?
- **List**
 - Which states signed the US Declaration of Independence?
- **Crosslingual**
 - E.g., questions in Spanish, documents in English
- **Series**
 - Next slide

TREC Series Questions

- What are prions made of?
- Who discovered prions?
- What researchers have worked with prions?
- Who is the lead singer/musician in Nirvana?
- Who are the band members?
- When was the band formed?
- What is their biggest hit?
- What are their albums?
- What style of music do they play?
- What industry is Rohm and Haas in?
- Where is the company located?
- What is their annual revenue?
- How many employees does it have?
- What kind of insect is a boll weevil?
- What type of plant does it damage?
- What states have had problems with boll weevils?

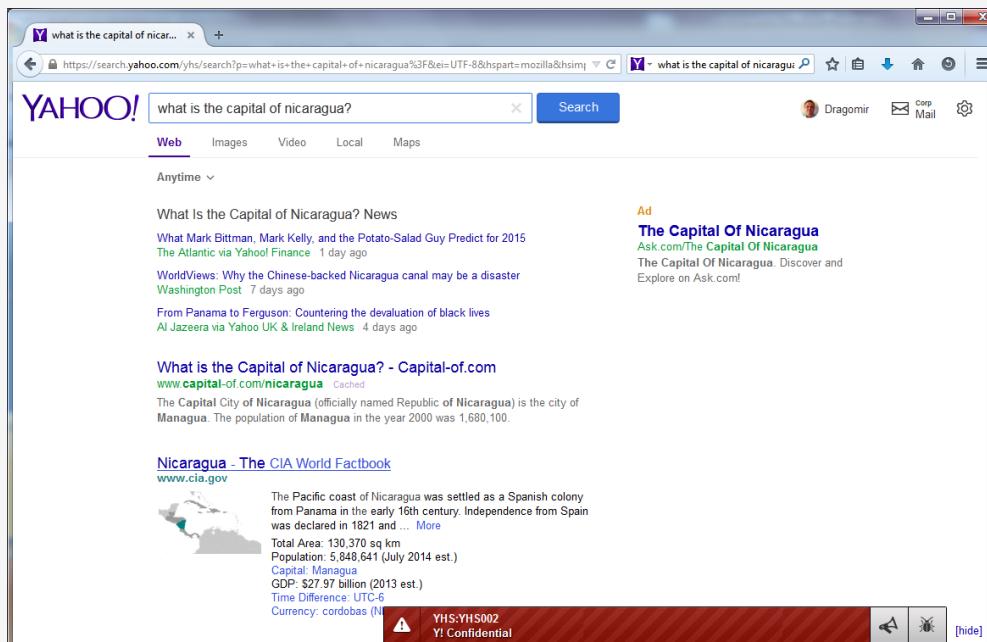
NLP

Introduction to NLP

System Architecture

System Architecture

- Many questions can be answered by traditional search engines



What is the Largest City in Northern Afghanistan?

... Afghanistan, Kabul 2,450 ... Administrative capital and **largest city** (1997 est ... Undetermined.

Panama, Panama City 450,668. ... of the Gauteng, **Northern** Province, Mpumalanga ...

www.infoplease.com/cgi-bin/id/A0855603

... died in Kano, northern Nigeria's **largest city**, during two days of anti-American riots

led by Muslims protesting the US-led bombing of **Afghanistan**, according to ...

www.washingtonpost.com/wp-dyn/print/world/

... air strikes on the **city**. ... the Taliban militia in **northern Afghanistan** in a significant

blow ... defection would be the **largest** since the United States ...

www.afgha.com/index.php - 60k

... Kabul is the capital and **largest city of Afghanistan**. . . . met. area pop. 2,029,889),

is the **largest city** in Uttar Pradesh, a state in **northern** India. . . .

school.discovery.com/homeworkhelp/worldbook/atozgeography/k/k1menu.html

... Gudermes Chechnya's second **largest** town. The attack ... location in **Afghanistan's** outlying

regions ... in the **city** of Mazar-i-Sharif, a **Northern** Alliance-affiliated ...

english.pravda.ru/hotspots/2001/09/17/

... Get Worse By RICK BRAGG Pakistan's **largest city** is getting a jump on the ... Region: Education

Offers Women in **Northern Afghanistan** a Ray of Hope. ...

www.nytimes.com/pages/world/asia/

... within three miles of the airport at Mazar-e-Sharif, the **largest city** in **northern**

Afghanistan, held since 1998 by the Taliban. There was no immediate comment ...

uk.fc.yahoo.com/photos/a/afghanistan.html

System Components

- Source identification
 - semi-structured vs. text sources
- Query modulation
 - best paraphrase of a NL question given the syntax of a search engine
 - Example: Who wrote Hamlet → author | wrote Hamlet
- Document retrieval
- Sentence ranking
 - n-gram matching, Okapi
- Answer extraction
 - question type classification
 - phrase chunking
- Answer ranking
 - question type, proximity to query words, frequency

What is the largest city in Northern Afghanistan?

Query modulation

(largest OR biggest) city "Northern Afghanistan"

Document retrieval

www.infoplease.com/cgi-bin/id/A0855603

www.washingtonpost.com/wp-dyn/print/world/

Sentence ranking

Gudermes, Chechnya's second **largest** town ... location in **Afghanistan's** outlying regions
within three miles of the airport at Mazar-e-Sharif, the **largest city** in **northern** Afghanistan

Answer extraction

Gudermes
Mazer-e-Sharif

Answer ranking

Mazer-e-Sharif
Gudermes

Question Type Classification

- Can help find the right answers in the text
- Example
 - Who wrote Anna Karenina?
 - Looking for a PERSON/INDIVIDUAL/WRITER

SYN-classes (IBM AnSel)

QA-token	Question type	Example
PLACE	Where?	In the Rocky Mountains
COUNTRY	Where? What country?	United Kingdom
STATE	Where? What state?	Massachusetts
PERSON	Who?	Albert Einstein
ROLE	Who?	Doctor
NAME	Who? What? Which?	The Shakespeare Festival
ORG	Who? What?	The U.S. Post Office
DURATION	How long?	For 5 centuries
AGE	How old?	30 years old
YEAR	When? What year?	1999
TIME	When? What time?	In the afternoon
DATE	When? What date?	July 4 th , 1776
VOLUME	How big? How large?	3 gallons
AREA	How big? How large?	4 square inches
LENGTH	How long? How big?	3 miles
WEIGHT	How heavy? How big?	25 tons
NUMBER	How many?	1,134.5
RATE	How much? What percentage?	50 per cent
MONEY	How much?	4 million dollars

UIUC Question Types

- **ENTITY:** entities
 - animal: animals
 - body: organs of body
 - color: colors
 - creative: inventions, books and other creative pieces
 - currency: currency names
 - dis.med.: diseases and medicine
 - event: events
 - food: food
 - instrument: musical instrument
 - lang: languages
 - letter: letters like a-z
 - other: other entities
 - plant: plants
 - product: products
 - religion: religions
 - sport: sports
 - substance: elements and substances
 - symbol: symbols and signs
 - technique: techniques and methods
 - term: equivalent terms
 - vehicle: vehicles
 - word: words with a special property
- **ABBREVIATION:** abbreviation
 - abb: abbreviation
 - exp: expression abbreviated
- **DESCRIPTION:** description and abstract concepts
 - definition: definition of sth.
 - description: description of sth.
 - manner: manner of an action
 - reason: reasons
- **HUMAN:** human beings
 - group: a group or organization of persons
 - ind: an individual
 - title: title of a person
 - description: description of a person
- **LOCATION:** locations
 - city: cities
 - country: countries
 - mountain: mountains
 - other: other locations
 - state: states
- **NUMERIC:** numeric values
 - code: postcodes or other codes
 - count: number of sth.
 - date: dates
 - distance: linear measures
 - money: prices
 - order: ranks
 - other: other numbers
 - period: the lasting time of sth.
 - percent: fractions
 - speed: speed
 - temp: temperature
 - size: size, area and volume
 - weight: weight

UIUC Examples

- **NUM:date**
 - When did Rococo painting and architecture flourish ?
- **LOC:country**
 - What country 's national passenger rail system is called Via ?
- **HUM:ind**
 - Who invented Make-up ?
- **DESC:desc**
 - What is the origin of the word `` attic '' ?
- **DESC:desc**
 - What did Delilah do to Samson 's hair ?
- **ENTY:animal**
 - What kind of animals were in the Paleozoic era ?
- **HUM:ind**
 - Which of the following was Rhodes Scholar ?
- **HUM:ind**
 - Who comprised the now-defunct comic book team known as the Champions ?
- **DESC:**
 - manner How do you make a paintball ?
- **LOC:**
 - state What U.S. state is Fort Knox in ?
- **ENTY:**
 - animal What is a female rabbit called ?
- **LOC:mount**
 - Where is the highest point in Japan ?
- **DESC:desc**
 - Where do chihuahuas come from ?
- **LOC:other**
 - Where does Barney Rubble go to work after he drops Fred off in the `` Flintstones '' cartoon series?

UIUC Papers On Question Classification

- Xin Li and Dan Roth. Experimental Data for Question Classification
 - <http://l2r.cs.uiuc.edu/~cogcomp/Data/QA/QC/>
- Xin Li, Dan Roth. Learning Question Classifiers: The Role of Semantic Information
 - <http://l2r.cs.uiuc.edu/~danr/Papers/LiRo05a.pdf>
 - <http://l2r.cs.uiuc.edu/~cogcomp/asoftware.php?skey=LBJ>
- Data set
 - training (http://l2r.cs.uiuc.edu/~cogcomp/Data/QA/QC/train_5500.label)
 - test (http://l2r.cs.uiuc.edu/~cogcomp/Data/QA/QC/TREC_10.label)

NLP

NLP

Techniques for Question Classification

- Classification task
 - Use standard techniques
- Regular expressions
 - WHO is|was -> PERSON
 - WHICH STATE -> STATE

Query Formulation

- Aka Query Modulation
- Determine which words to include in the IR query

No.	Operator	Paraphrase	TRDR
0	IDENTITY	What country is the " biggest producer " of tungsten	0.837
1	DEL_WH	country is the " biggest producer " of tungsten	1.755
2	DEL_AUX	What country the " biggest producer " of tungsten	1.322
3	DEL_ART	What country is " biggest producer " of tungsten	1.436
4	DEL_PREP	What country is the " biggest producer " tungsten	1.436
5	DEL_STOP	What country is the " biggest producer " of tungsten	0.837
6	REPL_IN	What (" administrative district " OR " political unit " OR people OR region OR " geographical area " OR " rural area ") is the " biggest producer " of tungsten	1.181
7	REPL_2N	What country is the "biggest producer" of (" metallic element" OR wolfram OR w OR " atomic number 74")	1.419

Passage Retrieval

- Features
 - Proper nouns that match the query
 - Near each other
 - Entities that match the expected answer type

Answer Retrieval

- Use NER to identify the matching phrases
 - E.g., “January 1, 1951” as a DATE
- Features
 - Distance to query words
 - Answer type
 - Wordnet similarity
 - Redundancy

Redundancy

- *What is the capital of Spain?*
 - *Madrid is the capital of Spain*
 - *En route to Spain's capital of Madrid*
 - *Madrid, Spain's capital city is situated almost at the geographical epicentre of the country*
 - *The capital of Spain is Madrid*
 - *Madrid, Spain's sunny capital*
 - *Madrid became Spain's capital*
 - *In 1561, it was elevated to status as Spain's capital city*
 - *Madrid has been the capital of Spain since 1562*
 - *Madrid, the physical and cultural capital of Spain*

Redundancy

- When did French revolutionaries storm the Bastille?
 - The storming of the Bastille occurred in Paris on the morning of 14 July 1789
 - The storming of the Bastille, 14 July 1789
 - The storming of the Bastille prison on July 14th 1789 was an event that paved the way to further civil disorder and upheaval in France.
 - French revolutionaries storm Bastille, 1789
 - The storming of the Bastille (Louis XVI's prison) by French revolutionaries took place on July 14th, 1789
- Who killed Mahathma Gandhi?
 - Mohandas Karamchand Gandhi (often called Mahatma Gandhi) was assassinated on 30 January 1948, shot at point-blank range by Nathuram Godse.
 - Nathuram Godse killed Gandhi
 - Godse killed Gandhi
 - Mahatma Gandhi was assassinated on 30 January 1948, shot at point-blank range by Nathuram Godse
 - Mohandas Gandhi was shot dead by Nathuram Godse, a Hindu extremist
 - Godse assassinated Mahatma Gandhi on January 30, 1948, approaching him during the evening prayer, bowing, and shooting him three times at close range with a Beretta semi-automatic pistol.

On the Web

- Significantly larger corpus
- No pre-annotation is possible
- Search engines only partially helpful
 - Stop words
 - Question types
 - Restrictions on queries
 - Issues with reliability, timeliness, inaccurate answers

NLP

NLP

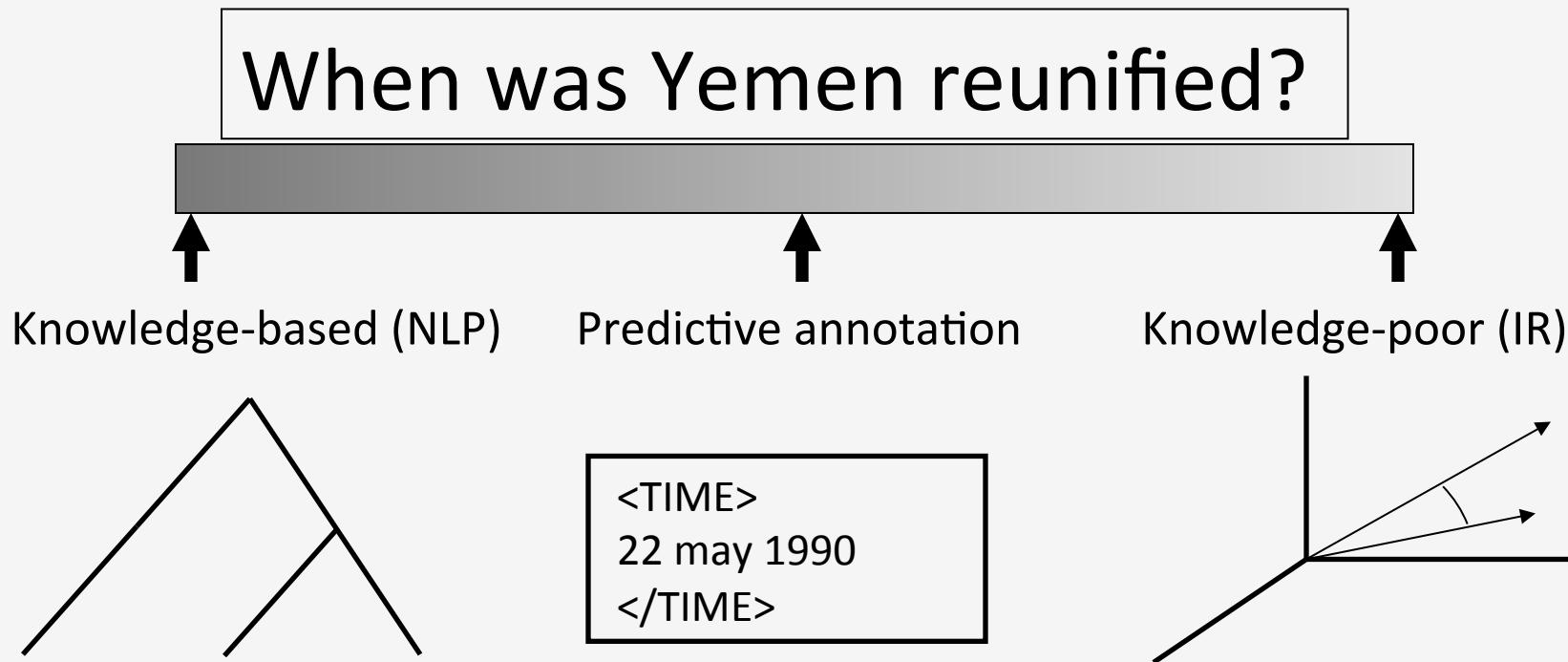
Introduction to NLP

Question Answering Systems

AnSel (Prager et al. 1999)

- IBM System
- Built for TREC
- Components
 - Predictive Annotation
 - Logistic Regression

Predictive Annotation



Predictive Annotation

```
<p><NUMBER>1</NUMBER></p>

<p><QUERY>Who is the author of the book, "The Iron Lady: A Biography of Margaret Thatcher"?</QUERY></p>

<p><PROCESSED_QUERY>@excwin(*dynamic* @weight(2.00001001 *Iron_Lady) @weight(200
Biography_of_Margaret_Thatcher) @weight(200 Margaret) @weight(100 author) @weight(100 book)
@weight(100 iron) @weight(100 lady) @weight(100 :) @weight(100 biography) @weight(100
thatcher) @weight(400 @syn(PERSON$ ORG$ NAME$ ROLE$) ) )</PROCESSED_QUERY></p>

<p><DOC>LA090290-0118</DOC></p>

<p><SCORE>1020.8114</SCORE></p>

<TEXT><p>THE IRON LADY; A <span class="NAME"> Biography of Margaret Thatcher </span> by
<span class="PERSON"> Hugo Young </span> (<span class="ORG"> Farrar , Straus & Giroux </span>
) The central riddle revealed here is why, as a woman <span class="PLACEDEF"> in a
man </span> 's world, <span class="PERSON"> Margaret Thatcher </span> evinces such an
exclusionary attitude toward women.</p></TEXT>
```

Some Observations

- In documents that contain the answers, the query terms tend to occur in close proximity to each other
- The answers to fact-seeking questions are usually phrases
- These phrases can be categorized by question type
- The phrases can be identified in text by pattern matching techniques

Feature Selection

Avgdst: the average distance in words between the beginning of the span and the words in the query that also appear in the passage. Example: given the question “Who was Johnny Mathis' high school track coach?” and the passage “*Tim O'Donohue, Woodbridge High School's varsity baseball coach, resigned Monday and will be replaced by assistant Johnny Ceballos, Athletic Director Dave Cowen said.*” and the span “*Tim O'Donohue*”, the value of **avgdst** is equal to 8.

Notinq: the number of words in the span that do not appear in the query. Example: **Notinq** (“*Woodbridge high school*”) = 1, because both “high” and “school” appear in the query while “Woodbridge” does not. It is set to –100 when the actual value is 0.

Frequency: number of times a given span appears in the hit list.

Sscore: passage relevance as computed by the search engine.

Number: position of the span among all spans returned. Example: “*Lou Vasquez*” was the first span returned by GuruQA on the sample question.

Rspanno: position of the span among all spans returned within the current passage.

Count: number of spans of any span class retrieved within the current passage.

Type: the position of the span type in the list of potential span types. Example: **Type** (“*Lou Vasquez*”) = 1, because the span type of “*Lou Vasquez*”, namely “PERSON” appears first in the list of potential span types, “PERSON ORG NAME ROLE”.

Span	Type	Number	Rspanno	Count	Notinq	Type	Avgdst	Sscore	TOTAL
<i>Lou Vasquez</i>	PERSON	1	1	6	2	1	16	0.02507	-9.93
Tim O'Donohue	PERSON	17	1	4	2	1	8	0.02257	-12.57
Athletic Director Dave Cowen	PERSON	23	6	4	4	1	11	0.02257	-15.87
Johnny Ceballos	PERSON	22	5	4	1	1	9	0.02257	-19.07
Civic Center Director Martin Durham	PERSON	13	1	2	5	1	16	0.02505	-19.36
Johnny Hodges	PERSON	25	2	4	1	1	15	0.02256	-25.22
Derrick Evans	PERSON	33	4	4	2	1	14	0.02256	-25.37
NEWSWIRE Johnny Majors	PERSON	30	1	4	2	1	17	0.02256	-25.47
Woodbridge High School	ORG	18	2	4	1	2	6	0.02257	-28.37
Evan	PERSON	37	6	4	1	1	14	0.02256	-29.57
Gary Edwards	PERSON	38	7	4	2	1	17	0.02256	-30.87
O.J. Simpson	NAME	2	2	6	2	3	12	0.02507	-37.40
South Lake Tahoe	NAME	7	5	6	3	3	14	0.02507	-40.06
Washington High	NAME	10	6	6	1	3	18	0.02507	-49.80
Morgan	NAME	26	3	4	1	3	12	0.02256	-52.52
Tennesseefootball	NAME	31	2	4	1	3	15	0.02256	-56.27
Ellington	NAME	24	1	4	1	3	20	0.02256	-59.42
assistant	ROLE	21	4	4	1	4	8	0.02257	-62.77
the Volunteers	ROLE	34	5	4	2	4	14	0.02256	-71.17
Johnny Mathis	PERSON	4	4	6	-100	1	11	0.02507	-211.33
Mathis	NAME	14	2	2	-100	3	10	0.02505	-254.16
coach	ROLE	19	3	4	-100	4	4	0.02257	-259.67

IONAUT (Abney et al. 2000)

- Passage retrieval
 - Uses START (Salton, Buckley)
- Entity recognition
 - Uses Cass (Abney) – partial parser
- Entity classification
 - Simple patterns for 8 question types

Mulder (Kwok et al. 2001)

- First large-scale Web QA system
- Components
 - Maximum entropy parser (Charniak)
 - PC-Kimmo for unknown words
 - Link parser (Sleator and Temperley)
 - Google
- Tokenization
 - phrases in quotes
- Query transformations
 - “When did Nixon visit China” → “Nixon visited China”

NSIR (Radev et al. 2002)

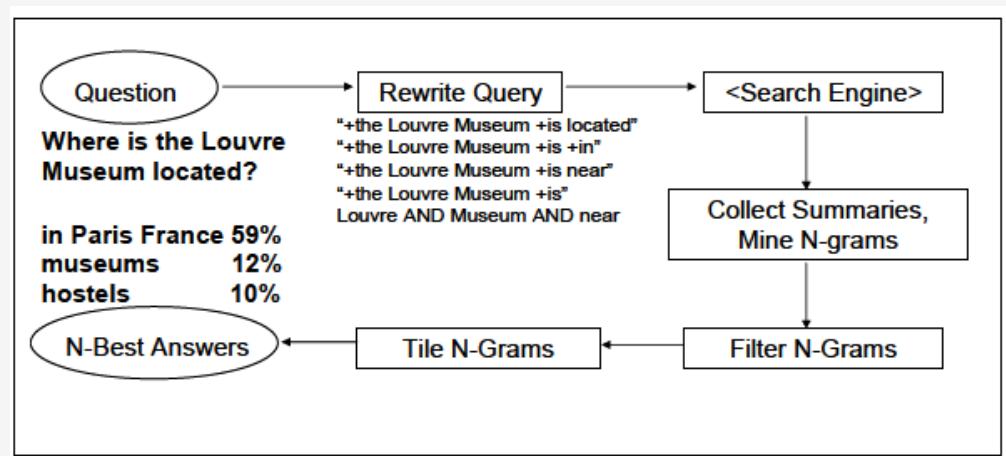
- Probabilistic phrase reranking
 - $P(\text{qtype}|\text{signature})$
 - Signature = POS sequence (e.g., “NNP NNP” for “Bill Gates”)
- Search engines
 - AlltheWeb, NorthernLight, Altavista, Google

AskMSR (Banko et al. 2002)

- Assumption
 - Someone has already answered this question on the Web

- Components
 - Query rewriting
 - Snippet retrieval
 - N-gram ranking

- Tiling matches
 - Combining A B C and B C D into A B C D
 - E.g., “Mr. Charles” and “Charles Dickens” into “Mr. Charles Dickens”



Echihabi and Marcu

- Based on the noisy-channel model
- Find the sentence S that maximizes $p(q|S)$
- Requires simplifying the sentences

NLP

NLP

LCC (e.g. Moldovan et al. 2003)

- Uses logic form transformations
- Uses axioms for inference
 - e.g. Lexical chains

Example:

Heavy selling of Standard & Poor's 500-stock index futures in Chicago relentlessly beat stocks downward.

LF:

heavyJJ(x1) & sellingNN(x1) & ofIN(x1,x6) &
StandardNN(x2) & &CC(x13,x2,x3) & PoorNN(x3)
&'sPOS(x6,x13) & 500-stockJJ(x6) & indexNN(x4)
& futureNN(x5) & nnNNC(x6,x4,x5) & inIN(x1,x8)
& ChicagoNN(x8) & relentlesslyRB(e12) &
beatVB(e12,x1,x9) & stocksNN(x9) & downwardRB(e12)

Q1394: What country did the game of croquet originate in?

Answer: Croquet is a 15th-century French sport that has largely been dominated by older, wealthier people who play at exclusive clubs.

Lexical chains:

- (1) game:n#3 → HYPERNYM → recreation:n#1 → HYPONYM → sport:n#1
- (2) originate.in:v#1 → HYPONYM → stem:v#1 → GLOSS → origin:n#1 → GLOSS → be:v#1

QASM (Radev & al. 2001)

- Noisy channel model
- Convert natural language question into query
 - What country is the biggest producer of tungsten?
 - (biggest OR largest) producer tungsten
- Channel operators
 - DELETE, e.g., delete prepositions, stop words
 - REPLACE, e.g., replace a noun phrase with a WordNet expansion
 - DISJUNCT, e.g., replace a noun phrase with a disjunction

Ravinchandran and Hovy 2002

- Characteristics
 - Automatically learn surface patterns
 - Starts with a seed
 - Query Web
 - Find patterns that contain both the question and the answer terms
- Example
 - Mozart was born in 1756
 - <NAME> was born on <BIRTHDATE>

Watson (Ferrucci et al. 2010)

- Jeopardy winner (2011)
- Architecture
 - Uses "DeepQA": a technology that enables computer systems to directly and precisely answer natural language questions over an open and broad range of knowledge
 - 10 racks of IBM Power 750 servers running Linux
 - 16 terabytes of RAM
 - 2,880 processor cores
 - Capable of operating at 80 teraflops.
 - Mostly in Java but also some C++ and Prolog
 - Integrated using UIMA

Ferrucci et al. 2010. Building Watson: An Overview of the DeepQA Project. AI Magazine. Fall 2010. 59-79.

<http://www.pcmag.com/article2/0,2817,2380351,00.asp>

<http://blog.reddit.com/2011/02/ibm-watson-research-team-answers-your.html>

Watson (Ferrucci et al. 2000)

- Knowledge sources
 - 200 million pages of structured and unstructured content consuming four terabytes of disk storage
 - Including Wikipedia, WordNet, Yago
- Betting strategy
 - Buzz if at least 50% certain
- Performance
 - Watson answers 66 correct and 9 incorrect.
 - Watson's two day winning streak was \$77,147. Ken Jennings ended with \$24,000 and Brad Rutter with \$21,600.

<http://www.quora.com/What-questions-were-asked-in-the-Jeopardy-episode-involving-Watson>

<http://theswimmingsubmarine.blogspot.com/2011/02/how-ibms-deep-question-answering.html>

<http://www.theatlantic.com/technology/archive/2011/02/is-it-time-to-welcome-our-new-computer-overlords/71388/>

Watson (Ferrucci et al. 2010)

- Question types
 - 2,500 of them
 - 200 of them are very common
- Jeopardy Question Archive
 - <http://j-archive.com/>

QA Challenges

- Word Sense Disambiguation
- Co-reference Resolution
- Semantic Role Labeling
- Temporal questions
- Categories on Jeopardy

NLP

NLP

Introduction to NLP

Summarization

Text Summarization

Health Benefits

- Eating a diet rich in vegetables and fruits as part of an overall healthy diet may reduce risk for heart disease, including heart attack and stroke.
- Eating a diet rich in some vegetables and fruits as part of an overall healthy diet may protect against certain types of cancers.
- Diets rich in foods containing fiber, such as some vegetables and fruits, may reduce the risk of heart disease, obesity, and type 2 diabetes.
- Eating vegetables and fruits rich in potassium as part of an overall healthy diet may lower blood pressure, and may also reduce the risk of developing kidney stones and help to decrease bone loss.
- Eating foods such as vegetables that are lower in calories per cup instead of some other higher-calorie food may be useful in helping to lower calorie intake.

Nutrients

- Most vegetables are naturally low in fat and calories. None have cholesterol. (Sauces or seasonings may add fat, calories, or cholesterol.)
- Vegetables are important sources of many nutrients, including potassium, dietary fiber, folate (folic acid), vitamin A, and vitamin C.
- Diets rich in potassium may help to maintain healthy blood pressure. Vegetable sources of potassium include sweet potatoes, white potatoes, white beans, tomato products (paste, sauce, and juice), beet greens, soybeans, lima beans, spinach, lentils, and kidney beans.
- Dietary fiber from vegetables, as part of an overall healthy diet, helps reduce blood cholesterol levels and may lower risk of heart disease. Fiber is important for proper bowel function. It helps reduce constipation and diverticulosis. Fiber-containing foods such as vegetables help provide a feeling of fullness with fewer calories.
- Folate (folic acid) helps the body form red blood cells. Women of childbearing age who may become pregnant should consume adequate folate from foods, and in addition 400 mcg of synthetic folic acid from fortified foods or supplements. This reduces the risk of neural tube defects, spina bifida, and anencephaly during fetal development.
- Vitamin A keeps eyes and skin healthy and helps to protect against infections.
- Vitamin C helps heal cuts and wounds and keeps teeth and gums healthy. Vitamin C aids in iron absorption.

Summary

Eating vegetables is healthy.

News Summarization

NewsInEssence: Web-based News Summarization - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Home Search Favorites History

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

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

Interactive Multi-source News Summarization

[Home](#) [Current Clusters](#) [Google Cluster](#) [Summarize Cluster](#) [Track Cluster](#) [User Cluster Archive](#) [CIDR Cluster Archive](#) [Google Cluster Archive](#)

[Help](#) [About NewsInEssence](#) [Contact Us](#)

[CLAIR](#) [MEAD](#) [summarization.com](#)

[7 Articles from 6 Sources] [5 Summaries]

Pressure grows on Bush to globalise Iraq effort

Bangkok Post Friday 1 August, 2003 - US sets mid-2004 target for Iraq elections, two soldiers killed. BBC NEWS World Middle East Bush under fire over Iraq. President George W Bush is coming under increasing pressure from his own Republican Party to disclose how much financing will be required to cover the costs of occupying and reconstructing Iraq.

Recent NIE News Clusters (more)

- 'Ananova - Tensions high as cortene approaches Najaf'
24 articles, 4 summaries: 09/02, 6:10 AM
- 'Bush Makes Push for Manufacturing Jobs'
18 articles, 4 summaries: 09/02, 6:10 AM
- 'Israeli Strike Kills Hamas Member September 1, 2003 21:44:37'
12 articles, 4 summaries: 09/02, 6:10 AM

NIE Headlines
Build your own cluster of articles.

NewsTroll from URL:
URL must be from [CNN](#), [Yahoo](#), [MSNBC](#), [BBC](#), or [USA Today](#).
<http://www.hinduonnet.c>

NewsTroll from query:
lugar bremer bush iraq

Recent User News Clusters (more)

- 'Japan launches asteroid probe'
4 articles, 3 summaries: 09/02, 11:28 AM
- 'Death tax traps 50 more'
1 article, 3 summaries: 09/02, 10:32 AM
- 'Spam peddlers hijack computers'
7 articles, 3 summaries: 09/02, 10:23 AM

[Advanced Options](#)

NIE News Clusters (Archive)

- 'Ananova - Tensions high as cortene approaches Najaf'
24 articles, 4 summaries: 09/02, 6:10 AM
- 'Bush Makes Push for Manufacturing Jobs'
18 articles, 4 summaries: 09/02, 6:10 AM
- 'Israeli Strike Kills Hamas Member September 1, 2003 21:44:37'
12 articles, 4 summaries: 09/02, 6:10 AM
- 'Taliban ambush two Afghan patrols'
9 articles, 4 summaries: 09/02, 6:10 AM
- 'FOXNews.com'
7 articles, 4 summaries: 09/02, 6:10 AM
- 'BBC SPORT Cricket Cricket 'to return to Kashmir'
7 articles, 4 summaries: 09/02, 6:10 AM
- 'Vivendi mulling Bronfman, GE offers; could decide Tuesday'

Pressure grows on Bush to globalise Iraq effort

produced on 09/02, 6:10 AM

2% Summary

Bangkok Post Friday 1 August, 2003 - US sets mid-2004 target for Iraq elections, two soldiers killed ([1:1](#)) BBC NEWS World Middle East Bush under fire over Iraq ([2:1](#)) President George W Bush is coming under increasing pressure from his own Republican Party to disclose how much financing will be required to cover the costs of occupying and reconstructing Iraq. ([2:2](#)) The Bush administration, which is already spending \$4bn a month of US taxpayers' money on the military costs in Iraq alone, looks to be gearing up for a big international appeal for contributions towards the cost of rebuilding Iraq. ([2:10](#))

Bush is facing growing calls from within his own party and from men running for his job to bring more international troops into Iraq amid mounting US casualties and costs. ([4:3](#)) A peacekeeping battalion from Kazakhstan has been deployed

Internet

Book Summaries

- Cliff's notes
- Book a minute (<http://www.rinkworks.com/bookaminute/>)
 - Which book is this?

(Some BOYS crash on an ISLAND.)

Ralph We need a fire. (*They make a fire. It goes out.*)

Ralph We need a fire. (*They make a fire. It goes out.*)

Ralph We need a fire.

Jack Forget the fire. Let's kill each other.

Other Boys Yeah! (*They do.*)

THE END

Movie Summaries

- **Titanic**
 - Beginning with genuine footage of the departure of the Titanic on its fateful voyage, this epic movie tells the events of that tragic night from the perspective of fictional survivor Rose. As an old lady of 100 she recounts her story of duty, love and disaster to a salvage crew searching for a lost gem.
 - Winner of 11 Oscars, James Cameron's effects-driven blockbuster puts a human face on a tragedy of epic proportions by wedding the historical tale of the doomed ocean liner with a fictional romance between two of the ship's ill-fated passengers. Society girl Rose Dewitt Bukate and penniless artist Jack Dawson struggle to survive both the sinking ship -- and the wrath of Rose's wealthy fiancé.

Search Engine Snippets

The screenshot shows a Google search results page for the query "Who wrote Cloud Atlas". The results are presented as snippets, each with a title, a link to a source, and a brief summary.

- Cloud Atlas (novel) - Wikipedia, the free encyclopedia**
[en.wikipedia.org/wiki/Cloud_Atlas_\(novel\)](https://en.wikipedia.org/wiki/Cloud_Atlas_(novel)) ▾ Wikipedia ▾
Cloud Atlas is a 2004 novel, the third book by British author David Mitchell. It consists of six nested stories that take the reader from the remote South Pacific in ...
- David Mitchell (author) - Wikipedia, the free encyclopedia**
[en.wikipedia.org/wiki/David_Mitchell_\(author\)](https://en.wikipedia.org/wiki/David_Mitchell_(author)) ▾ Wikipedia ▾
David Mitchell (author) For other people named David Mitchell, see David Mitchell (disambiguation). David Stephen Mitchell (born 12 January 1969) is an English novelist. He has written six novels, two of which, number9dream (2001) and Cloud Atlas (2004), were shortlisted for the Booker Prize.
- Cloud Atlas (film) - Wikipedia, the free encyclopedia**
[en.wikipedia.org/wiki/Cloud_Atlas_\(film\)](https://en.wikipedia.org/wiki/Cloud_Atlas_(film)) ▾ Wikipedia ▾
Adapted from the 2004 novel Cloud Atlas by David Mitchell, the film has ... It was nominated for a Golden Globe Award for Best Original Score for In the same interview, Lana Wachowski stated "(p)eople will try to will Cloud Atlas to be rejected. ... Best Original Score, Tom Tykwer, Johnny Klimek and Reinhold Heil, Won.
- The Bone Clocks by Cloud Atlas author David Mitchell ...**
www.slate.com/.../the_bone_clocks_by_cloud_atlas_author_david_mitchell... ▾ Slate ▾
Sep 2, 2014 - First things first: Is David Mitchell's new novel The Bone Clocks another Cloud Atlas, a multi-era-spanning masterpiece of literary ambition and ...
- Cloud Atlas Author David Mitchell What a Bloody Mess We ...**
www.motherjones.com/.../david-mitchell-cloud-atlas-bone-... ▾ Mother Jones ▾
Oct 11, 2014 - British novelist David Mitchell is best known as the guy who wrote the great novel that was made into the challenging movie Cloud Atlas.
- David Mitchell (Author of Cloud Atlas) - Goodreads**
www.goodreads.com/author/show/6538289.David_Mitchell ▾ Goodreads ▾
★★★★★ Rating: 4 · 229,493 votes
David Mitchell was born in Southport, Merseyside, in England, raised in Malvern, Worcestershire, and educated at the University of Kent, studying for a degree in ...
- Ways In Which The Movie 'Cloud Atlas' Has Changed My ...**
www.google.com/imgres?imgurl=http://www.randomhouse.com/catalog/authphoto_330/20870_mitchell_david.jpg&img...id=GWH7Uk4OEvO_zM&zoom=1&tbnh=160&tbnw=120&usg=_FFkg-xc0F8j1o_aygDLnPITCi0=&docid=WiMuI-WYw1dTeM8&itg=1

Genres

- headlines
- outlines
- minutes
- biographies
- abridgments
- sound bites
- movie summaries
- chronologies, etc.

[Mani and Maybury 1999]

Types of Summaries

- Input
- Output
- Purpose
 - Indicative, informative, and critical summaries
- Form
 - Extracts (representative paragraphs/sentences/phrases)
 - Abstracts: “a concise summary of the central subject matter of a document” [Paice90].
- Dimensions
 - Single-document vs. multi-document
- Context
 - Query-specific vs. generic

Stages

- Three stages (typically)
 - content identification
 - conceptual organization
 - realization

BAGHDAD, Iraq (CNN) 6 July 2004 -- Three U.S. Marines have died in al Anbar Province west of Baghdad, the Coalition Public Information Center said Tuesday.

According to CPIC, "Two Marines assigned to [1st] Marine Expeditionary Force were killed in action and one Marine died of wounds received in action Monday in the Al Anbar Province while conducting security and stability operations."

Al Anbar Province -- a hotbed for Iraqi insurgents -- includes the restive cities of Ramadi and Fallujah and runs to the Syrian and Jordanian borders.

Meanwhile, officials said eight people died Monday in a U.S. air raid on a house in Fallujah that American commanders said was used to harbor Islamic militants.

A senior U.S. military official told CNN the target was a group of people suspected of planning suicide attacks using vehicles.

The strike was the latest in a series of raids on the city to target what U.S. military spokesmen have called safehouses for the network led by fugitive Islamic militant leader Abu Musab al-Zarqawi.

A statement from Allawi said: "The people of Iraq will not tolerate terrorist groups or those who collaborate with any other foreign fighters such as the Zarqawi network to continue their wicked ways.

"The sovereign nation of Iraq and our international partners are committed to stopping terrorism and will continue to hunt down these evil terrorists and weed them out, one by one. I call upon all Iraqis to close ranks and report to the authorities on the activities of these criminal cells."

American planes dropped two 1,000-pound bombs and four 500-pound bombs on the house about 7:15 p.m. (11:15 a.m. ET), according to a statement from the U.S.-led Multi-National Force-Iraq.

At least four previous air raids have targeted suspected Zarqawi safehouses in Fallujah.

BAGHDAD, Iraq (CNN) 6 July 2004 -- Three U.S. Marines have died in al Anbar Province west of Baghdad, the Coalition Public Information Center said Tuesday.

According to CPIC, "Two Marines assigned to [1st] Marine Expeditionary Force were killed in action and one Marine died of wounds received in action Monday in the Al Anbar Province while conducting security and stability operations."

Al Anbar Province -- a hotbed for Iraqi insurgents -- includes the restive cities of Ramadi and Fallujah and runs to the Syrian and Jordanian borders.

Meanwhile, officials said eight people died Monday in a U.S. air raid on a house in Fallujah that American commanders said was used to harbor Islamic militants.

A senior U.S. military official told CNN the target was a group of people suspected of planning suicide attacks using vehicles.

The strike was the latest in a series of raids on the city to target what U.S. military spokesmen have called safehouses for the network led by fugitive Islamic militant leader Abu Musab al-Zarqawi.

A statement from Allawi said: "The people of Iraq will not tolerate terrorist groups or those who collaborate with any other foreign fighters such as the Zarqawi network to continue their wicked ways.

"The sovereign nation of Iraq and our international partners are committed to stopping terrorism and will continue to hunt down these evil terrorists and weed them out, one by one. I call upon all Iraqis to close ranks and report to the authorities on the activities of these criminal cells."

American planes dropped two 1,000-pound bombs and four 500-pound bombs on the house about 7:15 p.m. (11:15 a.m. ET), according to a statement from the U.S.-led Multi-National Force-Iraq.

At least four previous air raids have targeted suspected Zarqawi safehouses in Fallujah.

Human Summarization and Abstracting

- What professional abstractors do
- Ashworth (1973):
 - “To take an original article, understand it and pack it neatly into a nutshell without loss of substance or clarity presents a challenge which many have felt worth taking up for the joys of achievement alone. These are the characteristics of an art form”.

Extractive Summarization

- Selecting units of the original text
 - Usually sentences
 - No simplification
 - No rewriting
- Baseline
 - Extract the first few sentences

NLP

NLP

Introduction to NLP

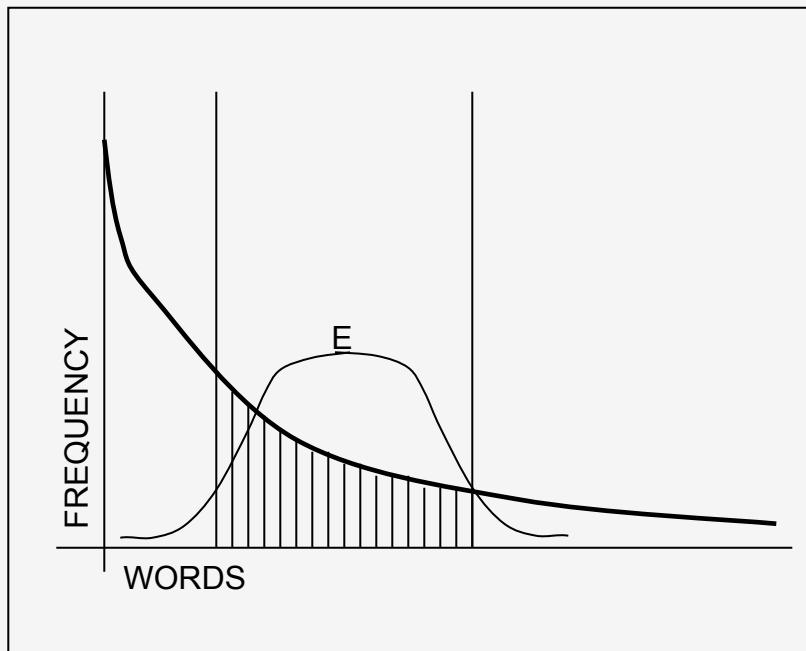
Summarization Techniques 1/3

Baxendale (1958)

- Positional method
 - Analysis of 200 paragraphs
 - Pick the first and last sentences of the paragraph
 - That's where the topic sentences are located
 - Naïve but decent approach

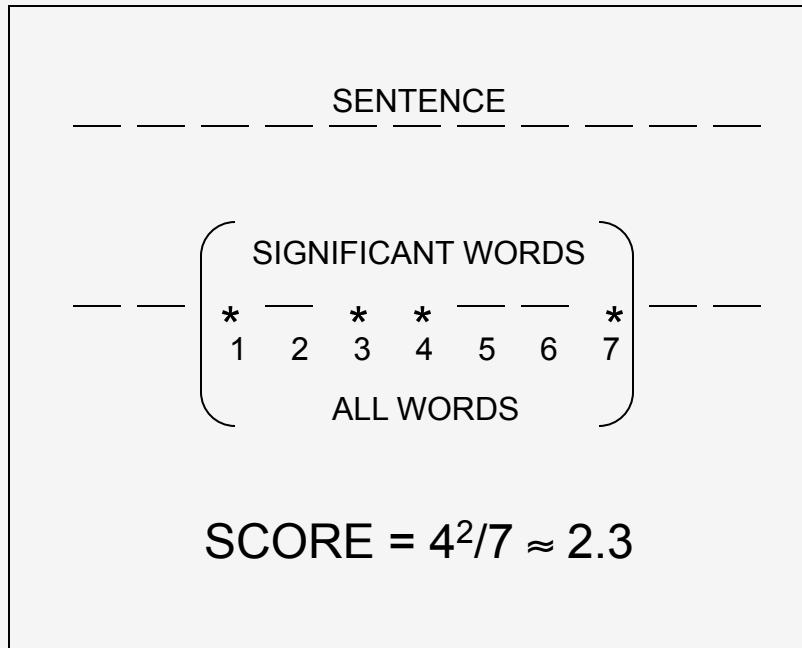
Luhn (1958)

- Technical documents
- Stemming
- Stop words are removed
- Frequency of content terms



Luhn (1958)

- Sentence-level significance factor
- Look for concentrations of salient content terms



Edmundson (1969)

- Technical documents
- Position and frequency
- Cue words (bonus and stigma words)
 - Significant, hardly, impossible
- Document structure
 - Is the sentence a title or heading or right under one of these
- Linear combination of the four features
$$\alpha_1 C + \alpha_2 K + \alpha_3 T + \alpha_4 L$$

Frump (deJong 1979, 1982)

- Knowledge-based
- Slot-filling based on UPI news stories
- Based on 50 sketchy scripts
- Inputs matched to scripts based on manually selected keywords
- Difficult to port to other domains
- Missing scripts for many inputs

Frump

\$demonstration script

- The demonstrators arrive at the demonstration location.
- The demonstrators march.
- Police arrive on the scene.
- The demonstrators communicate with the target of the demonstration.
- The demonstrators attack the target of the demonstration.
- The demonstrators attack the police.

Paice (1990)

- Survey up to 1990
- Techniques that (mostly) failed
 - Syntactic criteria (Earl 1970)
 - Indicator phrases
- Problems with extracts
 - Lack of balance
 - Lack of cohesion

Paice (1990)

- Lack of balance
 - later approaches based on text rhetorical structure
- Lack of cohesion
 - anaphoric reference
 - lexical or definite reference
 - rhetorical connectives
 - recognition of anaphors [Liddy et al. 87]
 - Example: “that” is
 - *nonanaphoric* if preceded by a research-verb (e.g., “demonstrat-”),
 - *nonanaphoric* if followed by a pronoun, article, quantifier,...,
 - *external* if no later than 10th word,
 - else
 - *internal*

Brandow et al. (1995)

- ANES: commercial news from 41 publications
- “Lead” achieves acceptability of 90% vs. 74.4% for “intelligent” summaries
- 20,997 documents
- words selected based on tf*idf (term frequency * inverse document frequency)
- sentence-based features:
 - signature words
 - location
 - anaphora words
 - length of abstract
- Sentences with no signature words are included if between two selected sentences
- Evaluation done at 60, 150, and 250 word length
- Non-task-driven evaluation:
 - “Most summaries judged less-than-perfect would not be detectable as such to a user”

Kupiec et al. (1995)

- First trainable method
 - 20% extract
 - 188 documents from scientific journals
 - Naïve Bayes classifier
- New features
 - Sentence length ($|S| > 5$)
 - Presence of uppercase words (except common acronyms)
 - Thematic words
 - Set of 26 manually fixed phrases
 - Sentence position in paragraph

Kupiec et al. (1995)

- Uses Naïve Bayesian classifier

$$P(s \in S | F_1, F_2, \dots, F_k) = \frac{P(F_1, F_2, \dots, F_k | s \in S) P(s \in S)}{P(F_1, F_2, \dots, F_k)}$$

- Assuming statistical independence

$$P(s \in S | F_1, F_2, \dots, F_k) = \frac{\prod_{j=1}^k P(F_j | s \in S) P(s \in S)}{\prod_{j=1}^k P(F_j)}$$

Kupiec et al. (1995)

- Performance:
 - For 25% summaries, 84% precision
 - For smaller summaries, 74% improvement over Lead

Summons (McKeown & Radev 1995)

- First work on multi-document summarization
- Approach
 - Knowledge-based
 - Information extraction (MUC templates)
 - Text generation

Summons (McKeown & Radev 1995)

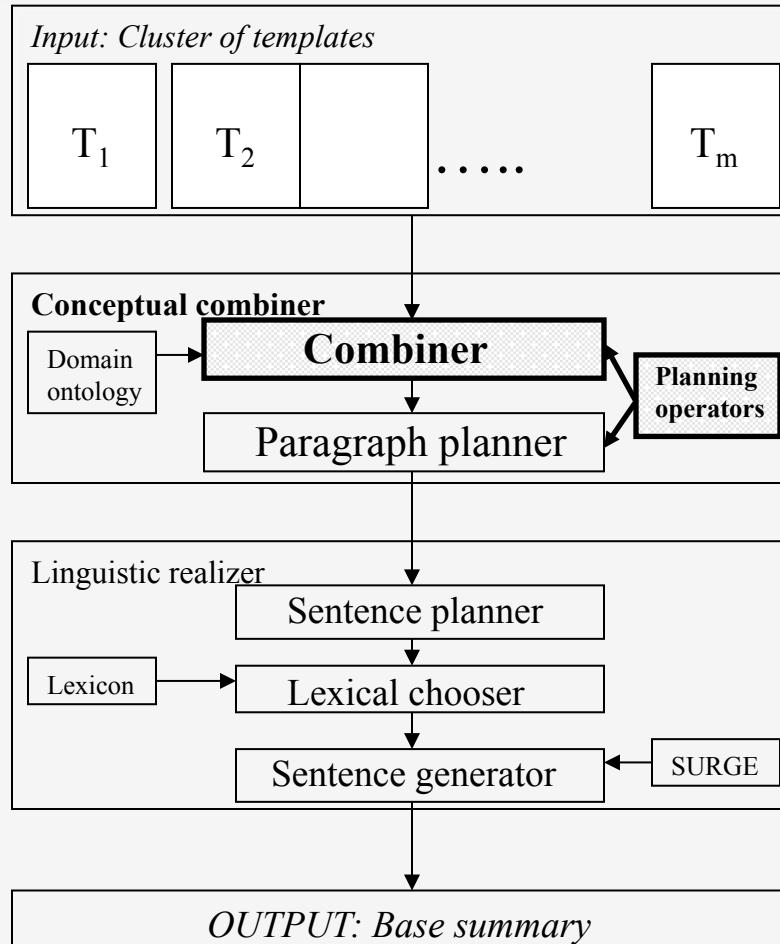
NICOSIA, Cyprus (AP) – Two bombs exploded near government ministries in Baghdad, but there was no immediate word of any casualties, Iraqi dissidents reported Friday. There was no independent confirmation of the claims by the Iraqi National Congress. Iraq's state-controlled media have not mentioned any bombings.

Multiple sources and disagreement

Explicit mentioning of “no information”.

Summons

MESSAGE: ID	TST3-MUC4-0010
MESSAGE: TEMPLATE	2
INCIDENT: DATE	30 OCT 89
INCIDENT: LOCATION	EL SALVADOR
INCIDENT: TYPE	ATTACK
INCIDENT: STAGE OF EXECUTION	ACCOMPLISHED
INCIDENT: INSTRUMENT ID	
INCIDENT: INSTRUMENT TYPE	
PERP: INCIDENT CATEGORY	TERRORIST ACT
PERP: INDIVIDUAL ID	"TERRORIST"
PERP: ORGANIZATION ID	"THE FMLN"
PERP: ORG. CONFIDENCE	REPORTED: "THE FMLN"
PHYS TGT: ID	
PHYS TGT: TYPE	
PHYS TGT: NUMBER	
PHYS TGT: FOREIGN NATION	
PHYS TGT: EFFECT OF INCIDENT	
PHYS TGT: TOTAL NUMBER	
HUM TGT: NAME	
HUM TGT: DESCRIPTION	"1 CIVILIAN"
HUM TGT: TYPE	CIVILIAN: "1 CIVILIAN"
HUM TGT: NUMBER	1: "1 CIVILIAN"
HUM TGT: FOREIGN NATION	
HUM TGT: EFFECT OF INCIDENT	DEATH: "1 CIVILIAN"
HUM TGT: TOTAL NUMBER	



Summons

MESSAGE: ID SECSOURCE: SOURCE SECSOURCE: DATE PRIMSOURCE: SOURCE INCIDENT: DATE INCIDENT: LOCATION INCIDENT: TYPE HUM TGT: NUMBER PERP: ORGANIZATION ID	TST-REU-0001 Reuters March 3, 1996 11:30 March 3, 1996 Jerusalem Bombing “killed: 18” “wounded: 10”	MESSAGE: ID SECSOURCE: SOURCE SECSOURCE: DATE PRIMSOURCE: SOURCE INCIDENT: DATE INCIDENT: LOCATION INCIDENT: TYPE HUM TGT: NUMBER PERP: ORGANIZATION ID	TST-REU-0002 Reuters March 4, 1996 07:20 Israel Radio March 4, 1996 Tel Aviv Bombing “killed: at least 10” “wounded: more than 100”
MESSAGE: ID SECSOURCE: SOURCE SECSOURCE: DATE PRIMSOURCE: SOURCE INCIDENT: DATE INCIDENT: LOCATION INCIDENT: TYPE HUM TGT: NUMBER PERP: ORGANIZATION ID	TST-REU-0003 Reuters March 4, 1996 14:20 March 4, 1996 Tel Aviv Bombing “killed: at least 13” “wounded: more than 100” “Hamas”	MESSAGE: ID SECSOURCE: SOURCE SECSOURCE: DATE PRIMSOURCE: SOURCE INCIDENT: DATE INCIDENT: LOCATION INCIDENT: TYPE HUM TGT: NUMBER PERP: ORGANIZATION ID	TST-REU-0004 Reuters March 4, 1996 14:30 March 4, 1996 Tel Aviv Bombing “killed: at least 12” “wounded: 105”

Summons

Reuters reported that 18 people were killed on *Sunday* in a bombing in Jerusalem. *The next day*, a bomb in Tel Aviv killed at least 10 people and wounded 30 according to Israel radio. Reuters reported that *at least 12 people* were killed and *105* wounded *in the second incident*. *Later the same day*, Reuters reported that Hamas has claimed responsibility for the act.

Summons

- If there are two templates
 - AND
 - the location is the same
 - AND
 - the time of the second template is after the time of the first template
 - AND
 - the source of the first template is different from the source of the second template
 - AND
 - at least one slot differs
 - THEN
 - combine the templates using the contradiction operator...

Summons

Change of perspective

Precondition:

The same source reports a change in a small number of slots

March 4th, Reuters reported that a bomb in Tel Aviv killed at least 10 people and wounded 30. *Later the same day, Reuters* reported that *exactly 12 people* were *actually* killed and *105* wounded.

Summons

Contradiction

Precondition:

Different sources report contradictory values for a small number of slots

The afternoon of February 26, 1993, Reuters reported that a suspected bomb killed *at least six people* in the World Trade Center. *However*, Associated Press announced that *exactly five people* were killed in the blast.

Summons

Refinement

On Monday morning, Reuters announced that a suicide bomber killed at least 10 people in Tel Aviv. *In the afternoon*, Reuters reported that *Hamas* claimed responsibility for the act.

Agreement

The morning of March 1st 1994, *both* UPI and Reuters reported that a man was kidnapped in the Bronx.

Summons

Generalization

According to UPI, three terrorists were arrested in Medellín last Tuesday. Reuters announced that the police arrested two drug traffickers in Bogotá the next day.

A total of five criminals were arrested in Colombia last week.

NLP

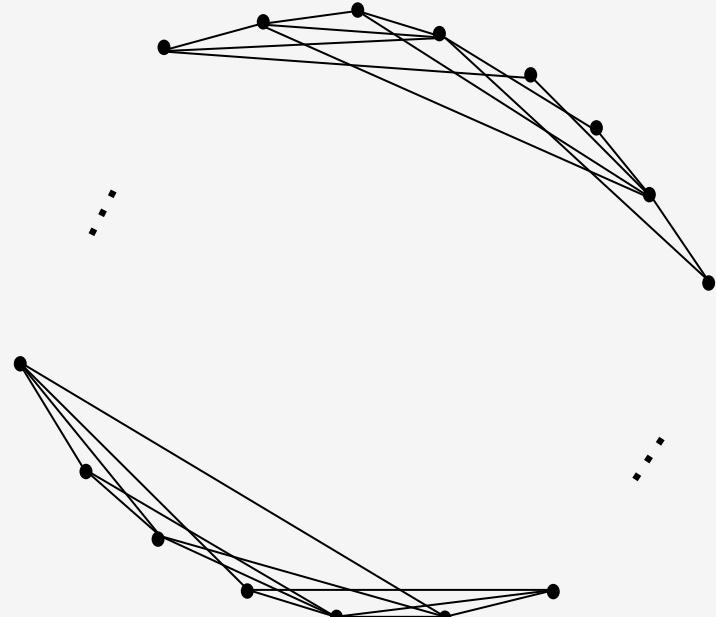
NLP

Introduction to NLP

Summarization Techniques 1/3
(cont'd)

Mitra/Allen/Salton

- Encyclopedia articles
- Semantic hyperlinks
 - Between pairs of paragraphs with lexical similarity above a threshold
 - Paths linking highly connected paragraphs are more likely to contain information central to the topic of the article



NLP

Introduction to NLP

Summarization Techniques 2/3

Mani/Bloedorn (1997, 1999)

- Graph-based method for identifying similarities and differences between documents
- Single event or sequence of events
- Content entities (nodes) and relations (edges)
- Relation types
 - SAME, ADJACENT, ALPHA (WordNet, NetOwl), COREF
- Spreading activation

1.25: "We are clear: the liberation of all our comrades, or we die with all the hostages," a rebel who did not give his name told a local radio station in a telephone call from inside the compound.

1.28: Many leaders of the Tupac Amaru which is smaller than Peru's Maoist Shining Path movement are in jail. 1.29: Its chief Victor Polay, was captured in June 1992 and is serving a life sentence, as is his lieutenant, Peter Cardenas.

1.30: Other top commanders conceded defeat and surrendered in July 1993.

Barzilay & M. Elhadad (1997)

- Lexical chains

Mr. Kenny is the person that invented the anesthetic machine which uses micro-computers to control the rate at which an anesthetic is pumped into the blood. Such machines are nothing new. But his device uses two micro-computers to achieve much closer monitoring of the pump feeding the anesthetic into the patient.

Barzilay and M. Elhadad (1997)

- WordNet-based
- Three types of relations:
 - extra-strong (repetitions)
 - strong (WordNet relations)
 - medium-strong (link between synsets is longer than one + some additional constraints)

Barzilay and M. Elhadad (1997)

- Scoring chains:
 - Length
 - Homogeneity index:
 $= 1 - \# \text{ distinct words in chain}$

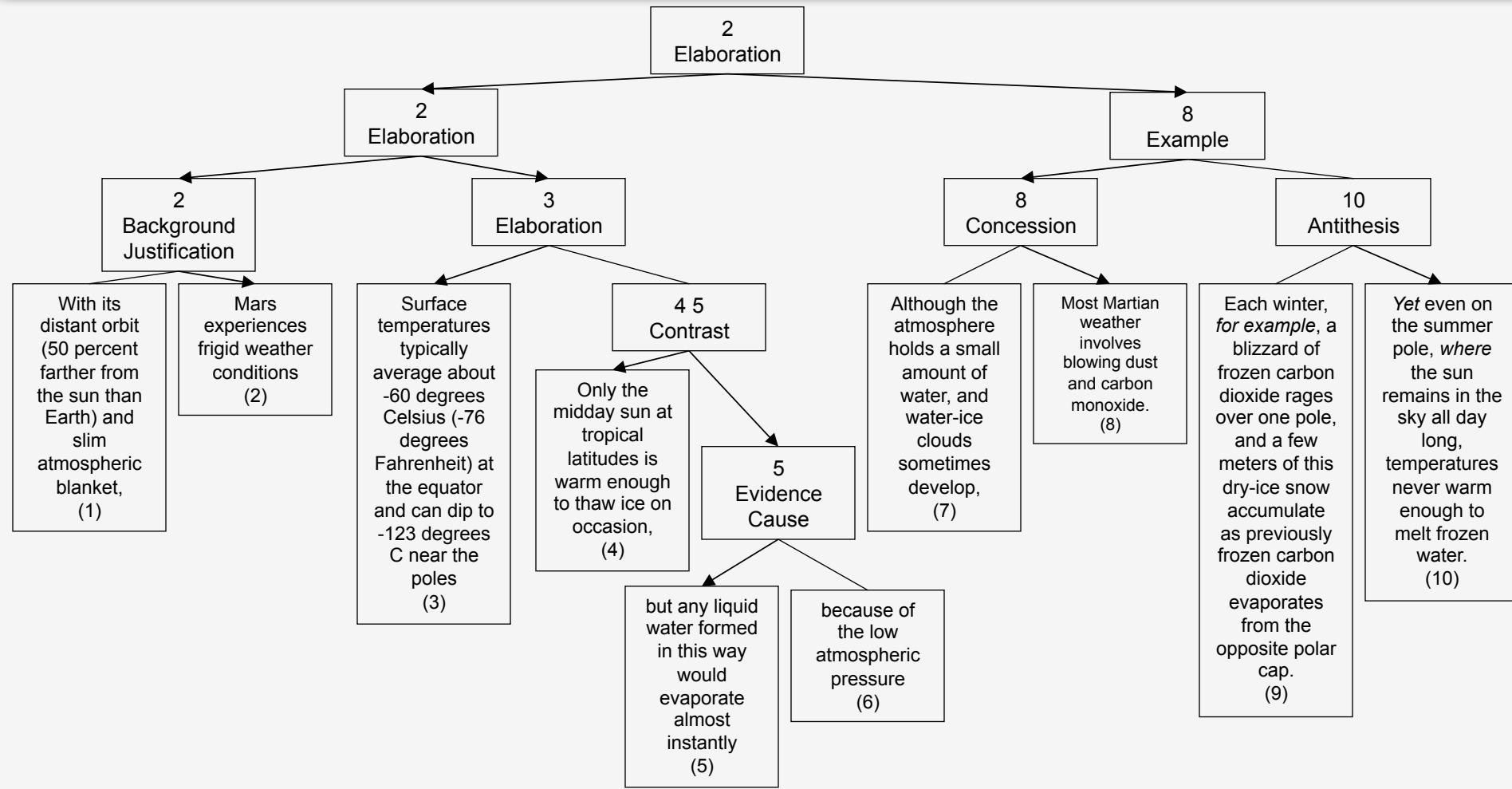
Score = Length * Homogeneity

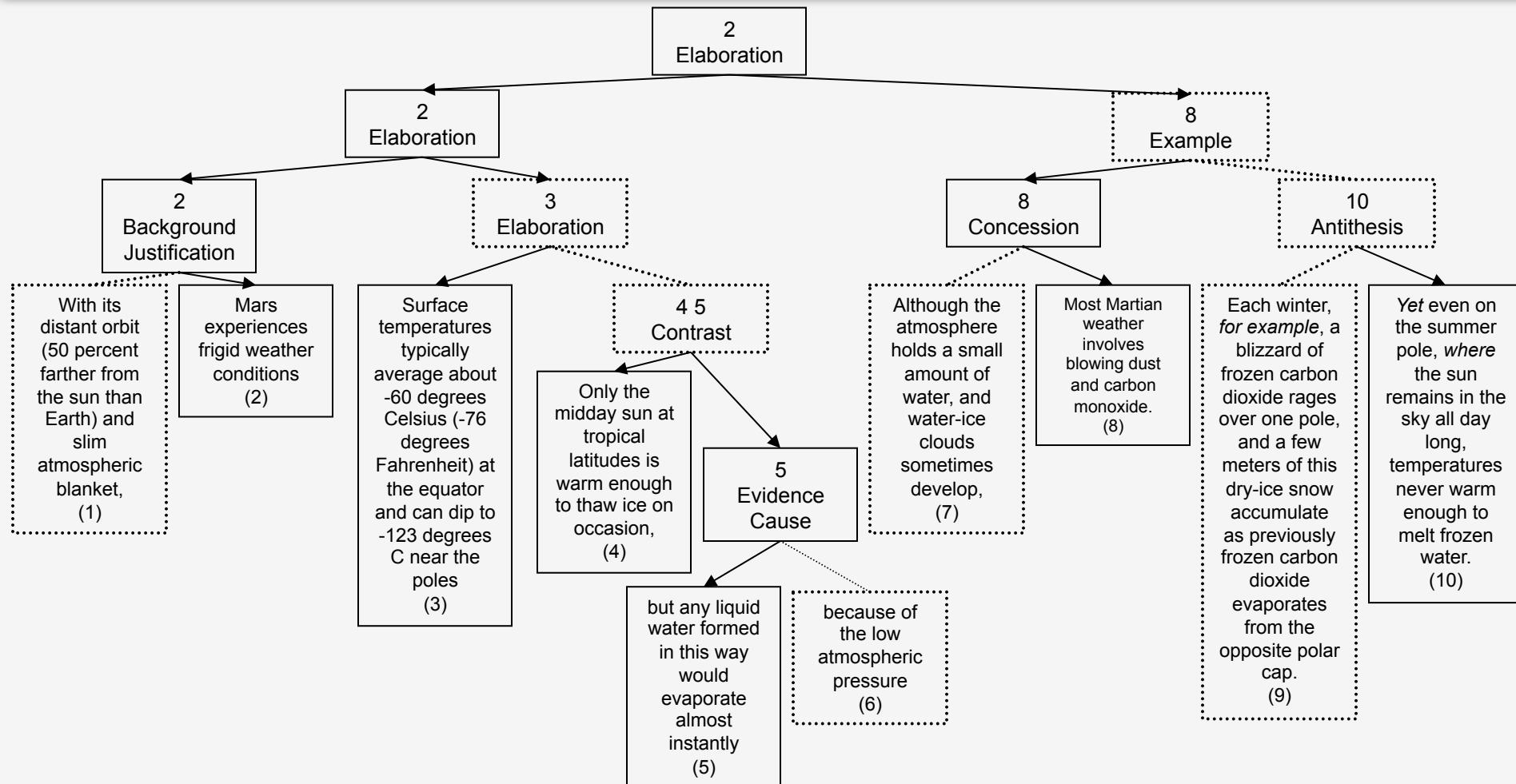
Score > Average + 2 * st.dev.

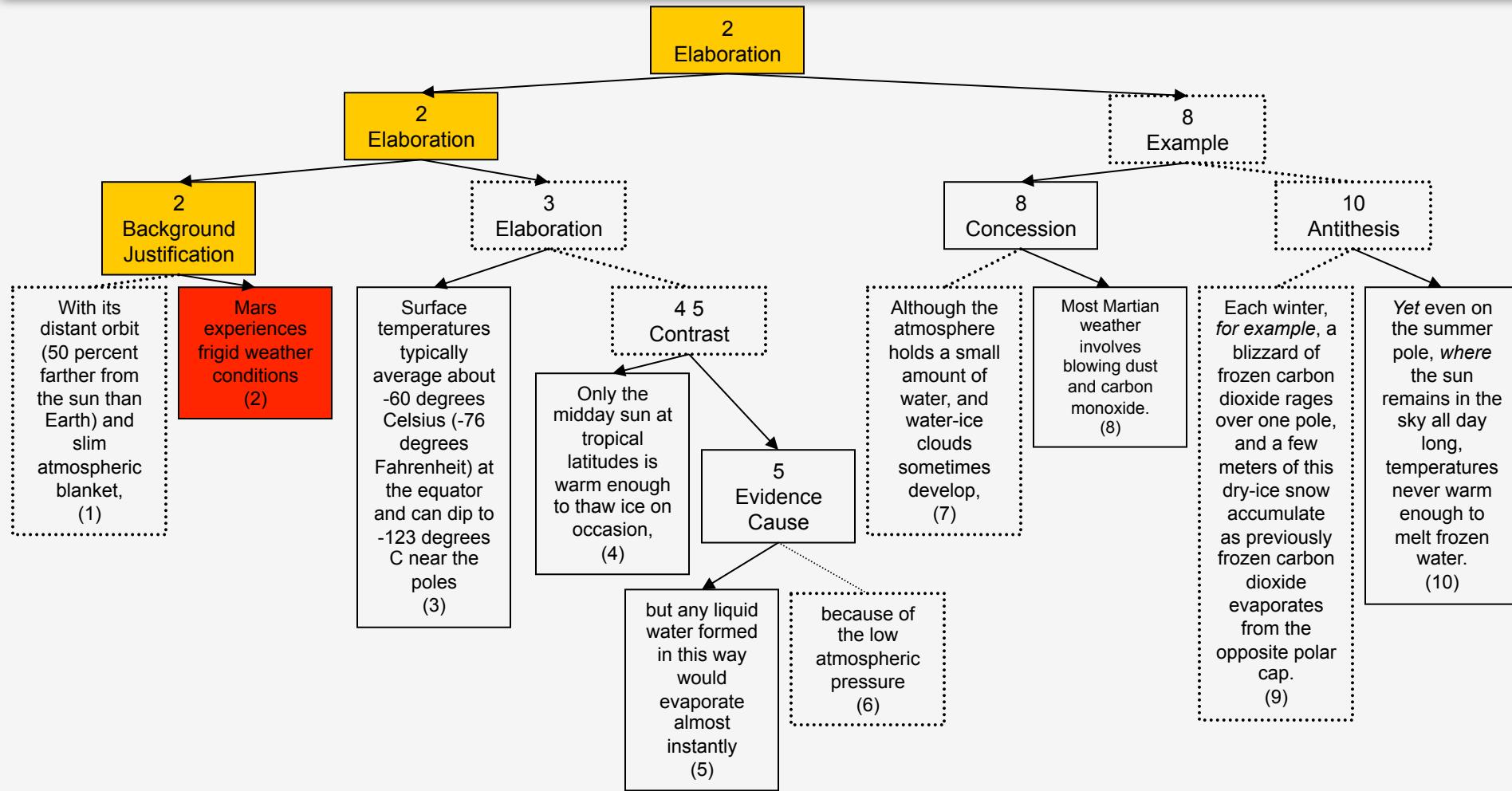
Marcu (1997-1999)

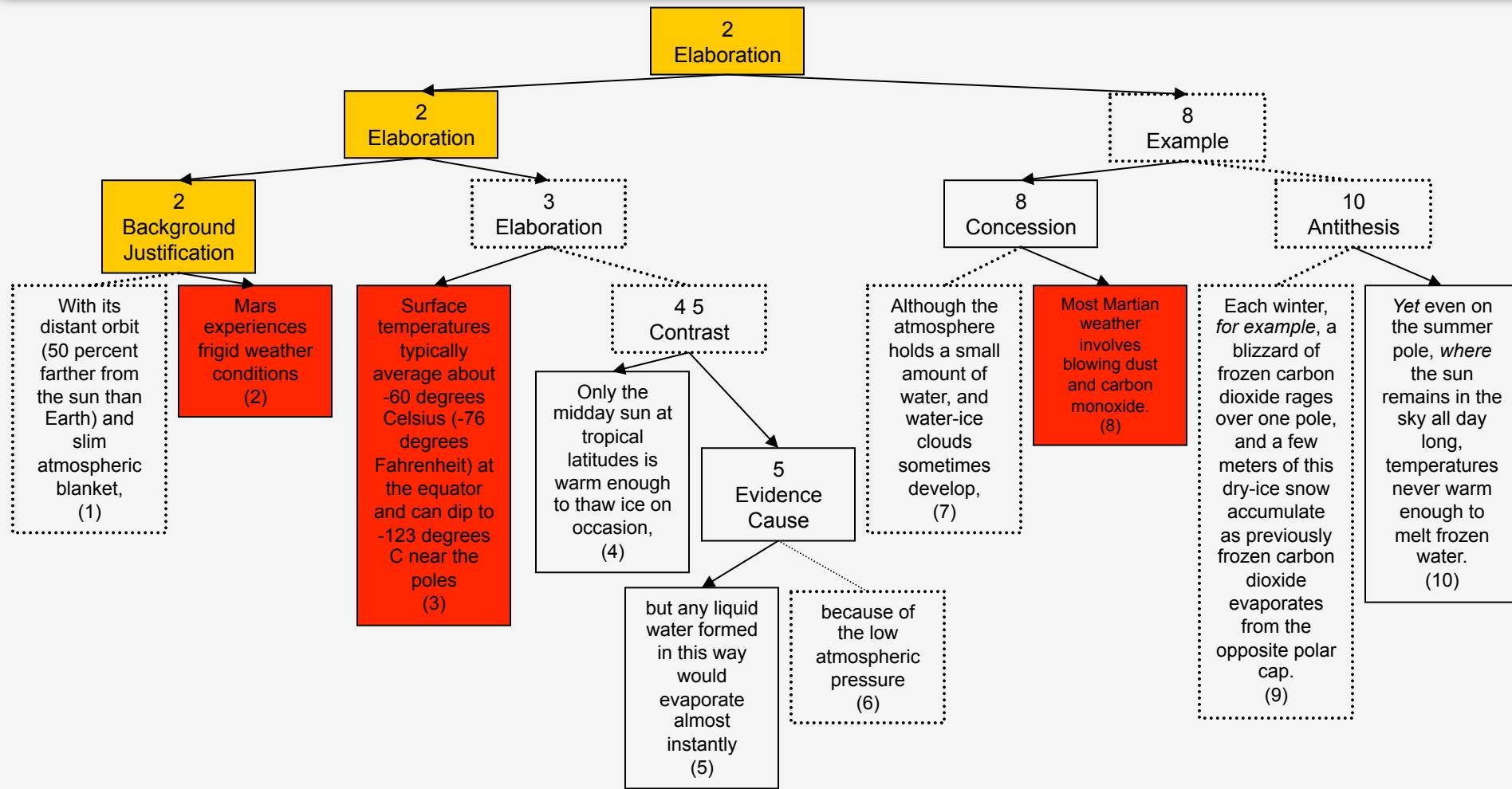
- Focuses on text coherence
- Based on Rhetorical Structure Theory (Mann and Thompson 1988)
 - Nucleus+Satellite [The truth is that the pressure to smoke in junior high is greater than it will be any other time of one's life:]_N [we know that 3,000 teens start smoking each day.]_S
- Example: evidence
 - N+S combination increases reader's belief in N

With its distant orbit (50 percent farther from the sun than Earth) and slim atmospheric blanket, (1)	Mars experiences frigid weather conditions (2)	Surface temperatures typically average about -60 degrees Celsius (-76 degrees Fahrenheit) at the equator and can dip to -123 degrees C near the poles (3)	Only the midday sun at tropical latitudes is warm enough to thaw ice on occasion, (4)	but any liquid water formed in this way would evaporate almost instantly (5)	because of the low atmospheric pressure (6)	Although the atmosphere holds a small amount of water, and water-ice clouds sometimes develop, (7)	Most Martian weather involves blowing dust and carbon monoxide. (8)	Each winter, <i>for example</i> , a blizzard of frozen carbon dioxide rages over one pole, and a few meters of this dry-ice snow accumulate as previously frozen carbon dioxide evaporates from the opposite polar cap. (9)	Yet even on the summer pole, where the sun remains in the sky all day long, temperatures never warm enough to melt frozen water. (10)
---	--	---	---	--	---	--	---	---	---



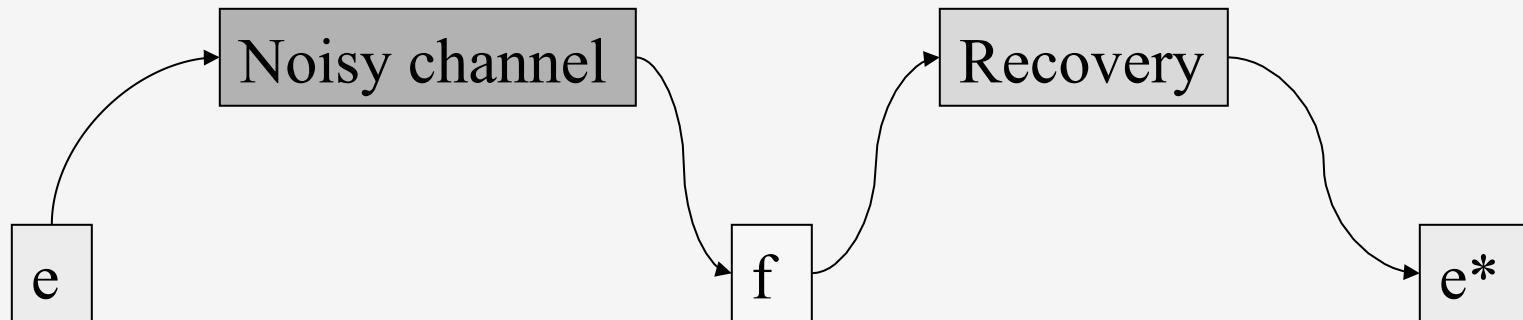






Noisy Channel Models

- Source/target language
- Coding process



Berger and Mittal 2000

- Source language
 - full document
- Target language
 - Summary
- Gisting (OCELOT)

$$g^* = \underset{g}{\operatorname{argmax}} p(g|d) = \underset{g}{\operatorname{argmax}} p(g) \cdot p(d|g)$$

Berger & Mittal 2000

- Training on 100K summary+document pairs
- Testing on 1046 pairs
- Use Viterbi-type search
- Evaluation: word overlap (0.2–0.4)
- No word ordering

Berger & Mittal 2000

Sample output:

Audubon society atlanta area savannah georgia chatham
and local birding savannah keepers chapter of the audubon
georgia and leasing

Carbonell and Goldstein (1998)

- Maximal marginal relevance
- Greedy selection method
- Query-based summaries
- Law of diminishing returns

C = doc collection

Q = user query

R = IR(C, Q, θ)

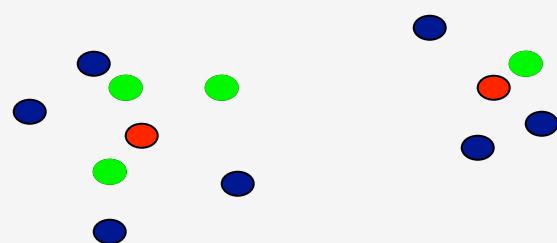
S = already retrieved documents

Sim = similarity metric used

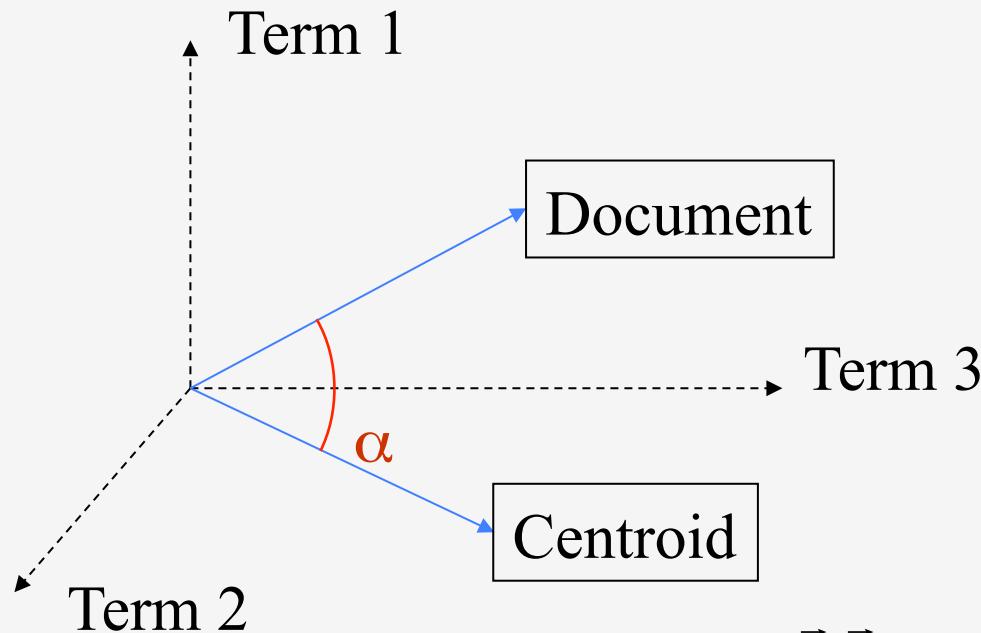
$$\text{MMR} = \operatorname{argmax}_{D_i \in R \setminus S} [\lambda \operatorname{Sim}_1(D_i, Q) - (1-\lambda) \max_{D_j \in S} \operatorname{Sim}_2(D_i, D_j)]$$

Mead (Radev et al. 2000)

- Salience-based extractive summarization
- In 6 languages
- Centroid-based method
- Single and multi document
- Vector space model
- Features: position, length, centroid
- Cross-document structure theory
- Reranker – similar to MMR
- Open source library (www.summarization.com/mead)



Mead



$$\cos(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\| \|\vec{y}\|} = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2} \sqrt{\sum_{i=1}^n y_i^2}}$$

Mead

- Input
 - Cluster of d documents with n sentences (compression rate = r)
- Output
 - $(n * r)$ sentences from the cluster with the highest scores

$$SCORE(s) = \sum_i (w_c C_i + w_p P_i + w_f F_i)$$

NewsInEssence (Radev et al. 2001)

- Web-based multi-document news summarization system

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

File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites History Print Mail E-mail Dell Home Real.com Messenger

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

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

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

Interactive Multi-source News Summarization

Home Current Clusters Create Cluster Save Cluster Track Cluster User Cluster Archive CDR Cluster Archive Google Cluster Archive

Help About NewsInEssence 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 - 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 Lowe's Key about 10 minutes earlier 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
- 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* 12 articles, 4 summaries: 07/02, 7:40 PM

NewsTroll from URL: url: news.troll.com/cgi-bin/troll.pl?MSNBC BBC or USA Today [http://www.msnbc.com/] Search

NewsTroll from query: fireworks bonita flames [Search]

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 (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)

Internet

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

File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites History Print Mail E-mail Dell Home Real.com Messenger

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

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

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

Interactive Multi-source News Summarization

Home Current Clusters Create Cluster Save Cluster Track Cluster User Cluster Archive CDR Cluster Archive Google Cluster Archive

Help About NewsInEssence Contact Us CLAIR MEAD summarization.com

Swede charged with plans to hijack plane

An organiser of an Islamic conference in Birmingham says it is "highly unlikely" that the man who got a seat on a British-bound flight in Sweden, was due to attend the conference. Security officers at Stockholm's Vasteras Airport say they found a handgun in a toiletries bag when they scanned the 29-year-old man's hand luggage.

Create A New Cluster

Recent User Clusters (more)

- Attacks expose Afghanistan vulnerability* 3 articles, 1 summary: 09/10, 4:41 PM

Recent CDR Clusters (more)

- Bush Continues to lobby foreign leaders against Iraq* 29 articles, 4 summaries: 09/10, 8:01 PM

Recent Google Headlines (more)

- Fatah says it will prevent attacks on Israeli civilians* 5 articles, 4 summaries: 09/11, 12:00 AM

FULL SUMMARY

User Clusters (Archive)

■ Disarm Iraq, Bush to urge U.N.* 11 articles, 5 summaries: 09/07, 2:09 AM

■ Cakewalk to Kabul* 10 articles, 2 summaries: 09/06, 7:41 PM

■ Confidence crisis in Canadian ice hockey* 3 articles: 09/09, 9:05 PM

■ Arafat: Palestinians on verge of statehood* 11 articles: 09/08, 2:39 PM

■ Karzai attacker from Taliban area*

Internet

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

File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites History Mail Print Edit Dell Home Real.com Messenger

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

User Clusters (Archive) (Create Your Own Cluster)

- **'Attacks expose Afghanistan vulnerability'**
3 articles, 1 summary: 09/10, 4:41 PM
- **'Microsoft hit by hacker attack'**
3 articles, 3 summaries: 09/09, 9:23 PM
- **'Confidence crisis in Canadian ice hockey'**
3 articles: 09/09, 9:05 PM
- **'Arafat: Palestinians on verge of statehood'**
7 articles, 1 summary: 09/09, 1:53 PM
- **'U.S. questions at Qaeda witness seeking reduced sentence'**
1 article, 1 summary: 09/10, 1:12 PM
- **'Knifeeman overpowered on Air Seychelles flight'**
1 article, 1 summary: 09/09, 1:05 PM
- **'9:11: American Idol seizes the day'**
1 article, 4 summaries: 09/07, 4:22 PM
- **'Iraq says war would be an American mistake'**
1 article, 2 summaries: 09/07, 3:01 AM
- **'US justifies Afghan wedding bombing'**
4 articles, 3 summaries: 09/07, 2:42 AM
- **'Powell: Cheney's slow 6 months'**
6 articles: 09/07, 2:39 AM

CIDR Clusters (more)

- **'Bush Continues to lobby foreign leaders against Iraq'**
29 articles, 4 summaries: 09/10, 8:01 PM
- **'CNN.com - Cheney: Saddam working on nuclear weapons - September 10, 2002'**
13 articles, 4 summaries: 09/10, 8:01 PM
- **'CNN.com - Ashcroft, Ridge: Americans should be**

Google Headlines (more)

- **'Fatah says it will prevent attacks on Israeli civilians'**
5 articles, 4 summaries: 09/11, 12:00 AM
- **'White House Puts Nation On "High Risk" Terror Alert'**
4 articles, 4 summaries: 09/11, 12:00 AM
- **'At least 50 dead after luxury train comes off bridge in India'**
4 articles, 4 summaries: 09/11, 12:00 AM

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

File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites History Mail Print Edit Dell Home Real.com Messenger

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

N.C.: 7 articles, 4 summaries: 09/10, 8:01 PM

'Swede charged with plans to hijack plane'
produced on 08/30, 1:52 PM

10% Summary

An organiser of an Islamic conference in Birmingham says it is "highly unlikely" that the man charged with trying to hijack a Ryanair flight in Sweden, was due to attend the conference. (4.2) Security officers at Stockholm's Västerås Airport say they found a handgun in a toiletries bag when they scanned the 29-year-old man's hand luggage. (4.3) The man - who was born in Sweden to Tunisian parents - had been trying to board a Ryanair flight to London's Stansted airport. (4.4) Swedish Police say he was with a party travelling to an Islamic conference in Birmingham and believe the man was going to hijack the plane. (4.5)

A man arrested in Sweden on suspicion of intending to hijack a plane bound for Britain has denied any knowledge of a gun in his luggage. (1.2) The man was arrested at Västerås airport, 100km (60 miles) north-west of Stockholm, after allegedly being caught trying to board the plane for Stansted with a gun. (1.3) "We believe he was going to hijack the plane," police spokesman Ulf Palm said. (1.6) UK Foreign Secretary Jack Straw, arriving in Denmark for the EU summit, said: "The fact that he appeared to be about to hijack the plane shows that the threat of international terrorism remains, and highlights the need for us to take concerted international action against that threat." (1.9)

Passengers aboard a plane that became the target of an alleged hijack attempt have praised the vigilance of the airport team who dealt with the incident. (3.2) A man was caught allegedly trying to board their Ryanair flight from Stockholm to Stansted carrying a gun. (3.3)

Swedish airport authorities have been praised for their security measures after a man was arrested on suspicion of attempted hijacking, when trying to board a plane to London carrying a handgun. (2.2) A 29-year-old man of Tunisian origin is being held by Swedish police after he allegedly attempted to board the plane at Västerås Airport 100 kilometres (60 miles) northwest of the capital, Stockholm, with a gun in his hand luggage. (5.2) Police are holding the suspect on a preliminary charge of planning to hijack a plane. (5.6)

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

File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites History Mail Print Edit Dell Home Real.com Messenger

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

Birmingham, England, according to police in Västerås, 60 miles northwest of the capital, Stockholm. (2.6) Security officers at Västerås airport found the weapon in a toiletries bag when they scanned the man's hand luggage on Thursday, police spokesman Ulf Palm said. (2.7)

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/224956.htm	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/225741.htm	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/225342.htm	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/uk/11/09/05/07225318.htm	news.bbc.co.uk	08/30, 8:54 AM
<input checked="" type="checkbox"/>	5	Terror suspect 'not attending conference' [Use As Seed] http://news.cnn.com/2002/WORLD/europe/08/20/stockholm.guide/index.html	www.cnn.com	08/30, 8:57 AM
<input checked="" type="checkbox"/>	6	Swede charged with plans to hijack plane [Use As Seed] http://news.msnbc.com/news/801304.asp?sp=1	www.msnbc.com	08/30, 12:00 AM
<input checked="" type="checkbox"/>	7	Swede faces attempt hijack charge [Use As Seed] http://news.msnbc.com/news/801287.asp	www.msnbc.com	08/29, 12:00 AM

Redraw **Reset** Compression: 10% **Summarize**

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

Email: Time: **Go**

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

File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites History Mail Print Edit Dell Home Real.com Messenger

Address: http://www.newsinessence.com/nie.cgi?CID=20020911013532IIsummary

Links: Best of the Web Channel Guide Customize Links Free HotMail Internet Explorer News Internet Start Microsoft RealPlayer

http://news.bbc.co.uk/1/hi/world/europe/1954240.htm news.bbc.co.uk 09/09, 7:50 PM

Swiss President calls for UN membership [Use As Seed] <http://news.bbc.co.uk/1/hi/world/source/172756.htm> news.bbc.co.uk 01/01, 7:00 PM

Redraw **Reset** Compression: 10% **Summarize**

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

Email: Time: **Go**

Create a New Cluster:
You can create a cluster from a seed URL or search query.
If you know the URL of the news story you want to use as a seed or the query, enter it below.
Otherwise, you can browse a list of recent headlines at [NIE Headlines](#).

Seed URL: **Search from Seed** [Advanced Options](#)

Enter the URL of a seed story that you want to grow into a cluster.
You can use [NIE Headlines](#) to get a seed URL (marked by NIE).
Search by Seed currently requires that seed articles be from one of [cnn.com](#), [news.yahoo.com](#), [msnbc.com](#), [news.bbc.co.uk](#), or [usatoday.com](#).

Query Words: **Search by Query** [Advanced Options](#)

Other News Summarization Systems

- Newsblaster (McKeown et al. 2002)
 - Google News

A screenshot of the Columbia Newsblaster website as it would appear in Microsoft Internet Explorer. The title bar reads "Columbia Newsblaster: Summarizing All the News on the Web - Microsoft Internet Explorer provided by AT&T WorldNet Service". The menu bar includes File, Edit, View, Favorites, Tools, Help, Back, Forward, Stop, Home, Search, Favorites, History, Links, Customize Links, Free Hotmail, Windows Media, Windows, Like Music - Try AOL!, WorldNet Customer Care, WorldNet Home Page, and WorldNet Member Services. The address bar shows "http://newsblaster.cs.columbia.edu/". The main content area features a large banner for "Columbia Newsblaster" with the subtitle "Summarizing all the news on the Web". To the right, a date box says "Monday, June 30, 2003" and "Last update: 7:28 AM EST". On the left, a sidebar lists news categories: U.S., World, Finance, SciTech, Entertainment, Sports, View Today's Images, View Archives, About Newsblaster, Newsblaster in Press, Academic Papers, and Current Sources (with links to Business.com, Suntimes.com, and Dallasnews.com). The central column has a large image of a person holding a surfboard next to an Israeli flag on a beach. Headlines include "Palestinian militants declare truce: Israel starts pulling troops out of Gaza" and "Once Hailed, Soldiers in Iraq Now Feel Blame at Each Step". The right column features a "Top News" section with the headline "Now that court says sodomy's OK, all limits are off".

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

File Edit View Favorites Tools Help

Back Forward Stop Refresh Favorites History

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](#)

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 pulled the city's historic Manger Square. This morning 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, a senior military official 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 »

Text Version

About Google News

World »

Kenyan Women Accuse British Troops of Rape
New York Times - 3 hours ago
Kenya's president says he will sue Britain over accusations that Kenyan

U.S. »

Low-Profile Suburb Braces for Onslaught
Washington Post - 40 minutes ago
FORT MYERS, Fla. - A day after a Florida County judge


Reuters


CNN


Internet

NLP

NLP

Introduction to NLP

Summarization Techniques 3/3

Conroy and O'Leary (2001)

- Using Hidden Markov Models
- Takes into account the local dependencies between sentences
- Features
 - Position, number of terms, similarity to document terms
- HMM alternates between summary and non-summary states



Figure 1: Summary Extraction Markov Model to Extract 2 Lead Sentences and Additional Supporting Sentences

Osborne (2002)

- Don't assume feature independence
- Use maxent (log-linear) models
- Better than Naïve Bayes
- Features
 - Sentence length
 - Sentence position
 - Inside introduction
 - Inside conclusion

Lexrank (Erkan and Radev 2004)

- Single and multi-document summarization
- Lexical Centrality
 - Represent text as graph
 - Graph centrality
 - Graph clustering
 - Random walks

Lexrank

- 1 (d1s1) Iraqi Vice President Taha Yassin Ramadan announced today, Sunday, that Iraq refuses to back down from its decision to stop cooperating with disarmament inspectors before its demands are met.
- 2 (d2s1) Iraqi Vice president Taha Yassin Ramadan announced today, Thursday, that Iraq rejects cooperating with the United Nations except on the issue of lifting the blockade imposed upon it since the year 1990.
- 3 (d2s2) Ramadan told reporters in Baghdad that "Iraq cannot deal positively with whoever represents the Security Council unless there was a clear stance on the issue of lifting the blockade off of it."
- 4 (d2s3) Baghdad had decided late last October to completely cease cooperating with the inspectors of the United Nations Special Commission (UNSCOM), in charge of disarming Iraq's weapons, and whose work became very limited since the fifth of August, and announced it will not resume its cooperation with the Commission even if it were subjected to a military operation.
- 5 (d3s1) The Russian Foreign Minister, Igor Ivanov, warned today, Wednesday against using force against Iraq, which will destroy, according to him, seven years of difficult diplomatic work and will complicate the regional situation in the area.
- 6 (d3s2) Ivanov contended that carrying out air strikes against Iraq, who refuses to cooperate with the United Nations inspectors, ``will end the tremendous work achieved by the international group during the past seven years and will complicate the situation in the region."
- 7 (d3s3) Nevertheless, Ivanov stressed that Baghdad must resume working with the Special Commission in charge of disarming the Iraqi weapons of mass destruction (UNSCOM).
- 8 (d4s1) The Special Representative of the United Nations Secretary-General in Baghdad, Prakash Shah, announced today, Wednesday, after meeting with the Iraqi Deputy Prime Minister Tariq Aziz, that Iraq refuses to back down from its decision to cut off cooperation with the disarmament inspectors.
- 9 (d5s1) British Prime Minister Tony Blair said today, Sunday, that the crisis between the international community and Iraq ``did not end" and that Britain is still ``ready, prepared, and able to strike Iraq."
- 10 (d5s2) In a gathering with the press held at the Prime Minister's office, Blair contended that the crisis with Iraq ``will not end until Iraq has absolutely and unconditionally respected its commitments" towards the United Nations.
- 11 (d5s3) A spokesman for Tony Blair had indicated that the British Prime Minister gave permission to British Air Force Tornado planes stationed in Kuwait to join the aerial bombardment against Iraq.

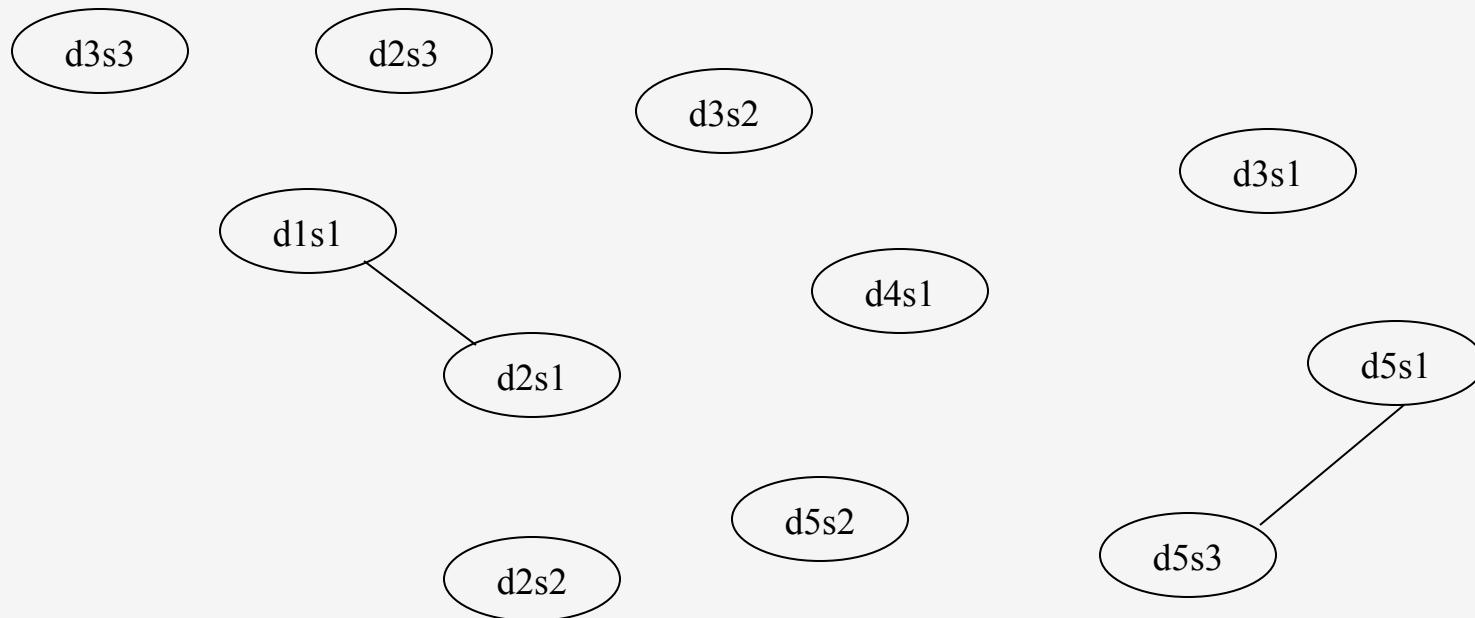
Lexrank

	1	2	3	4	5	6	7	8	9	10	11
1	1.00	0.45	0.02	0.17	0.03	0.22	0.03	0.28	0.06	0.06	0.00
2	0.45	1.00	0.16	0.27	0.03	0.19	0.03	0.21	0.03	0.15	0.00
3	0.02	0.16	1.00	0.03	0.00	0.01	0.03	0.04	0.00	0.01	0.00
4	0.17	0.27	0.03	1.00	0.01	0.16	0.28	0.17	0.00	0.09	0.01
5	0.03	0.03	0.00	0.01	1.00	0.29	0.05	0.15	0.20	0.04	0.18
6	0.22	0.19	0.01	0.16	0.29	1.00	0.05	0.29	0.04	0.20	0.03
7	0.03	0.03	0.03	0.28	0.05	0.05	1.00	0.06	0.00	0.00	0.01
8	0.28	0.21	0.04	0.17	0.15	0.29	0.06	1.00	0.25	0.20	0.17
9	0.06	0.03	0.00	0.00	0.20	0.04	0.00	0.25	1.00	0.26	0.38
10	0.06	0.15	0.01	0.09	0.04	0.20	0.00	0.20	0.26	1.00	0.12
11	0.00	0.00	0.00	0.01	0.18	0.03	0.01	0.17	0.38	0.12	1.00

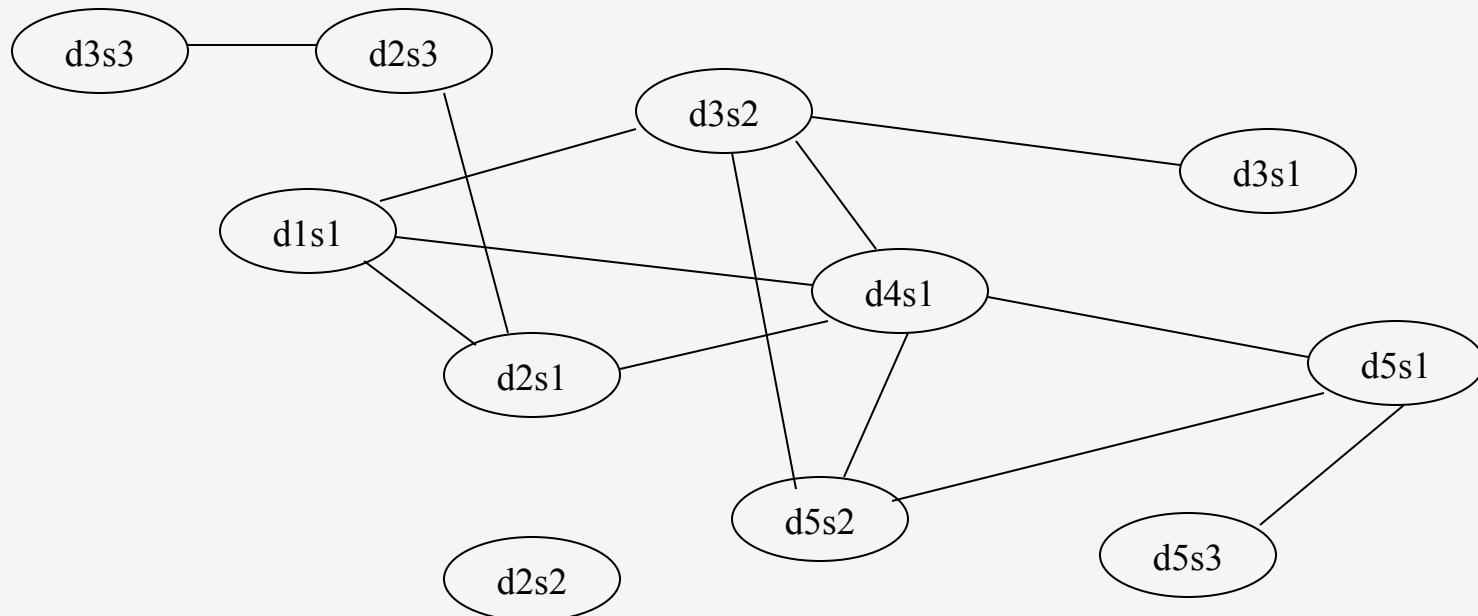
Lexrank

	1	2	3	4	5	6	7	8	9	10	11
1	1.00	0.45	0.02	0.17	0.03	0.22	0.03	0.28	0.06	0.06	0.00
2	0.45	1.00	0.16	0.27	0.03	0.19	0.03	0.21	0.03	0.15	0.00
3	0.02	0.16	1.00	0.03	0.00	0.01	0.03	0.04	0.00	0.01	0.00
4	0.17	0.27	0.03	1.00	0.01	0.16	0.28	0.17	0.00	0.09	0.01
5	0.03	0.03	0.00	0.01	1.00	0.29	0.05	0.15	0.20	0.04	0.18
6	0.22	0.19	0.01	0.16	0.29	1.00	0.05	0.29	0.04	0.20	0.03
7	0.03	0.03	0.03	0.28	0.05	0.05	1.00	0.06	0.00	0.00	0.01
8	0.28	0.21	0.04	0.17	0.15	0.29	0.06	1.00	0.25	0.20	0.17
9	0.06	0.03	0.00	0.00	0.20	0.04	0.00	0.25	1.00	0.26	0.38
10	0.06	0.15	0.01	0.09	0.04	0.20	0.00	0.20	0.26	1.00	0.12
11	0.00	0.00	0.00	0.01	0.18	0.03	0.01	0.17	0.38	0.12	1.00

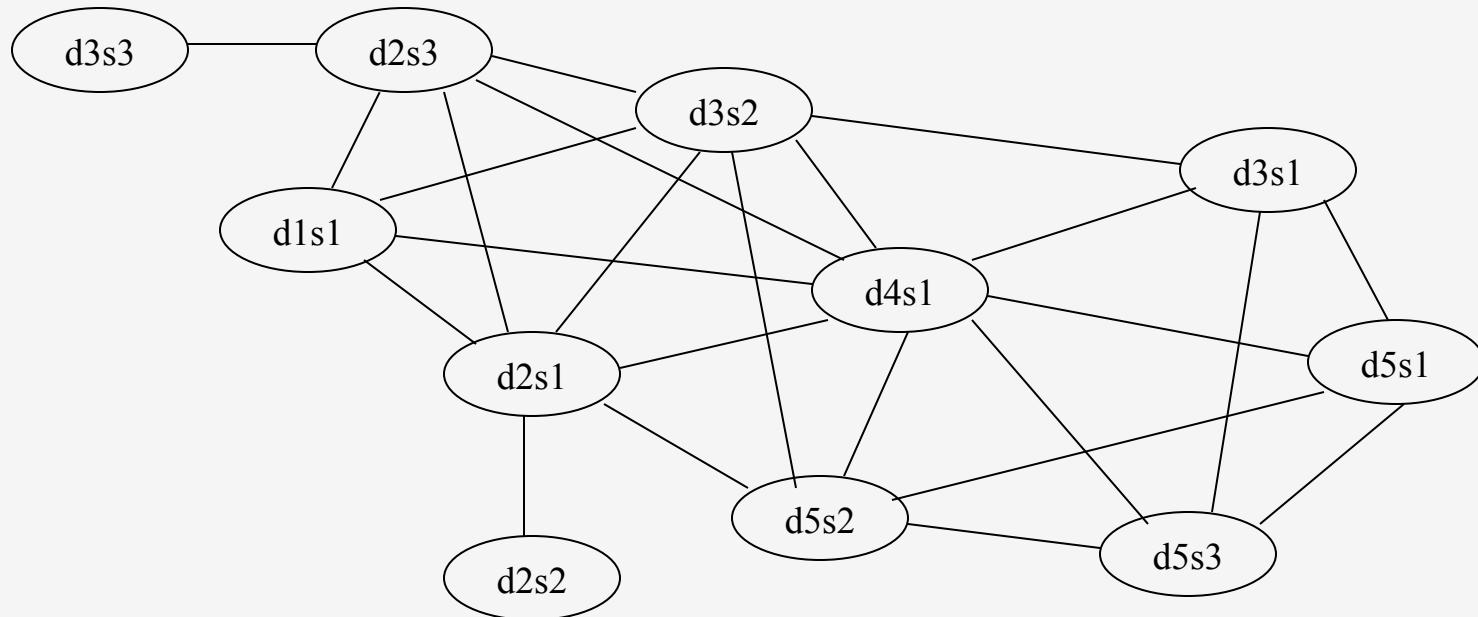
Cosine Centrality ($t=0.3$)



Cosine Centrality ($t=0.2$)

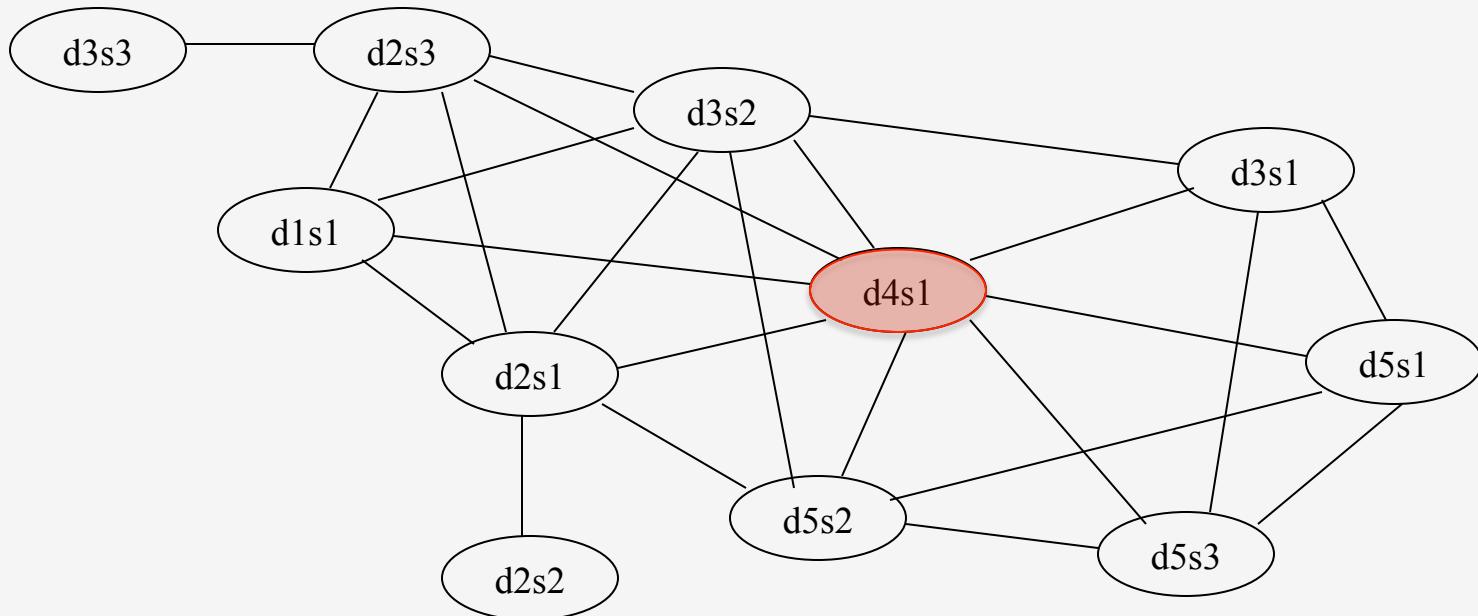


Cosine Centrality ($t=0.1$)



Sentences vote for the most central sentence!

Cosine Centrality ($t=0.1$)



Sentences vote for the most central sentence!

Lexrank (Advanced Material)

- Square connectivity matrix
- Directed vs. undirected
- An eigenvalue for a square matrix A is a scalar λ such that there exists a vector $x \neq 0$ such that $Ax = \lambda x$
- The normalized eigenvector associated with the largest λ is called the principal eigenvector of A
- A matrix is called a stochastic matrix when the sum of entries in each row sum to 1 and none is negative. All stochastic matrices have a principal eigenvector

Lexrank (Advanced Material)

- The connectivity matrix used in PageRank [Page & al. 1998] is irreducible [Langville & Meyer 2003]
- An iterative method (power method) can be used to compute the principal eigenvector
- That eigenvector corresponds to the stationary value of the Markov stochastic process described by the connectivity matrix
- This is also equivalent to performing a random walk on the matrix

Lexrank (Advanced Material)

- The stationary value of the Markov stochastic matrix can be computed using an iterative power method:

$$p = E^T p$$

$$(I - E^T)p = 0$$

- PageRank adds an extra twist to deal with dead-end pages. With a probability $1-\varepsilon$, a random starting point is chosen. This has a natural interpretation in the case of Web page ranking

$$p(v) = \frac{1 - \varepsilon}{n} + \varepsilon \sum_{u \in pr[v]} \frac{p(u)}{|su[u]|}$$

su = successor nodes
pr = predecessor nodes

- Eigenvector centrality: the paths in the random walk are weighted by the centrality of the nodes that the path connects

Gong and Liu (2001)

- Using Latent Semantic Analysis (LSA)
- Single and multi-document
- Not using WordNet
- Each document is represented as a word by sentence matrix (row=word, column=sentence)
- TF*IDF weights in the matrix
- SVD: $A = USV^T$
- The rows of V^T are independent topics
- Select sentences that cover these independent topics

NLP

NLP

Introduction to NLP

Summarization Evaluation

Evaluation Criteria

- Appropriate length
- Fidelity
- Salience
- Grammaticality
- Non-redundancy
- Referential well-formedness
- Structure and coherence

Ideal Evaluation

Information content

$$C = \text{Compression Ratio} = \frac{|S|}{|D|}$$

$$R = \text{Retention Ratio} = \frac{i(S)}{i(D)}$$

R>C

Types Of Evaluation Methods

- Extrinsic techniques (task-based)
 - Can you make the same decision using the summary as with the full text, but in less time
- Intrinsic techniques
 - Comparing summaries against gold standards

Precision and Recall

	Relevant	Non-relevant
System: relevant	A	B
System: non-relevant	C	D

$$\text{Precision: } P = \frac{A}{A + B}$$

$$\text{Recall: } R = \frac{A}{A + C}$$

$$F = \frac{2PR}{(P + R)}$$

Rouge (Lin and Hovy 2003)

- Very convenient for prototyping
- Based on Papineni et al's BLEU (used for MT)
- R stands for Recall
- ROUGE-n is a measure of n-gram overlap between a summary and a set of reference summaries
- ROUGE-L uses longest common subsequence instead of n-gram overlap
- Correlated with manual evaluations when averaged over many examples
- It can be easily gamed

Relative Utility (Radev et al. 2000)

	Ideal	System 1	System 2
S1	+	+	-
S2	+	+	+
S3	-	-	-
S4	-	-	+
S5	-	-	-
S6	-	-	-
S7	-	-	-
S8	-	-	-
S9	-	-	-
S10	-	-	-

	Ideal	System 1	System 2
S1	10(+)	10(+)	5
S2	8(+)	9(+)	8(+)
S3	2	3	4
S4	7	6	9(+)

$U(system, ideal) = \% \text{ of ideal utility covered by system summary}$

Relative Utility

$$\text{RU} = \frac{17}{\text{---}}$$

	Judge1	Judge2	Judge3
Sentence 1	10	10	5
Sentence 2	8	9	8
Sentence 3	2	3	4
Sentence 4	5	6	9

Relative Utility

$$\text{RU} = \frac{13}{17} = 0.765$$

	Judge 1	Judge 2	Judge 3
Sentence 1	10	10	5
Sentence 2	8	9	8
Sentence 3	2	3	4
Sentence 4	5	6	9

Normalized System Performance

	Judge 1	Judge 2	Judge 3	Average
Judge 1	1.000	1.000	0.765	0.883
Judge 2	1.000	1.000	0.765	0.883
Judge 3	0.722	0.789	1.000	0.756

$$D = \frac{(S-R)}{(J-R)}$$

Normalized system performance

System performance

Random performance

Interjudge agreement

Random Performance

average of all $\frac{n!}{(n(1-r))! (r*n)!}$ systems

$$D = \frac{(S-R)}{(J-R)}$$

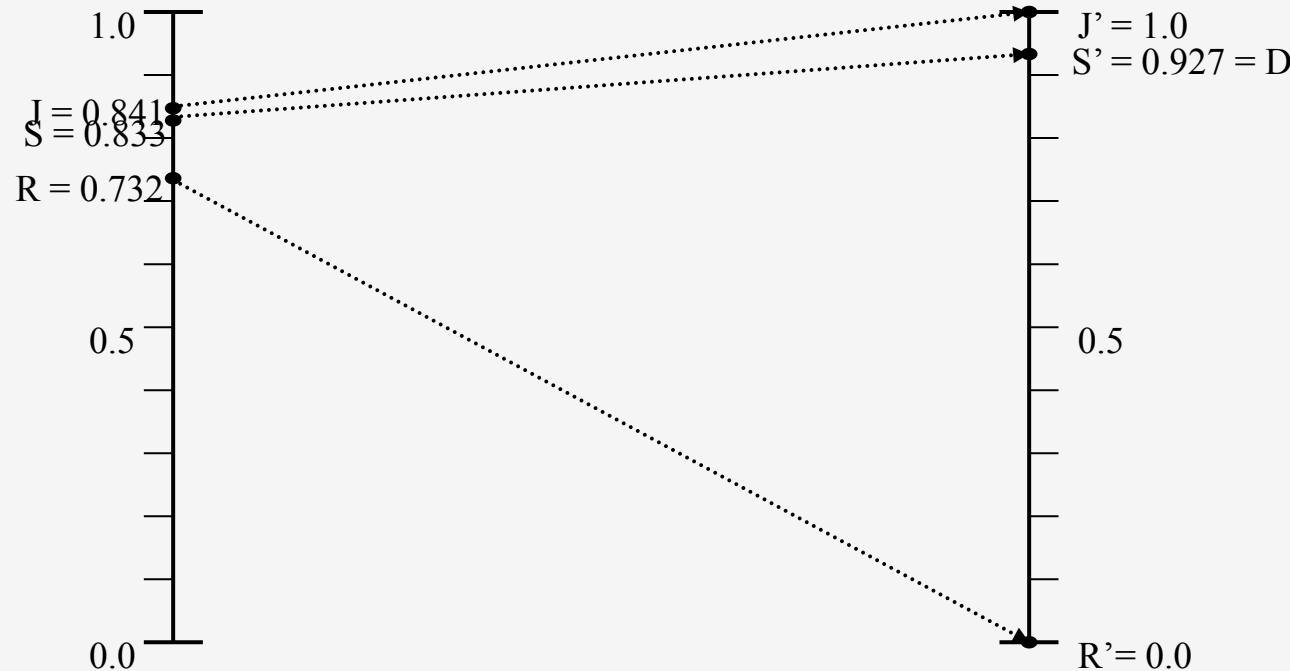

{12}
{13}
{14}
{23}
{24}
{34}

Example

$$D \{14\} = \frac{(S-R)}{(J-R)} = \frac{0.833 - 0.732}{0.841 - 0.732} = 0.927$$

$$D \{24\} = 0.963$$

Normalized Evaluation of {14}



ARTICLE 18853: ALGIERS, May 20 (AFP)

1. Eighteen decapitated bodies have been found in a mass grave in northern Algeria, press reports said Thursday, adding that two shepherds were murdered earlier this week.
2. Security forces found the mass grave on Wednesday at Chbika, near Djelfa, 275 kilometers (170 miles) south of the capital.
3. It contained the bodies of people killed last year during a wedding ceremony, according to Le Quotidien Liberte.
4. The victims included women, children and old men.
5. Most of them had been decapitated and their heads thrown on a road, reported the Es Sahafa.
6. Another mass grave containing the bodies of around 10 people was discovered recently near Algiers, in the Eucalyptus district.
7. The two shepherds were killed Monday evening by a group of nine armed Islamists near the Moulay Slissen forest.
8. After being injured in a hail of automatic weapons fire, the pair were finished off with machete blows before being decapitated, Le Quotidien d'Oran reported.
9. Seven people, six of them children, were killed and two injured Wednesday by armed Islamists near Medea, 120 kilometers (75 miles) south of Algiers, security forces said.
10. The same day a parcel bomb explosion injured 17 people in Algiers itself.
11. Since early March, violence linked to armed Islamists has claimed more than 500 lives, according to press tallies.

ARTICLE 18854: ALGIERS, May 20 (UPI)

1. Algerian newspapers have reported that 18 decapitated bodies have been found by authorities in the south of the country.
2. Police found the ``decapitated bodies of women, children and old men, with their heads thrown on a road'' near the town of Djelfa, 275 kilometers (170 miles) south of the capital Algiers.
3. In another incident on Wednesday, seven people -- including six children -- were killed by terrorists, Algerian security forces said.
4. Extremist Muslim militants were responsible for the slaughter of the seven people in the province of Medea, 120 kilometers (74 miles) south of Algiers.
5. The killers also kidnapped three girls during the same attack, authorities said, and one of the girls was found wounded on a nearby road.
6. Meanwhile, the Algerian daily Le Matin today quoted Interior Minister Abdul Malik Silal as saying that ``terrorism has not been eradicated, but the movement of the terrorists has significantly declined.''
7. Algerian violence has claimed the lives of more than 70,000 people since the army cancelled the 1992 general elections that Islamic parties were likely to win.
8. Mainstream Islamic groups, most of which are banned in the country, insist their members are not responsible for the violence against civilians.
9. Some Muslim groups have blamed the army, while others accuse ``foreign elements conspiring against Algeria.''

ARTICLE 18853: ALGIERS, May 20 (AFP)

1. Eighteen decapitated bodies have been found in a mass grave in northern Algeria, press reports said Thursday, adding that two shepherds were murdered earlier this week.
2. Security forces found the mass grave on Wednesday at Chbika, near Djelfa, 275 kilometers (170 miles) south of the capital.
3. It contained the bodies of people killed last year during a wedding ceremony, according to Le Quotidien Liberte.
4. The victims included women, children and old men.
5. Most of them had been decapitated and their heads thrown on a road, reported the Es Sahafa.
6. Another mass grave containing the bodies of around 10 people was discovered recently near Algiers, in the Eucalyptus district.
7. The two shepherds were killed Monday evening by a group of nine armed Islamists near the Moulay Slissen forest.
8. After being injured in a hail of automatic weapons fire, the pair were finished off with machete blows before being decapitated, Le Quotidien d'Oran reported.
9. Seven people, six of them children, were killed and two injured Wednesday by armed Islamists near Medea, 120 kilometers (75 miles) south of Algiers, security forces said.
10. The same day a parcel bomb explosion injured 17 people in Algiers itself.
11. Since early March, violence linked to armed Islamists has claimed more than 500 lives, according to press tallies.

ARTICLE 18854: ALGIERS, May 20 (UPI)

1. Algerian newspapers have reported that 18 decapitated bodies have been found by authorities in the south of the country.
2. Police found the ``decapitated bodies of women, children and old men, with their heads thrown on a road'' near the town of Jelfa, 275 kilometers (170 miles) south of the capital Algiers.
3. In another incident on Wednesday, seven people -- including six children -- were killed by terrorists, Algerian security forces said.
4. Extremist Muslim militants were responsible for the slaughter of the seven people in the province of Medea, 120 kilometers (74 miles) south of Algiers.
5. The killers also kidnapped three girls during the same attack, authorities said, and one of the girls was found wounded on a nearby road.
6. Meanwhile, the Algerian daily Le Matin today quoted Interior Minister Abdul Malik Silal as saying that ``terrorism has not been eradicated, but the movement of the terrorists has significantly declined.''
7. Algerian violence has claimed the lives of more than 70,000 people since the army cancelled the 1992 general elections that Islamic parties were likely to win.
8. Mainstream Islamic groups, most of which are banned in the country, insist their members are not responsible for the violence against civilians.
9. Some Muslim groups have blamed the army, while others accuse ``foreign elements conspiring against Algeria.'

Subsumption and Equivalence

- *Subsumption*: If the information content of sentence a (denoted as $I(a)$) is contained within sentence b, then a becomes informationally redundant and the content of b is said to subsume that of a:
 $I(a) \subset I(b)$
- *Equivalence*: If $I(a) \subset I(b) \wedge I(b) \subset I(a)$

Subsumption

- (1) John Doe was found guilty of the murder.
- (2) The court found John Doe guilty of the murder of Jane Doe last August and sentenced him to life.

Subsumption

	Article 1	Article 2	Article 3
S1	10	→ 10	5
S2	8	9	8
S3	2	3	4
S4	7	6	9

Subsumption (Cont'd)

$$SCORE(s) = \sum_i (w_c C_i + w_p P_i + w_f F_i) - w_R R_s$$

R_s = cross-sentence word overlap

$$R_s = \frac{2 * (\# \text{ overlapping words})}{(\# \text{ words in sentence 1} + \# \text{ words in sentence 2})}$$

$$w_R = \text{Max}_s (SCORE(s))$$

Pyramid

- Nenkova and Passonneau 2004
- Based on Semantic Content Units (SCU)
- Used for multi-document summarization
- Different surface realizations with equivalent meanings

Pyramid

A1 In 1998 two Libyans indicted in 1991 for the Lockerbie bombing were still in Libya.

B1 Two Libyans were indicted in 1991 for blowing up a Pan Am jumbo jet over Lockerbie, Scotland in 1988.

C1 Two Libyans, accused by the United States and Britain of bombing a New York bound Pan Am jet over Lockerbie, Scotland in 1988, killing 270 people, for 10 years were harbored by Libya who claimed the suspects could not get a fair trial in America or Britain.

D2 Two Libyan suspects were indicted in 1991.

SCU1 (w=4): two Libyans were officially accused of the Lockerbie bombing

A1 [two Libyans]1 [indicted]1

B1 [Two Libyans were indicted]1

C1 [Two Libyans,]1 [accused]1

D2 [Two Libyan suspects were indicted]1

SCU2 (w=3): the indictment of the two Lockerbie suspects was in 1991

A1 [in 1991]2

B1 [in 1991]2

D2 [in 1991.]2

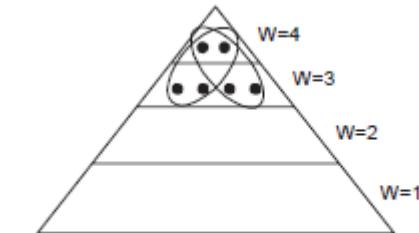


Figure 2: Two of six optimal summaries with 4 SCUs

Available Corpora

- DUC and TAC corpora
 - <http://duc.nist.gov>
 - <http://tac.nist.gov>
- SummBank corpus
 - <http://www.summarization.com/summbank>
- SUMMAC corpus
- NY Times corpus (from LDC)

NLP

NLP

Introduction to NLP

Sentence Simplification

Sentence Simplification

- Removing some parts of sentences
 - Quotes
 - Appositions
 - Adjectives and adverbs
 - Embedded clauses
 - Attribution clauses
- Applications
 - Subtitling
 - Headline generation
 - Mobile devices
 - Applications for the visually impaired

Knight and Marcu 2000

- Use structured (syntactic) information
- Two approaches
 - Noisy channel
 - Decision based

$$P_{\text{exp}}(\text{NP} \rightarrow \text{DT NN} \mid \text{NP} \rightarrow \text{DT NN}) = 0.8678$$

$$P_{\text{exp}}(\text{NP} \rightarrow \text{DT JJ NN} \mid \text{NP} \rightarrow \text{DT NN}) = 0.0287$$

$$P_{\text{exp}}(\text{NP} \rightarrow \text{DT NNP NN} \mid \text{NP} \rightarrow \text{DT NN}) = 0.0230$$

$$P_{\text{exp}}(\text{NP} \rightarrow \text{DT JJS NN} \mid \text{NP} \rightarrow \text{DT NN}) = 0.0115$$

$$P_{\text{exp}}(\text{NP} \rightarrow \text{DT NNP CD NN} \mid \text{NP} \rightarrow \text{DT NN}) = 0.0057$$

Knight and Marcu

- The documentation is typical of Epson quality; excellent.
 - Documentation is excellent.
- All of our design goals were achieved and the delivered performance matches the speed of the underlying device.
 - All design goals were achieved.
- Although the modules themselves may be physically and/or electronically incompatible, the cable-specific jacks on them provide industry-standard connections.
 - Cable-specific jacks provide industry-standard connections.
- Beyond the basic level, the operations of the three products vary widely.
 - The operations of the three products vary widely.
- Arborscan is reliable and worked accurately in testing, but it produces very large dxf files.
 - Arborscan produces very large dxf files.
- Many debugging features, including user-defined break points and variable-watching and message-watching windows, have been added.
 - Many debugging features have been added.

English Wikipedia

- **Anthony Charles Lynton Blair** (born 6 May 1953)^[1] is a former British Labour Party politician who served as the Prime Minister of the United Kingdom from 2 May 1997 to 27 June 2007. He was the Member of Parliament (MP) for Sedgefield from 1983 to 2007 and Leader of the Labour Party from 1994 to 2007. He resigned from all of these positions in June 2007.
- Tony Blair was elected Leader of the Labour Party in the leadership election of July 1994, following the sudden death of his predecessor, John Smith. Under his leadership, the party adopted the term "New Labour"^[2] and moved away from its traditional left wing position towards the centre ground.^{[3][4]} Blair subsequently led Labour to a landslide victory in the 1997 general election. At 43 years old, he became the youngest Prime Minister since Lord Liverpool in 1812. In the first years of the New Labour government, Blair's government implemented a number of 1997 manifesto pledges, introducing the minimum wage, Human Rights Act and Freedom of Information Act, and carrying out regional devolution, establishing the Scottish Parliament, the National Assembly for Wales, and the Northern Ireland Assembly.
- Blair's role as Prime Minister was particularly visible in foreign and security policy, including in Northern Ireland, where he was involved in the 1998 Good Friday Agreement. From the start of the War on Terror in 2001, Blair strongly supported the foreign policy of US President George W. Bush, notably by participating in the 2001 invasion of Afghanistan and 2003 invasion of Iraq. Blair is the Labour Party's longest-serving Prime Minister, the only person to have led the Labour Party to three consecutive general election victories, and the only Labour Prime Minister to serve consecutive terms more than one of which was at least four years long.
- He was succeeded as Leader of the Labour Party on 24 June 2007 and as Prime Minister on 27 June 2007 by Gordon Brown.^[5] On the day he resigned as Prime Minister, he was appointed the official Envoy of the Quartet on the Middle East. In May 2008, Blair launched his Tony Blair Faith Foundation.^[6] This was followed in July 2009 by the launching of the Faith and Globalisation Initiative with Yale University in the USA, Durham University in the UK and the National University of Singapore in Asia to deliver a postgraduate programme in partnership with the Foundation.^{[7][8]}

Simple English Wikipedia

- **Anthony Charles Lyton Blair**, usually called **Tony Blair**, is a former Prime Minister of the United Kingdom. He was born in Edinburgh. He was Prime Minister from May 1997 until June 2007 and was succeeded by the former Chancellor of the Exchequer, Gordon Brown.
- As leader of the Labour party, he won three general elections in the UK, in 1997, 2001 and 2005. He is married to Cherie Booth. They met on the top deck of a double-decker bus in 1975. Cherie claimed "It was a double-decker and we went upstairs. It was completely empty and by the time we got off we knew each other better than when we'd got on. And even better the next morning. He was a very good-looking young man, tall and slim, yet broad in the shoulders. A really strong body."^[1]. Cherie Blair is a lawyer, who graduated from the London School of Economics with a first class honours degree. Blair himself left Oxford University with a second class degree. They have four children, Euan, Nicky, Kathryn, and Leo. There was a controversy over Blair sending his eldest son Euan to a grant-maintained school. As a result of this, Alastair Campbell discovered Blair "standing stark naked reading the Daily Mail"^[2].
- He attributes his success in politics to a pair of lucky brogues which he wore for every single Prime Ministers Questions of his leadership. He claimed that "cheap shoes are a false economy".^[3]

NLP

NLP

Introduction to NLP

Collocations

Collocations (phrases)

- Dictionary definitions
 - Meaning of words in isolation
- “Know a word by the company that it keeps”
 - Firth 1935
- Examples
 - dead end
 - strong tea
 - Benazir Bhutto
 - Fabry disease

Collocations

- Properties
 - Common use
 - No general syntactic or semantic rules
 - Important for non-native speakers
- Collocation acquisition
 - Important for NLP

Types Of Multiword Sequences

- Idioms
- Free-word combinations
- Collocations

Examples

Idioms

To kick the bucket
Dead end
To catch up

Collocations

To trade actively
Table of contents
Orthogonal projection

Free-word combinations

To take the bus
The end of the road
To buy a house

Properties

- Arbitrariness: substitutions are usually not allowed:
 - Make an effort vs. *make an exertion
 - Running commentary vs. *running discussion
 - Commit treason vs. *commit treachery
- Language- and dialect-specific
 - Régler la circulation = direct traffic
 - Russian, German, Serbo-Croatian: direct translation of regulate is used
 - AE: set the table, make a decision
 - BE: lay the table, take a decision
 - “semcer le désarroi” – “to sow disarray” – “to wreak havoc”
- Common in technical language
- Recurrent in context

Uses

- Disambiguation (e.g, “bank”/“loan”, “river”)
- Translation
- Generation

Types of Collocations

- Grammatical
 - come to, put on; afraid that, fond of, by accident, witness to
- Semantic
 - only certain synonyms
- Flexible
 - find/discover/notice by chance

Base-Collocator Pairs

- Base – bears most of the meaning of the collocation. Writers think of the base first. Foreign language speakers search by base. For decoding purposes, it is more appropriate to store the collocation under the collocator.

Base	Collocator	Example
Noun	verb	Set the table
Noun	adjective	Warm greetings
Verb	adverb	Struggle desperately
Adjective	adverb	Sound asleep
Verb	preposition	Put on

Extracting Collocations

- Most-common bigrams?
- Drop function words?
- Look at POS sequences?

Extracting Collocations

- Mutual information

$$I(x;y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

- Larger means stronger
- What if $I(x;y) = 0$?
- What if $I(x;y) < 0$?

Yule's Coefficient

- A – frequency of pairs involving both W and X
- B – frequency of pairs involving W only
- C – frequency of pairs involving X only
- D – frequency of pairs involving neither

$$Y = \frac{AD - BC}{AD + BC}$$

$$-1 \leq Y \leq 1$$

Example

	W	w			
X	A=800	C=180		A	800
x	B=160	D=80		B	160
				C	180
				D	80
				AD-BC	35200
				AD+BC	92800
					0.38

Example From The Hansard Corpus (Brown, Lai, And Mercer) – “Prime”

French word	Mutual information
sein	5.63
bureau	5.63
trudeau	5.34
premier	5.25
résidence	5.12
intention	4.57
no	4.53
session	4.34

Flexible And Rigid Collocations

- Example (from Smadja): “free” and “trade”

Total	p-5	p-4	p-3	p-2	p-1	p+1	p+2	p+3	p+4	p+5
8031	7	6	13	5	7918	0	12	20	26	24

Xtract (Smadja)

- The Dow Jones Industrial Average
- The NYSE's composite index of all its listed common stocks fell *NUMBER* to *NUMBER*

Translating Collocations

- Brush up a lesson, repasser une leçon
- Bring about/осуществлять
- Hansards:
 - late spring
 - fin du printemps
 - Atlantic Canada Opportunities Agency
 - Agence de promotion économique du Canada atlantique

Links

- Sample phrasal collocations
 - [http://en.wiktionary.org/wiki/
Appendix:Collocations_of_do,_have,_make,_and_take](http://en.wiktionary.org/wiki/Appendix:Collocations_of_do,_have,_make,_and_take)
- List of English language idioms
 - [http://en.wikipedia.org/wiki/List_of_English-
language_idioms](http://en.wikipedia.org/wiki/List_of_English-language_idioms)
- Idiomsite
 - <http://www.idiomsite.com/>

NLP

NLP

Introduction to NLP

Information Retrieval

Introduction

- People search the Web daily
- Search engines
 - Google
 - Bing
 - Baidu
 - Yandex
- Information Retrieval is about search engines

Yahoo Search

ebola -- Yahoo Search Res... +

https://search.yahoo.com/yhs/search?p=ebola&ei=UTF-8&hsprt mozilla&hsmp=yhs-001

YAHOO! ebola

Search

Dragomir Corp Mail

Web Images Video Local Maps

Anytime ▾

Ebola Virus News

Liberia sees surge in new Ebola cases in border county
Reuters via Yahoo! News 2 hours ago

MONROVIA (Reuters) - An outbreak of Ebola cases in a western Liberia county threatens the country's goal of recording no new cases of the disease by the end of the year. From Dec. 1 to 25, some 49...

Malaria killing thousands more than Ebola in West Africa
Associated Press via Yahoo! News 19 hours ago

The forgotten epidemic? Malaria kills thousands more than Ebola
CBS News 2 hours ago

Ebola virus disease - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Ebola_virus_disease Cached

Ebola virus disease (EVD; also Ebola hemorrhagic fever, or EHF), or simply Ebola, is a disease of humans and other primates caused by ebolaviruses.

#5 of 10 Most Popular Galleries of 2014: Ebola Outbreak In West Africa - Yahoo

news.yahoo.com

A man is carried away to be tested for Ebola after collapsing on a street in Monrovia December 9, 2014. The death toll from the Ebola outbreak in West Africa has reached 10,000.

Ads

Ebola Virus Pictures
Lifescript.com/Health

Find Facts, Symptoms & Treatments.
Trusted By 50 Million Visitors.

The Ebola Outbreak
news.yahoo.com/ebola

Stay informed with Katie Couric Yahoo News reveals the real world.

Ebola Virus Treatment
bit.ly/EbolaTreatment

Find out How to Prevent and Treatment for Ebola Virus, Read it Here!

About The Disease Ebola
buyerpricer.com/aboutdiseas Ebola

Searching for About The Disease Ebola?
Find Info & Browse Results Now.

no bucket
Y! Confidential

[hide]

Amazon Search

The screenshot shows a Firefox browser window connecting to www.amazon.com. The search term "samsung galaxy" is entered in the search bar. The results page displays 1-16 of 9,697,207 results for "samsung galaxy". On the left, there are navigation links for departments like Cell Phones & Accessories and Computers & Accessories. A sidebar on the left allows refining by brand, with checkboxes for Samsung, Hongyada, Ancerson, shopping_shop2000, ACEFAST INC, EVTECH, Gravydeals, and Inteuro. The main content area features a "Year-End Deals" banner and a section for "Amazon's Samsung Store" with images of various Samsung products. Below this, a product listing for the "Samsung Galaxy Tab 4 (7-Inch, White)" is shown, including its price (\$499.99), Prime status, delivery date (Tuesday, Dec 30), and a 4-star rating from 1,120 reviews. Another product listing for the "Samsung Galaxy S Duos II S7582 White DUAL SIM Factory Unlocked International Ver" is partially visible at the bottom.

Connecting... | x

www.amazon.com/?ie=UTF8&field-keywords=samsung%20galaxy&index=blended&link_code=qs&sourceid=Mc

samsung galaxy

Firefox has prevented the outdated plugin "Adobe Flash" from running on www.amazon.com.

amazon Try Prime Your Amazon.com Today's Deals Gift Cards Sell Help

Shop by Department Search All samsung galaxy Go

Hello, Sign in Your Account Try Prime Cart Wish List

Year-End Deals Continue Blocking Allow...

1-16 of 9,697,207 results for "samsung galaxy"

Show results for

Cell Phones & Accessories >

- Unlocked Cell Phones
- Cell Phones
- + See more

Computers & Accessories >

- Computer Tablets
- + See All 32 Departments

Related Searches: [samsung galaxy s3](#), [samsung galaxy s4](#), [samsung galaxy s5](#).

Amazon's Samsung Store

Samsung Unlocked Cell Phones Cell Phones & Accessories Computer Tablets Computers & Accessories

Samsung Galaxy Tab 4 (7-Inch, White) May 1, 2014

by Samsung

\$499.99 Click to see price ✓Prime

Get it by **Tuesday, Dec 30**

More Buying Choices \$131.49 used & new (40 offers)

★★★★★ 1,120

FREE Shipping and 1 more promotion

Product Description ... web, or watching movies, the Samsung Galaxy Tab 4 features a 7.0-inch ...

Electronics: See all 4,851,153 items

Samsung Galaxy S Duos II S7582 White DUAL SIM Factory Unlocked International Ver

by Samsung

£122.00

★★★★★ 214

Library of Congress Search

Search Results for "bulgaria..."

www.loc.gov/search/?in=&q=bulgaria&new=true&st=

LOC.GOV CONGRESS.GOV COPYRIGHT.GOV

LIBRARY OF CONGRESS

Discover Services Visit Education Connect About

All Formats

Library of Congress > Search

Results for "bulgaria" 1 - 25 of 1,942

Refined by:

Refine your search Sort By Relevance View List

Available Online 1,942
 All Items 40,021

Original Formats

Web Pages	1,041
Books	600
Manuscripts/Mixed Material	327
Photos, Prints, Drawings	200
Legislation	164
Archived Web Sites	55

Bulgaria
1 negative : glass ; 5 x 7 in. or smaller | Photograph possibly shows an arch built for the celebration of the 25th anniversary of the accession to the Bulgarian throne of Ferdinand I (1861-1948). Arch has dates 1887-1912. (Source: Flickr Commons project, 2011)
Contributor: Bain News Service
Original Format: Photos, Prints, Drawings
Date: 1912

Bulgaria.
"Base 802231 (R01234) 6-94". Also issued with shaded relief. Includes note. Available also through the Library of Congress Web site as a raster image.

Examples Of Search Engines

- Conventional (library catalog)
 - Search by keyword, title, author, etc.
- Text-based (Lexis-Nexis, Google, Yahoo!)
 - Search by keywords. Limited search using queries in natural language.
- Image-based
 - shapes, colors, keywords
- Question answering systems (ask.com)
 - Search in (restricted) natural language
- Clustering systems (Vivísimo, Clusty)
- Research systems (Lemur, Nutch)

Sample Queries

- How to get rid of stretch marks
- Dodge
- Kourtney Kardashian
- How many calories are in pumpkn pie
- Angelina Jolie and Brad Pitt
- How to vote
- Derek Jeter
- Interstellar trailer
- What is Ebola?

The Size Of The World Wide Web

- The size of the indexed world wide web pages (by 2014)
 - Indexed by Google: about 45 Billion pages
 - Indexed by Bing: about 25 Billion pages

<http://www.worldwidewebsize.com/>

Web Statistics

- Twitter hits 400 million tweets per day
 - June, 2012. Dick Costolo, CEO at Twitter
- Over 2.5 billion photos uploaded to Facebook each month (2010)
 - blog.facebook.com
- Google's clusters process a total of more than 20 petabytes of data per day.
 - 2008. Jeffrey Dean from Google

Challenges

- Dynamically generated content
- New pages get added all the time
- The size of the blogosphere doubles every 6 months

Characteristics Of User Queries

- Sessions
 - users revisit their queries
- Very short queries
 - typically 2 words long
- A large number of typos
- A small number of popular queries
 - A long tail of infrequent ones
- Almost no use of advanced query operators
 - with the exception of double quotes

Information Retrieval

- Baseline Process
 - Given a collection of documents
 - And a user's query
 - Find the most relevant documents

Key Terms Used in IR

- **Query**
 - a representation of what the user is looking for – can be a list of words or a phrase.
- **Document**
 - an information entity that the user wants to retrieve
- **Collection**
 - a set of documents
- **Index**
 - a representation of information that makes querying easier
- **Term**
 - word or concept that appears in a document or a query

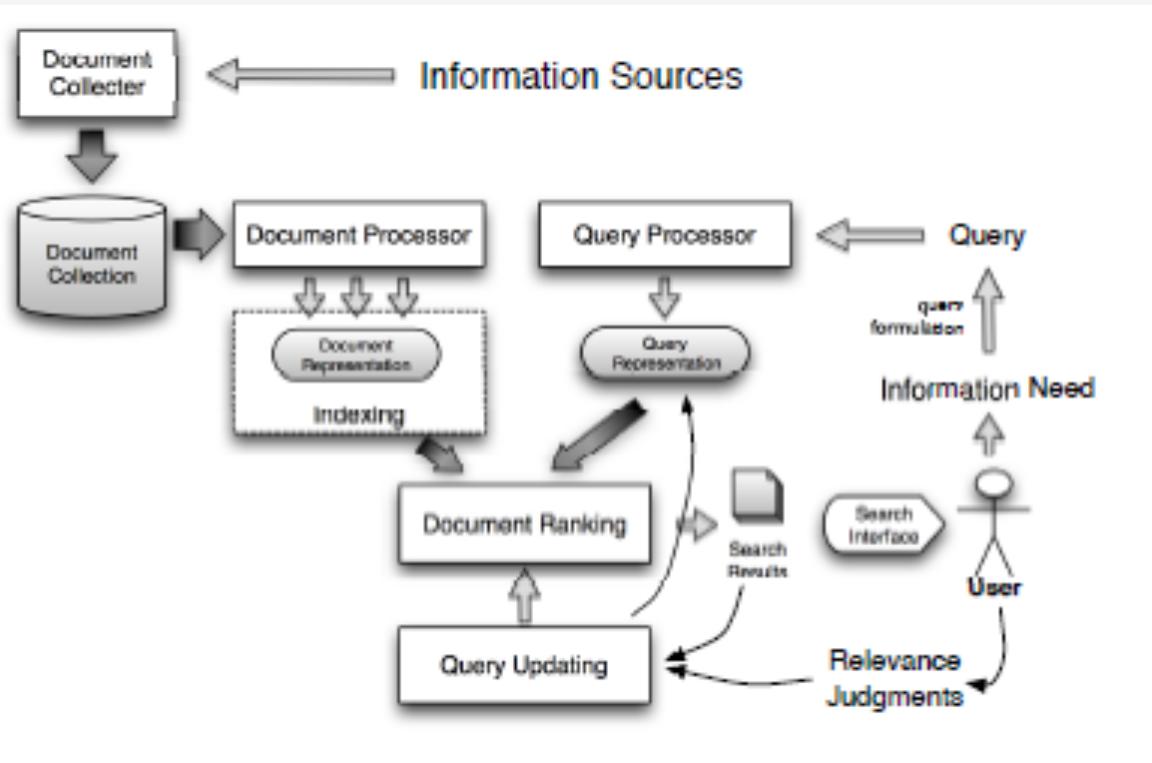
Documents

- Not just printed paper
- Can be records, pages, sites, images, people, movies
- Document encoding (Unicode)
- Document representation
- Document preprocessing (e.g., removing metadata)
- Words, terms, types, tokens

Search Engine Architecture

- Decide what to index
- Collect it
- Index it (efficiently)
- Keep the index up to date
- Provide user-friendly query facilities

Search Engine Architecture



Document Representations

- Term-document matrix ($m \times n$)
- Document-document matrix ($n \times n$)
- Typical example in a medium-sized collection
 - $n=3,000,000$ documents
 - $m=50,000$ terms
- Typical example on the Web
 - $n=30,000,000,000$
 - $m=1,000,000$
- Boolean vs. integer-valued matrices

Storage Issues

- Imagine a medium-sized collection with $n=3,000,000$ and $m=50,000$
- How large a term-document matrix will be needed?
- Is there any way to do better? Any heuristic?

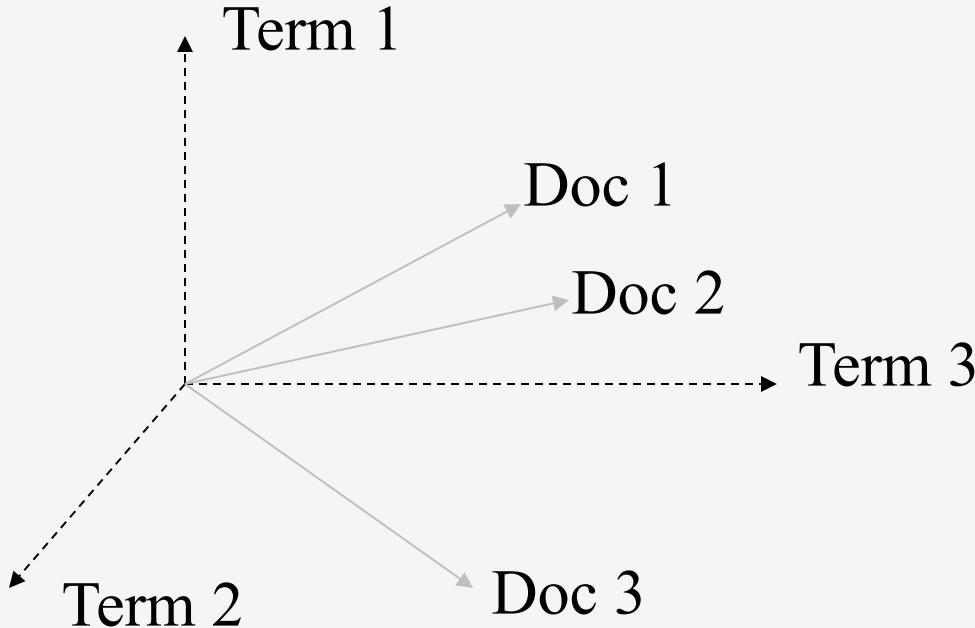
Inverted Index

- Instead of an incidence vector, use a **posting table**
 - VERMONT: D1, D2, D6
 - MASSACHUSETTS: D1, D5, D6, D7
- Use linked lists to be able to insert new document postings in order and to remove existing postings.
- Can be used to compute document frequency
- Keep everything sorted! This gives you a logarithmic improvement in access.

Basic Operations On Inverted Indexes

- Conjunction (AND)
 - iterative merge of the two postings: $O(x+y)$
- Disjunction (OR)
 - very similar
- Negation (NOT)
 - can we still do it in $O(x+y)$?
 - Example: VERMONT AND NOT MASSACHUSETTS
 - Example: MASSACHUSETTS OR NOT VERMONT
- Recursive operations
- Optimization
 - start with the smallest sets

The Vector Model



Queries as Documents

- Advantages:
 - Mathematically easier to manage
- Problems:
 - Different lengths
 - Syntactic differences
 - Repetitions of words (or lack thereof)

Vector Queries

- Each document is represented as a vector
- Non-efficient representation
- Dimensional compatibility

\mathbf{W}_1	\mathbf{W}_2	\mathbf{W}_3	\mathbf{W}_4	\mathbf{W}_5	\mathbf{W}_6	\mathbf{W}_7	\mathbf{W}_8	\mathbf{W}_9	\mathbf{W}_{10}
C ₁	C ₂	C ₃	C ₄	C ₅	C ₆	C ₇	C ₈	C ₉	C ₁₀

The Matching Process

- Document space
- Matching is done between a document and a query (or between two documents)
- Distance vs. similarity measures.
 - Euclidean distance (define)
 - Manhattan distance (define)
 - Word overlap
 - Jaccard coefficient

Similarity Measures

- The Cosine measure (normalized dot product)

$$\sigma(D, Q) = \frac{|D \cap Q|}{\sqrt{|D| \cdot |Q|}} = \frac{\sum (d_i \cdot q_i)}{\sqrt{\sum (d_i)^2} \cdot \sqrt{\sum (q_i)^2}}$$

- The Jaccard coefficient

$$\sigma(D, Q) = \frac{|D \cap Q|}{|D \cup Q|}$$

Exercise

- Compute the cosine scores
 - $\sigma(D_1, D_2)$
 - $\sigma(D_1, D_3)$
- for the documents
 - $D_1 = \langle 1, 3 \rangle$
 - $D_2 = \langle 100, 300 \rangle$
 - $D_3 = \langle 3, 1 \rangle$
- Compute the corresponding Euclidean distances, Manhattan distances, and Jaccard coefficients.

Phrase-based Queries

- Examples
 - “New York City”
 - “Ann Arbor”
 - “Barack Obama”
- We don’t want to match
 - York is a city in New Hampshire

Positional Indexing

- Keep track of all words and their positions in the documents
- To find a multi-word phrase, look for the matching words appearing next to each other

Document Ranking

- Compute the similarity between the query and each of the documents
- Use cosine similarity
- Use TF*IDF weighting
- Return the top K matches to the user

IDF: Inverse Document Frequency

- Motivation
- Example

N : number of documents

d_k : number of documents containing term k

f_{ik} : absolute frequency of term k in document i

w_{ik} : weight of term k in document i

$$\text{idf}_k = \log_2(N/d_k) + 1 = \log_2 N - \log_2 d_k + 1$$

NLP

NLP

Introduction to NLP

Evaluation of IR

Evaluation

- Size of index
- Speed of indexing
- Speed of retrieval
- Accuracy
- Timeliness
- Ease of use
- Expressiveness of search language

Contingency Table

		retrieved	not retrieved	
relevant	w=tp	x=fn	$n_1 = w + x$	
	y=fp	z=tn		
	$n_2 = w + y$			N

Precision and Recall

Recall:

$$\frac{W}{W+X}$$

Precision:

$$\frac{W}{W+Y}$$

Issues

- Why not use accuracy $A=(w+z)/N$?
- Average precision
- Report when $P=R$
- F measure:
 - $F=(\beta^2+1)PR/(\beta^2P+R)$
- F1 measure:
 - $F1 = 2/(1/R+1/P)$: harmonic mean of P and R

Sample TREC query

```
<top>
<num> Number: 305
<title> Most Dangerous Vehicles

<desc> Description:
Which are the most crashworthy, and least crashworthy,
passenger vehicles?

<narr> Narrative:
A relevant document will contain information on the crashworthiness of
a given vehicle or vehicles that can be used to draw a comparison with
other vehicles. The document will have to describe/compare vehicles,
not drivers. For instance, it should be expected that vehicles preferred
by 16-25 year-olds would be involved in more crashes, because that age
group is involved in more crashes. I would view number of fatalities
per 100 crashes to be more revealing of a vehicle's crashworthiness
than the number of crashes per 100,000 miles, for example.

</top>
```

LA031689-0177	LA042790-0172
FT922-1008	LA021790-0136
LA090190-0126	LA092289-0167
LA101190-0218	LA111189-0013
LA082690-0158	LA120189-0179
LA112590-0109	LA020490-0021
FT944-136	LA122989-0063
LA020590-0119	LA091389-0119
FT944-5300	LA072189-0048
LA052190-0048	FT944-15615
LA051689-0139	LA091589-0101
FT944-9371	LA021289-0208
LA032390-0172	

<DOCNO> LA031689-0177 </DOCNO>
<DOCID> 31701 </DOCID>
<DATE><P>March 16, 1989, Thursday, Home Edition </P></DATE>
<SECTION><P>Business; Part 4; Page 1; Column 5; Financial Desk </P></SECTION>
<LENGTH><P>586 words </P></LENGTH>
<HEADLINE><P>AGENCY TO LAUNCH STUDY OF FORD BRONCO II AFTER HIGH RATE OF ROLL-OVER ACCIDENTS </P></HEADLINE>
<BYLINE><P>By LINDA WILLIAMS, Times Staff Writer </P></BYLINE>
<TEXT>
<P>The federal government's highway safety watchdog said Wednesday that the Ford Bronco II appears to be involved in more fatal roll-over accidents than other vehicles in its class and that it will seek to determine if the vehicle itself contributes to the accidents. </P>
<P>The decision to do an engineering analysis of the Ford Motor Co. utility-sport vehicle grew out of a federal accident study of the Suzuki Samurai, said Tim Hurd, a spokesman for the National Highway Traffic Safety Administration. NHTSA looked at Samurai accidents after Consumer Reports magazine charged that the vehicle had basic design flaws. </P>
<P>Several Fatalities </P>
<P>However, the accident study showed that the "Ford Bronco II appears to have a higher number of single-vehicle, first event roll-overs, particularly those involving fatalities," Hurd said. The engineering analysis of the Bronco, the second of three levels of investigation conducted by NHTSA, will cover the 1984-1989 Bronco II models, the agency said. </P>
<P>According to a Fatal Accident Reporting System study included in the September report on the Samurai, 43 Bronco II single-vehicle roll-overs caused fatalities, or 19 of every 100,000 vehicles. There were eight Samurai fatal roll-overs, or 6 per 100,000; 13 involving the Chevrolet S10 Blazers or GMC Jimmy, or 6 per 100,000, and six fatal Jeep Cherokee roll-overs, for 2.5 per 100,000. After the accident report, NHTSA declined to investigate the Samurai. </P>
...
</TEXT>
<GRAPHIC><P> Photo, The Ford Bronco II "appears to have a higher number of single-vehicle, first event roll-overs," a federal official said. </P></GRAPHIC>
<SUBJECT>
<P>TRAFFIC ACCIDENTS; FORD MOTOR CORP; NATIONAL HIGHWAY TRAFFIC SAFETY ADMINISTRATION; VEHICLE INSPECTIONS; RECREATIONAL VEHICLES; SUZUKI MOTOR CO; AUTOMOBILE SAFETY </P>
</SUBJECT>
</DOC>

TREC (cont'd)

- <http://trec.nist.gov/tracks.html>
- <http://trec.nist.gov/presentations/presentations.html>

Most Used Reference Collections

- Generic retrieval
 - OHSUMED, CRANFIELD, CACM
- Text classification
 - Reuters, 20newsgroups
- Question answering
 - TREC-QA
- Web
 - DOTGOV, wt100g
- Blogs
 - Buzzmetrics datasets
- TREC ad hoc collections, 2–6 GB
- TREC Web collections, 2–100GB

Comparing Two Systems

- Comparing A and B
- One query?
- Average performance?
- Need: A to consistently outperform B

[Example from James Allan]

The Sign Test

- **Example 1:**
 - $A > B$ (12 times)
 - $A = B$ (25 times)
 - $A < B$ (3 times)
 - $p < 0.035$ (significant at the 5% level)
- **Example 2:**
 - $A > B$ (18 times)
 - $A < B$ (9 times)
 - $p < 0.122$ (not significant at the 5% level)
- **External link:**
 - [http://www.fon.hum.uva.nl/Service/Statistics/
Sign Test.html](http://www.fon.hum.uva.nl/Service/Statistics/Sign_Test.html)

Other Tests

- Student t-test: takes into account the actual performances, not just which system is better
 - [http://www.fon.hum.uva.nl/Service/Statistics/
Student_t_Test.html](http://www.fon.hum.uva.nl/Service/Statistics/Student_t_Test.html)
 - http://www.socialresearchmethods.net/kb/stat_t.php
- Wilcoxon Matched-Pairs Signed-Ranks Test
 - [http://www.fon.hum.uva.nl/Service/Statistics/
Signed_Rank_Test.html](http://www.fon.hum.uva.nl/Service/Statistics/Signed_Rank_Test.html)

NLP

NLP

Introduction to NLP

Text Classification

Classification

- Assigning documents to predefined categories
 - topics, languages, users
- A given set of classes C
 - Given x , determine its class in C
- Hierarchical vs. flat
- Overlapping (soft) vs non-overlapping (hard)

Classification

- Ideas: manual classification using rules
 - e.g., Columbia AND University → Education
 - Columbia AND “South Carolina” → Geography
- Popular techniques
 - generative (k-nn, Naïve Bayes) vs. discriminative (SVM, regression)
- Generative
 - model joint prob $p(x,y)$ and use Bayesian prediction to compute $p(y|x)$
- Discriminative
 - model $p(y|x)$ directly.

Representations For Document Classification (And Clustering)

- Typically: vector-based
 - Words: “cat”, “dog”, etc.
 - Features: document length, author name, etc.
- Each document is represented as a vector in an n -dimensional space
- Similar documents appear nearby in the vector space (distance measures are needed)

Naïve Bayesian classifiers

- Naïve Bayesian classifier

$$P(d \in C | F_1, F_2, \dots, F_k) = \frac{P(F_1, F_2, \dots, F_k | d \in C) P(d \in C)}{P(F_1, F_2, \dots, F_k)}$$

- Assuming statistical independence

$$P(d \in C | F_1, F_2, \dots, F_k) = \frac{\prod_{j=1}^k P(F_j | d \in C) P(d \in C)}{\prod_{j=1}^k P(F_j)}$$

- Features = words (or phrases) typically

Issues with Naïve Bayes

- Where do we get the values $P(d \in C)$
 - use maximum likelihood estimation (N_i/N)
- Same for the conditionals
 - these are based on a multinomial generator and the MLE estimator is $(T_{ji}/\sum T_{ji})$
- Smoothing is needed
 - why
 - Laplace smoothing $((T_{ji}+1)/\sum(T_{ji}+1))$
- Implementation
 - how to avoid floating point underflow

Spam Recognition

Return-Path: <ig_esq@rediffmail.com>
X-Sieve: CMU Sieve 2.2
From: "Ibrahim Galadima" <ig_esq@rediffmail.com>
Reply-To: galadima_esq@netpiper.com
To: webmaster@aclweb.org
Subject: Gooday

DEAR SIR

FUNDS FOR INVESTMENTS

THIS LETTER MAY COME TO YOU AS A SURPRISE SINCE I HAD
NO PREVIOUS CORRESPONDENCE WITH YOU

I AM THE CHAIRMAN TENDER BOARD OF INDEPENDENT
NATIONAL ELECTORAL COMMISSION INEC I GOT YOUR
CONTACT IN THE COURSE OF MY SEARCH FOR A RELIABLE
PERSON WITH WHOM TO HANDLE A VERY CONFIDENTIAL
TRANSACTION INVOLVING THE ! TRANSFER OF FUND VALUED AT
TWENTY ONE MILLION SIX HUNDRED THOUSAND UNITED STATES
DOLLARS US\$20M TO A SAFE FOREIGN ACCOUNT

SpamAssassin

- <http://spamassassin.apache.org/>
- http://spamassassin.apache.org/tests_3_3_x.html
- Examples:
 - body Incorporates a tracking ID number
 - body HTML and text parts are different
 - header Date: is 3 to 6 hours before Received: date
 - body HTML font size is huge
 - header Attempt to obfuscate words in Subject:
 - header Subject =~ /`urgent(?:[\s\W]*(dollar) | \{1,40}(?:alert| response| assistance| proposal| reply| warning| noti(?:ce| fication)| greeting| matter))/i`

Feature Selection: The X^2 Test

- For a term t :

		I_t	
	0	1	
C	0	k_{00}	k_{01}
	1	k_{10}	k_{11}
- C=class, i_t = feature
- Testing for independence: $P(C=0, I_t=0)$ should be equal to $P(C=0) P(I_t=0)$
 - $P(C=0) = (k_{00}+k_{01})/n$
 - $P(C=1) = 1-P(C=0) = (k_{10}+k_{11})/n$
 - $P(I_t=0) = (k_{00}+K_{10})/n$
 - $P(I_t=1) = 1-P(I_t=0) = (k_{01}+k_{11})/n$

Feature Selection: The X^2 Test

$$X^2 = \frac{n(k_{11}k_{00} - k_{10}k_{01})^2}{(k_{11} + k_{10})(k_{01} + k_{00})(k_{11} + k_{01})(k_{10} + k_{00})}$$

- High values of X^2 indicate lower belief in independence.
- In practice, compute X^2 for all words and pick the top k among them.

Feature Selection: Mutual Information

- No document length scaling is needed
- Documents are assumed to be generated according to the multinomial model
- Measures amount of information: if the distribution is the same as the background distribution, then $MI=0$
- $X = \text{word}; Y = \text{class}$

$$MI(X, Y) = \sum_x \sum_y P(x, y) \log \frac{P(x, y)}{P(x)P(y)}$$

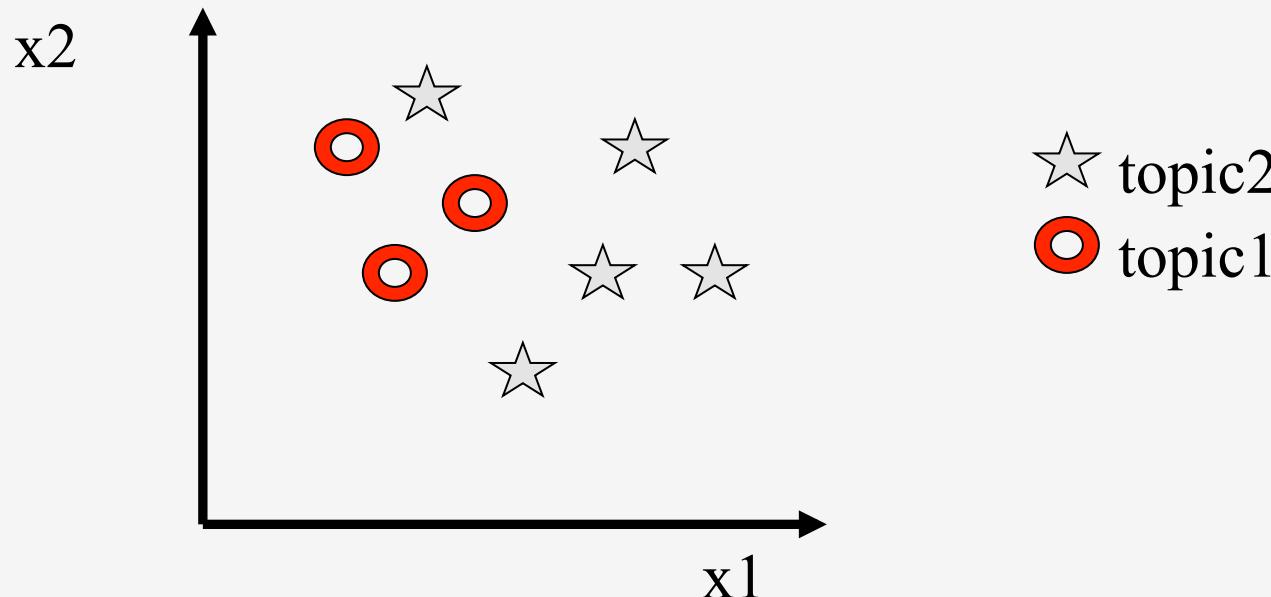
Well-known Datasets

- 20 newsgroups
 - <http://qwone.com/~jason/20Newsgroups/>
- Reuters-21578
 - <http://www.daviddlewis.com/resources/testcollections/reuters21578/>
 - Cats: grain, acquisitions, corn, crude, wheat, trade...
- WebKB
 - <http://www-2.cs.cmu.edu/~webkb/>
 - course, student, faculty, staff, project, dept, other
- RCV1
 - <http://www.daviddlewis.com/resources/testcollections/rcv1/>
 - Larger Reuters corpus

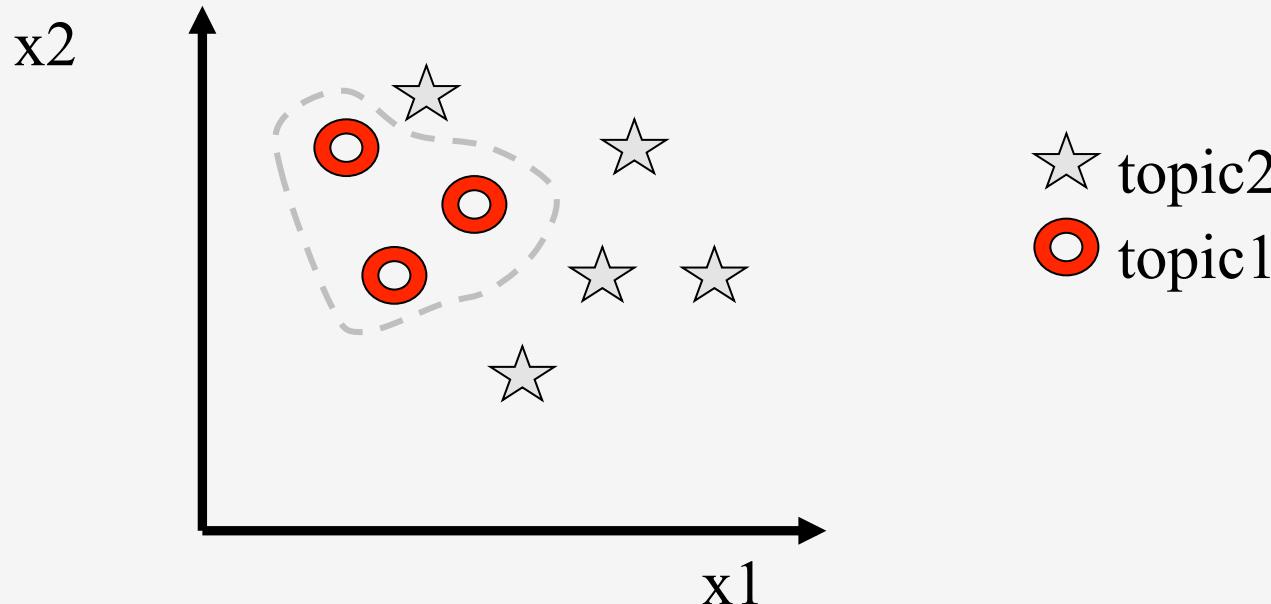
Evaluation Of Text Classification

- Microaveraging
 - average over classes
- Macroaveraging
 - uses pooled table

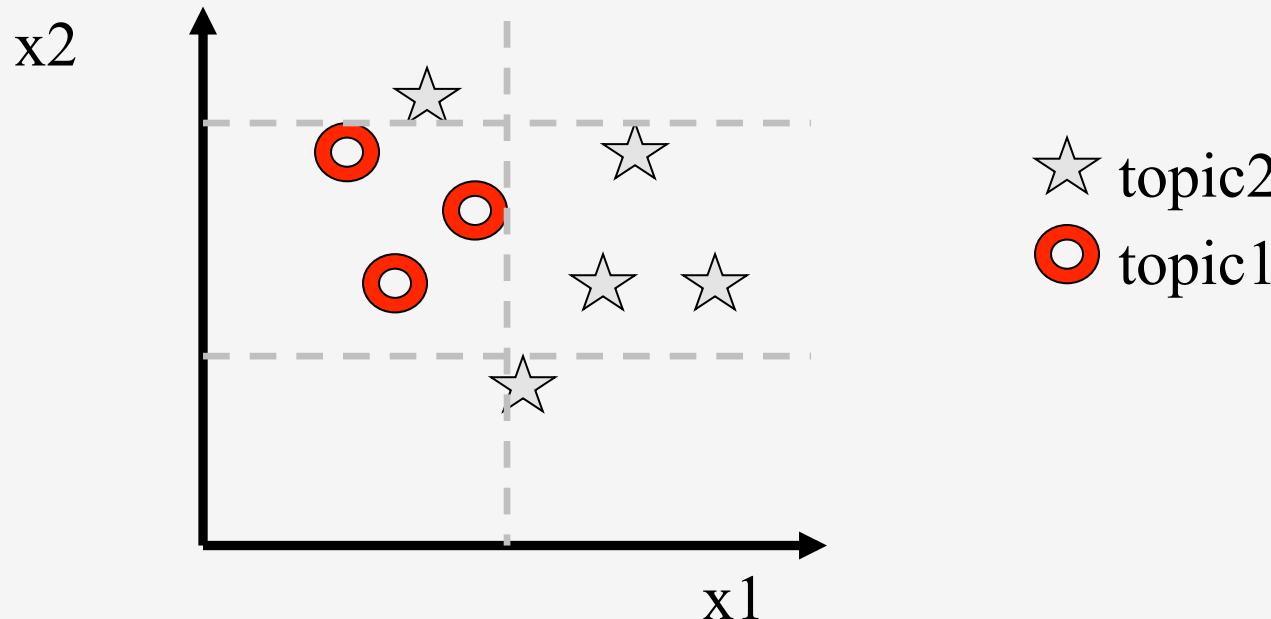
Vector Space Classification



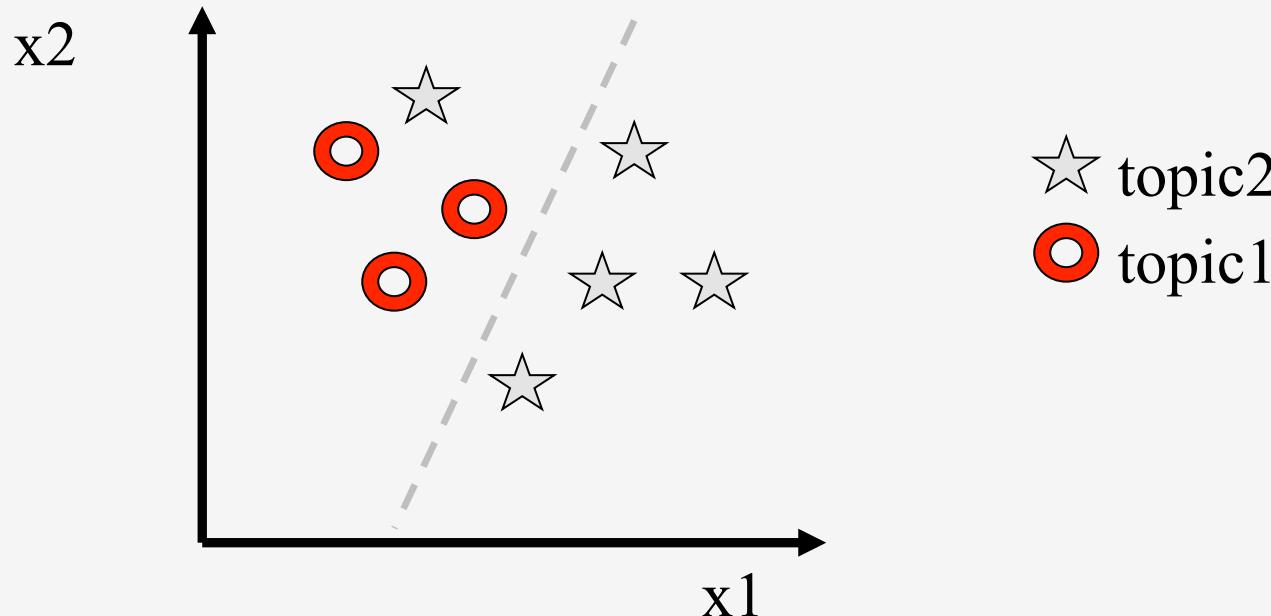
Decision Surfaces



Decision Trees



Linear Boundary



Vector Space Classifiers

- Using centroids
- Boundary
 - line that is equidistant from two centroids

Linear Separators

- Two-dimensional line:

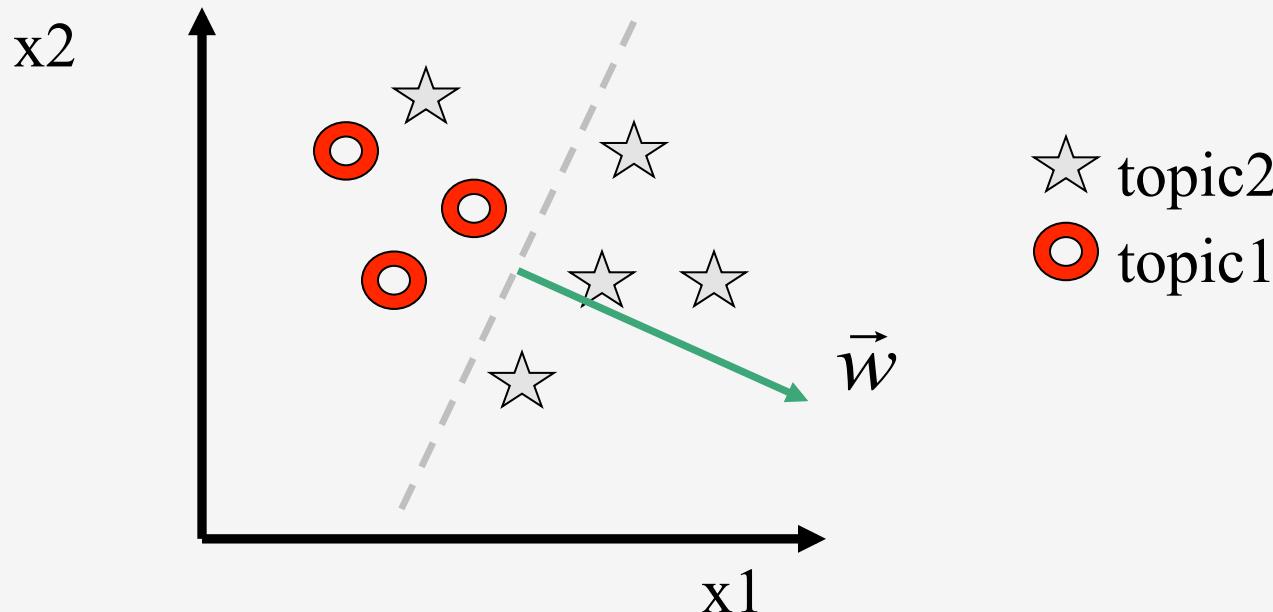
$w_1x_1 + w_2x_2 = b$ is the linear separator

$w_1x_1 + w_2x_2 > b$ for the positive class

- In n-dimensional spaces:

$$\vec{w}^T \vec{x} = b$$

Decision Boundary



Example

- Bias $b=0$
- Document is “A D E H”
- Its score will be

$$\begin{aligned} & 0.6*1+0.4*1+0.4*1+(-0.5)*1 \\ & = 0.9 > 0 \end{aligned}$$

w_i	x_i	w_i	x_i
0.6	A	-0.7	G
0.5	B	-0.5	H
0.5	C	-0.3	I
0.4	D	-0.2	J
0.4	E	-0.2	K
0.3	F	-0.2	L

Perceptron Algorithm

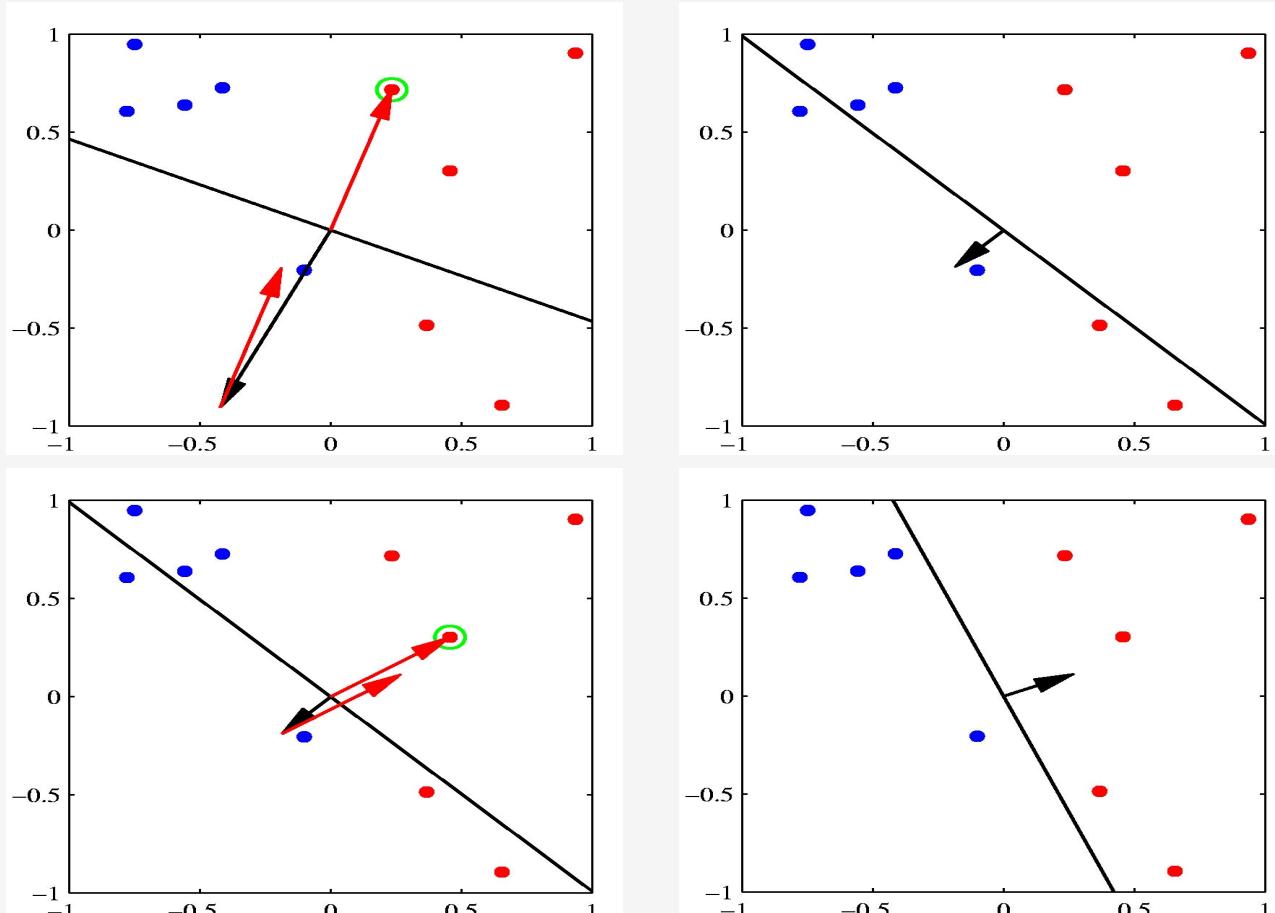
Input: $S = ((\vec{x}_1, y_1), \dots, (\vec{x}_n, y_n)), \vec{x}_i \in \Re^N, y_i \in \{-1, 1\}$
 $\eta \in \Re$

Algorithm:

```
     $\vec{w}_0 = \vec{0}, k = 0$ 
    FOR  $i = 1$  TO  $n$ 
        IF  $y_i(\vec{w}_k \bullet \vec{x}_i) \leq 0$ 
             $\vec{w}_{k+1} = \vec{w}_k + \eta y_i \vec{x}_i$ 
             $k = k + 1$ 
    END
END
```

Output: \vec{w}_k

10.04 Text Classification



Generative Models: knn

- Assign each element to the closest cluster
- K-nearest neighbors

$$\text{score}(c, d_q) = b_c + \sum_{d \in kNN(d_q)} s(d_q, d)$$

- Very easy to program
- Issues:
 - choosing k, b?
- Demo:
 - <http://www-2.cs.cmu.edu/~zhuxj/courseproject/knndemo/KNN.html>

NLP

NLP

Introduction to NLP

Text Clustering

Clustering

- Exclusive/overlapping clusters
- Hierarchical/flat clusters
- The cluster hypothesis
 - Documents in the same cluster are relevant to the same query
 - How do we use it in practice?

Example

Clusty Search » jaguar - Mozilla Firefox

File Edit View History Bookmarks ScrapBook Tools Help

http://clusty.com/search?v%3afile=viv_698%4031%3aPKOn4R&v%3afame=tree&v%3astate=%28ro

web news images wikipedia blogs jobs more »

jaguar Search advanced preferences

clusters sources sites

All Results (223)

- Parts (53)
- Jaguar Cars (40)
- Club (36)
- Pictures (23)
- Classic (15)
- Dealer (13)
- Zookeeper, Denver (8)
- Panthera onca (6)
 - Cat (4)
 - Other Topics (2)
- Type Jaguar (3)
- Motor (8)

more | all clusters

find in clusters: Find

Font size: A A A A

Find: radev Next Previous Highlight all Match case

FoxyT Open Notebook

Waiting for wikipedia1.clusty.com...

Cluster **Panthera onca** contains 6 documents.

Jaguar

Visit the Official **Jaguar** Site for more info and to find a dealer. - www.JaguarUSA.com

Jaguar

The **jaguar** (*Panthera onca*) is a large member of the cat family native to warm regions of the [Americas](#). It is closely related to the [lion](#), [tiger](#), and [leopard](#) of the [Old World](#), and is the largest species of the cat family found in the Americas.
en.wikipedia.org/wiki/Jaguar - [cache] - Wikipedia, MSN, Ask

Jaguar

Jaguar may refer to: A **jaguar** (*Panthera onca*), a large felid native to South and Central America Grumman F10 **Jaguar** a military aircraft SEPECAT **Jaguar** , a military ... aircraft **Jaguar Cars** , British automobile maker **Jaguar** Racing , a former ...
[en.wikipedia.org/wiki/Jaguar_\(disambiguation\)](http://en.wikipedia.org/wiki/Jaguar_(disambiguation)) - [cache] - Wikipedia

Jaguar

Panthera onca. MYSTERIOUS CAT OF THE AMAZON. Of all the big cats, the **jaguar** remains the least studied. While some information comes from the wild, most of what is known about **jaguars** has been learned ...
www.bluelion.org/jaguar.htm - [cache] - MSN, Ask

Jaguar (Panthera onca)

Jaguar (Panthera onca) facts, photos and videos. ... The **Jaguar** is the largest cat in the Western Hemisphere and the third largest cat in ...
www.thebigzoo.com/Animals/Jaguar.asp - [cache] - Ask

jaguar - Definitions from [Dictionary.com](#)

k-means

- Iteratively determine which cluster a point belongs to, then adjust the cluster centroid, then repeat
- Needed: small number k of desired clusters
- hard decisions

k-means

```
1 initialize cluster centroids to arbitrary vectors
2 while further improvement is possible do
3   for each document  $d$  do
4     find the cluster  $c$  whose centroid is closest to  $d$ 
5     assign  $d$  to cluster  $c$ 
6   end for
7   for each cluster  $c$  do
8     recompute the centroid of cluster  $c$  based on its
      documents
9   end for
10  end while
```

Example

- Cluster the following vectors into two groups:
 - A = <1,6>
 - B = <2,2>
 - C = <4,0>
 - D = <3,3>
 - E = <2,5>
 - F = <2,1>

Demos

- http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletKM.html
- http://cgm.cs.mcgill.ca/~godfried/student_projects/bonnef_k-means
- <http://www.cs.washington.edu/research/imagedatabase/demo/kmcluster>
- <http://www.cc.gatech.edu/~dellaert/FrankDellaert/Software.html>
- <http://www-2.cs.cmu.edu/~awm/tutorials/kmeans11.pdf>
- <http://web.archive.org/web/20110223234358/http://www.ece.neu.edu/groups/rpl/projects/kmeans/>

Evaluation of Clustering

- Purity
 - considering the majority class in each cluster
- RAND index
 - See next slide

Purity

- Three clusters

XXXOO

OOOX%

% % % % XX

- Purity:
 - $(3+3+4)/16=62.5\%$

Rand Index

- Accuracy when preserving object-object relationships.
- $RI = (TP + TN) / (TP + FP + FN + TN)$
- In the example:

$$TP + FP = \binom{5}{2} + \binom{5}{2} + \binom{6}{2} = 35$$

$$TP = \binom{3}{2} + \binom{3}{2} + \binom{4}{2} + \binom{2}{2} = 13$$

$$FP = 35 - 13 = 22$$

Rand Index

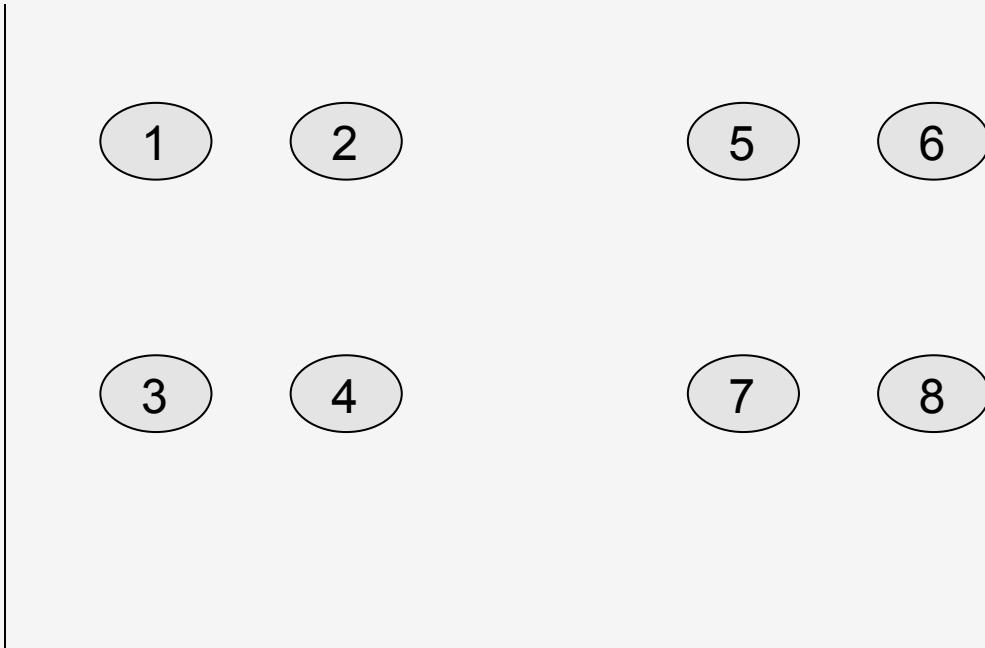
Same cluster		
Same class	TP=13	FN=21
	FP=22	TN=64

$$RI = (TP+TN)/(TP+TN+FP+FN) = (13+64)/(13+64+22+21) = 0.64$$

Hierarchical Clustering Methods

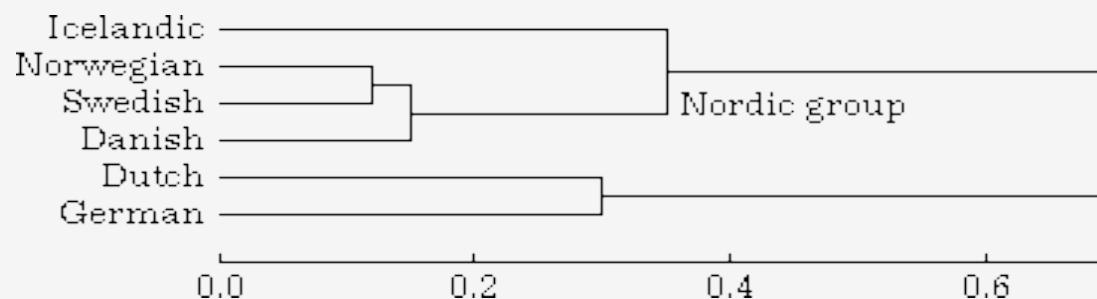
- Single-linkage
 - One common pair is sufficient
 - disadvantages: long chains
- Complete-linkage
 - All pairs have to match
 - Disadvantages: too conservative
- Average-linkage

Hierarchical Clustering



Hierarchical Agglomerative Clustering Dendrograms

E.g., language similarity:



<http://odur.let.rug.nl/~kleiweg/clustering/clustering.html>

Clustering Using Dendograms

Example: cluster the following sentences:

A B C B A
A D C C A D E
C D E F C D A
E F G F D A
A C D A B A

REPEAT

- Compute pairwise similarities
- Identify closest pair
- Merge pair into single node

UNTIL only one node left

Q: what is the equivalent Venn diagram representation?

NLP

NLP

Introduction to NLP

Information Retrieval Toolkits

Open Source IR Toolkits

- Smart (Cornell)
- MG (RMIT & Melbourne, Australia; Waikato, New Zealand),
- Lemur (CMU/Univ. of Massachusetts)
- Terrier (Glasgow)
- Clairlib (University of Michigan)
- Lucene/SOLR (Apache)

Smart

- The most influential IR system/toolkit
- Developed at Cornell since 1960's
- Vector space model with lots of weighting options
- Written in C
- The Cornell/AT&T groups have used the Smart system to achieve top TREC performance

MG

- A highly efficient toolkit for retrieval of text and images
- Developed by people at Univ. of Waikato, Univ. of Melbourne, and RMIT in 1990's
- Written in C, running on Unix
- Vector space model with lots of compression and speed up tricks
- People have used it to achieve good TREC performance

Lemur/Indri

- An IR toolkit emphasizing language models
- Developed at CMU and Univ. of Massachusetts in 2000's
- Written in C++, highly extensible
- Vector space and probabilistic models including language models
- Achieving good TREC performance with a simple language model

Lucene

- Open Source IR toolkit
- Initially developed by Doug Cutting in Java
- Now has been ported to some other languages
- Good for building IR/Web applications
- Many applications have been built using Lucene (e.g., Nutch and SOLR)

NLP

NLP

Introduction to NLP

Sentiment Analysis

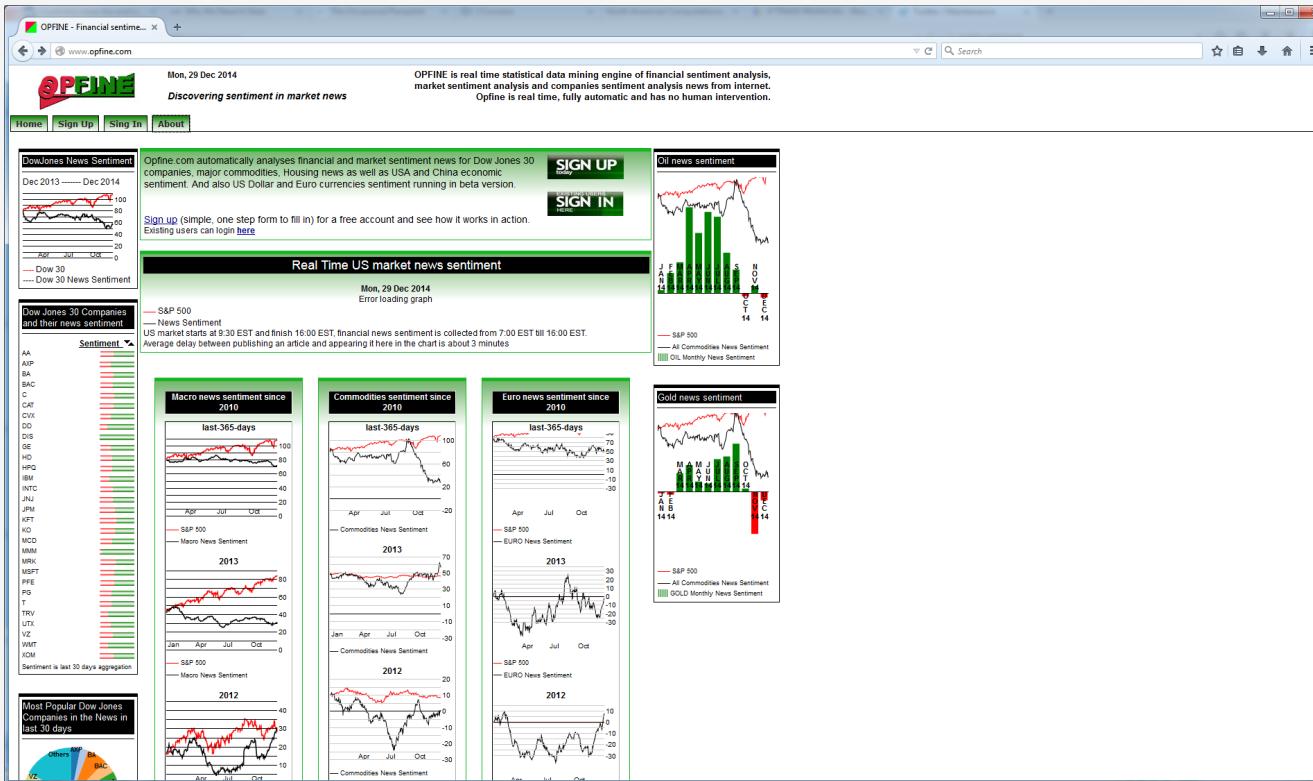
Reviews of 1Q84 by Haruki Murakami

- “1Q84 is a tremendous feat and a triumph . . . A must-read for anyone who wants to come to terms with contemporary Japanese culture.”
—Lindsay Howell, *Baltimore Examiner*
- “Perhaps one of the most important works of science fiction of the year . . . 1Q84 does not disappoint . . . [It] envelops the reader in a shifting world of strange cults and peculiar characters that is surreal and entrancing.”
—Matt Staggs, Suvudu.com
- Ambitious, sprawling and thoroughly stunning . . . Orwellian dystopia, sci-fi, the modern world (terrorism, drugs, apathy, pop novels)—all blend in this dreamlike, strange and wholly unforgettable epic.”
—Kirkus Reviews (starred review)

Reviews of 1Q84 by Haruki Murakami

- “1Q84 is a **tremendous feat** and a **triumph** . . . A **must-read** for anyone who wants to come to terms with contemporary Japanese culture.”
—Lindsay Howell, *Baltimore Examiner*
- “Perhaps one of the most important works of science fiction of the year . . . 1Q84 does not disappoint . . . [It] envelops the reader in a shifting world of strange cults and peculiar characters that is surreal and **entrancing**.”
—Matt Staggs, Suvudu.com
- **Ambitious**, sprawling and thoroughly **stunning** . . . Orwellian dystopia, sci-fi, the modern world (terrorism, drugs, apathy, pop novels)—all blend in this dreamlike, strange and wholly **unforgettable epic**.”
—Kirkus Reviews (starred review)

Sentiment About Companies



Other Examples

- Movie reviews
- Product reviews
- Debates
 - www.createdebate.com

Introduction

- Many posts, blogs
- Expressing personal opinions
- Research questions
 - Subjectivity analysis
 - Polarity analysis (positive/negative, number of stars)
 - Viewpoint analysis (Chelsea vs. Manchester United, republican vs. democrat)
- Sentiment target
 - entity
 - aspect

Introduction

- Level of granularity
 - Document
 - Sentence
 - Attribute
- Opinion words
 - Base
 - Comparative (better, slower)

Introduction

- Just counting negative words is not enough
- Negation analysis

Reviews of 1Q84 by Haruki Murakami

- “*1Q84* is a tremendous feat and a triumph . . . A must-read for anyone who wants to come to terms with contemporary Japanese culture.”
—Lindsay Howell, *Baltimore Examiner*
- “Perhaps one of the most important works of science fiction of the year . . . *1Q84* does not disappoint . . . [It] envelops the reader in a shifting world of strange cults and peculiar characters that is surreal and entrancing.”
—Matt Staggs, Suvudu.com
- Ambitious, sprawling and thoroughly stunning . . . Orwellian dystopia, sci-fi, the modern world (terrorism, drugs, apathy, pop novels)—all blend in this dreamlike, strange and wholly unforgettable epic.”
—*Kirkus Reviews* (starred review)

Reviews of 1Q84 by Haruki Murakami

- “*1Q84* is a tremendous feat and a triumph . . . A must-read for anyone who wants to come to terms with contemporary Japanese culture.”
—Lindsay Howell, *Baltimore Examiner*
- “Perhaps one of the most important works of science fiction of the year . . . *1Q84* **does not disappoint** . . . [It] envelops the reader in a shifting world of strange cults and peculiar characters that is surreal and entrancing.”
—Matt Staggs, Suvudu.com
- Ambitious, sprawling and thoroughly stunning . . . Orwellian dystopia, sci-fi, the modern world (terrorism, drugs, apathy, pop novels)—all blend in this dreamlike, strange and wholly unforgettable epic.”
—*Kirkus Reviews* (starred review)

Product Reviews

The screenshot shows a Google search result for a Samsung Galaxy S5 product review. The page title is "Samsung Galaxy S5 Android Phone 16 GB - Shimmery White - Verizon - CDMA". It displays a price of \$50 online and \$50 nearby, along with a 4-star rating from 2,839 reviews. Below this, there's a section titled "Reviews" showing 2,839 reviews, with a navigation bar for star ratings (1, 2, 3, 4 stars) followed by a green "5 stars" bar. A section titled "What people are saying" lists various positive reviews for the phone's features like battery, size, camera, and screen. At the bottom, a user review by "Larrcom10" is shown, providing a detailed description of their experience with the phone.

Samsung Galaxy S5 Android Phone 16 GB - Shimmery White - Verizon - CDMA

\$50 online, \$50 nearby ★★★★☆ 2,839 reviews

Reviews

2,839 reviews

1 2 3 4 stars 5 stars

What people are saying

Feature	Review
battery	"This device has great battery life!"
size	"Awesome phone, light weight, easy to use."
camera	"Love the great pics you can take."
features	"Great phone - love all the features and ease of use."
screen	"Great phone, love the screen!!!"
design	"Love this phone, camera and overall functionality."
speaker/headset	"Fast processor, great receptions and clear speaker."

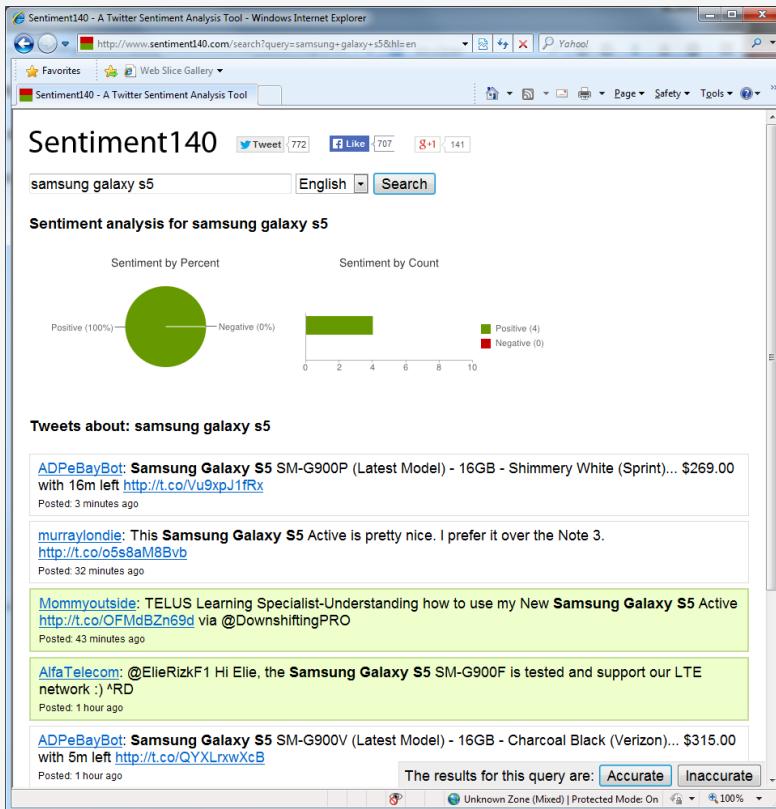
Write a review

Larrcom10 - Full review provided by Best Buy

★★★★☆ so far I am on the fence - May 11, 2014

Ok I had an s3. I have had this phone for 3 weeks. Fingerprint reader worthless it tries 5 times but then you end up typing your password. My power button seems to only turn the phone on first push when it wants to. I tried to return it thinking defective but missed the 14 day return by 2 days. so I have to deal with samsung directly. Not happy about this. The screen scratches easy get a protector. There are more scratches in 3 weeks than 2 years on the s3. I don't think it's glass. Heart monitor worthless. Anyway I am not totally unhappy though. I could not go back to my s3 now. volume is great. Pictures look great. Battery better than s3. Fitness software is cool I work outside and it tracks my steps. I went for a bike ride today, put it on bicycle and it was

Twitter Sentiment



Problems

- Subtlety
- Concession
- Manipulation
- Sarcasm and irony

SA as a Classification Problem

- Set of features
 - Words
 - Presence is more important than frequency
 - Punctuation
 - Phrases
 - Syntax
- A lot of training data is available
 - E.g., movie review sentences and stars
 - <http://www.cs.cornell.edu/people/pabo/movie-review-data>
- Techniques
 - MaxEnt
 - SVM
 - Naive Bayes

Resources

- CMU Twitter parser
 - <http://www.ark.cs.cmu.edu/TweetNLP/>

NLP

NLP

Introduction to NLP

Sentiment Lexicons

Sentiment Lexicons

- SentiWordNet
 - <http://sentiwordnet.isti.cnr.it/>
- General Inquirer
 - 2,000 positive words and 2,000 negative words
 - <http://www.wjh.harvard.edu/~inquirer/>
- LIWC
 - <http://www.liwc.net/>
- MPQA subjectivity lexicon
 - http://www.cs.pitt.edu/mpqa/subj_lexicon.html

General Inquirer

- Annotations
 - Strong Power Weak Submit Active Passive Pleasur Pain Feel Arousal EMOT Virtue Vice Ovrst Undrst Academ Doctrin Econ@ Exch ECON Exprsv Legal Milit Polit@ POLIT Relig Role COLL Work Ritual SocRel Race Kin@ MALE Female Nonadlt HU ANI PLACE Social Region Route Aquatic Land Sky Object Tool Food Vehicle BldgPt CommObj NatObj BodyPt ComForm COM Say Need Goal Try Means Persist Complet Fail NatrPro Begin Vary Increas Decreas Finish Stay Rise Exert Fetch Travel Fall Think Know Causal Ought Perceiv Compare Eval@ EVAL Solve Abs@ ABS Quality Quan NUMB ORD CARD FREQ DIST Time@ TIME Space POS DIM Rel COLOR Self Our You Name Yes No Negate Intrj IAV DAV SV IPadj IndAdj PowGain PowLoss PowEnds PowAren PowCon PowCoop PowAuPt PowPt PowDoct PowAuth PowOth PowTot RcEthic RcRelig RcGain RcLoss RcEnds RcTot RspGain RspLoss RspOth RspTot AffGain AffLoss AffPt AffOth AffTot WltPt WltTran WltOth WltTot WlbGain WlbLoss WlbPhys WlbPsyc WlbPt WlbTot EnlGain EnlLoss EnlEnds EnlPt EnlOth EnlTot SklAsth SklPt SklOth SklTot TrnGain TrnLoss TranLw MeansLw EndsLw ArenaLw PtLw Nation Anomie NegAff PosAff SureLw If NotLw TimeSpc
- <http://www.webuse.umd.edu:9090/tags/>
 - Positive: able, accolade, accuracy, adept, adequate...
 - Negative: addiction, adversity, adultery, affliction, aggressive...

Dictionary-based Methods

- Start from known seeds
 - e.g., happy, angry
- Expand using WordNet
 - synonyms
 - hypernyms
- Random-walk based methods
 - words with known polarity as absorbing boundary

Automatic Extraction of Sentiment Words

- Semi-supervised methods

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the Semantic Orientation of Adjectives. ACL, 174–181

Molistic

- NACLO problem (2007)

Imagine that you have heard these sentences:

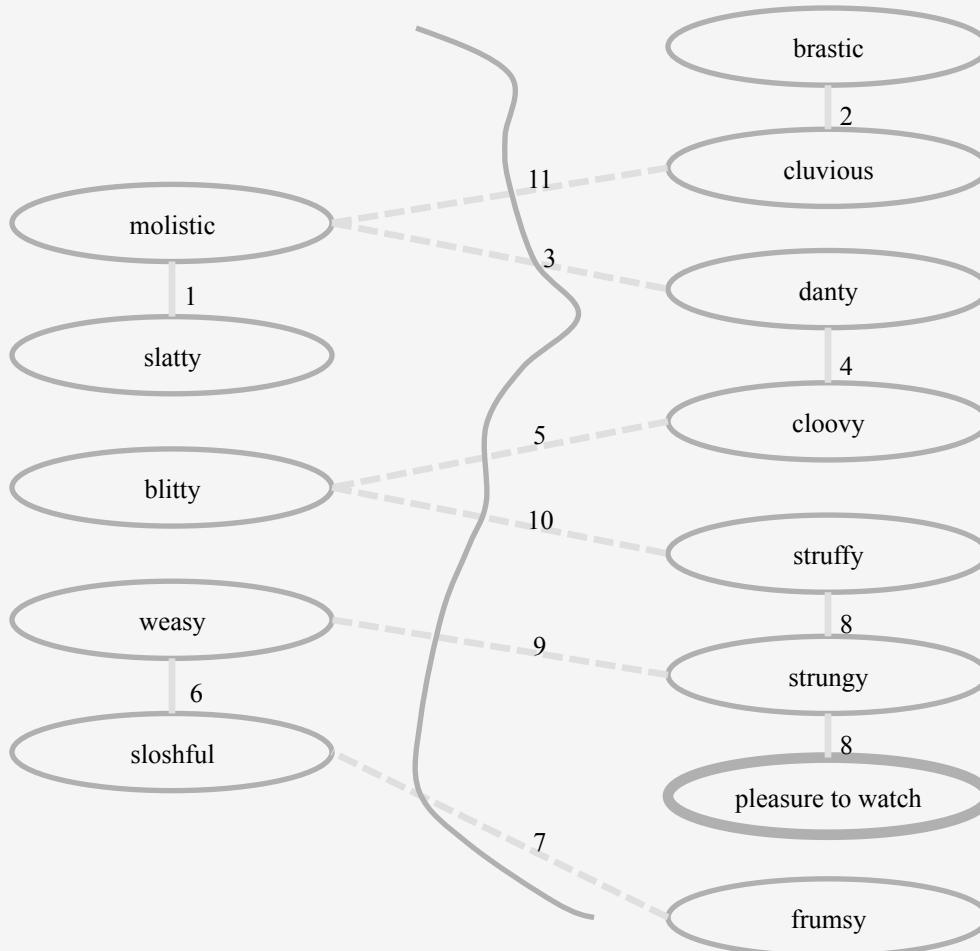
Jane is molistic and slatty.
Jennifer is cluvious and brastic.
Molly and Kyle are slatty but danty.
The teacher is danty and cloovy.
Mary is blitty but cloovy.
Jeremiah is not only sloshful but also weasy.
Even though frumsy, Jim is sloshful.
Strungy and struffy, Diane was a pleasure to watch.
Even though weasy, John is strungy.
Carla is blitty but struffy.
The salespeople were cluvious and not slatty.

A1. Then which of the following would you be likely to hear?

- a. Meredith is blitty and brastic.
- b. The singer was not only molistic but also cluvious.
- c. May found a dog that was danty but sloshful.

A2. What quality or qualities would you be looking for in a person?

- a. blitty
- b. weasy
- c. sloshful



PMI (Turney)

- PMI=pointwise mutual information
- Check how often a given unlabeled word appears with a known positive word (“excellent”)
- Same for a known negative word (“poor”)

$$\text{PMI}(word_1, word_2) = \log_2 \frac{\text{hits}(word_1 \text{ NEAR } word_2)}{\text{hits}(word_1)\text{hits}(word_2)}$$

Dataset

- [http://www.cs.jhu.edu/~mdredze/datasets/
sentiment/](http://www.cs.jhu.edu/~mdredze/datasets/sentiment/)

NLP

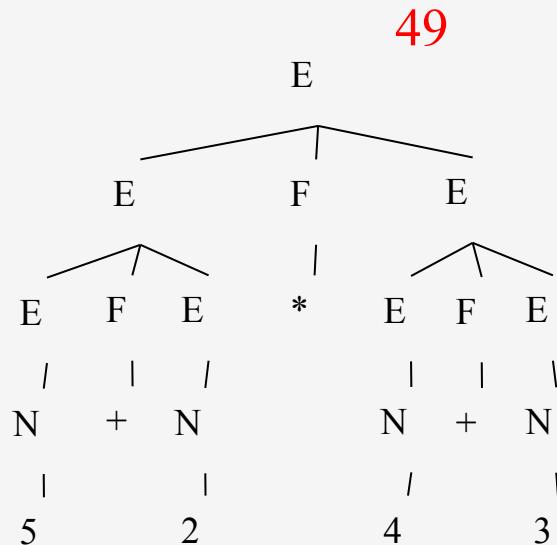
NLP

Introduction to NLP

Semantics

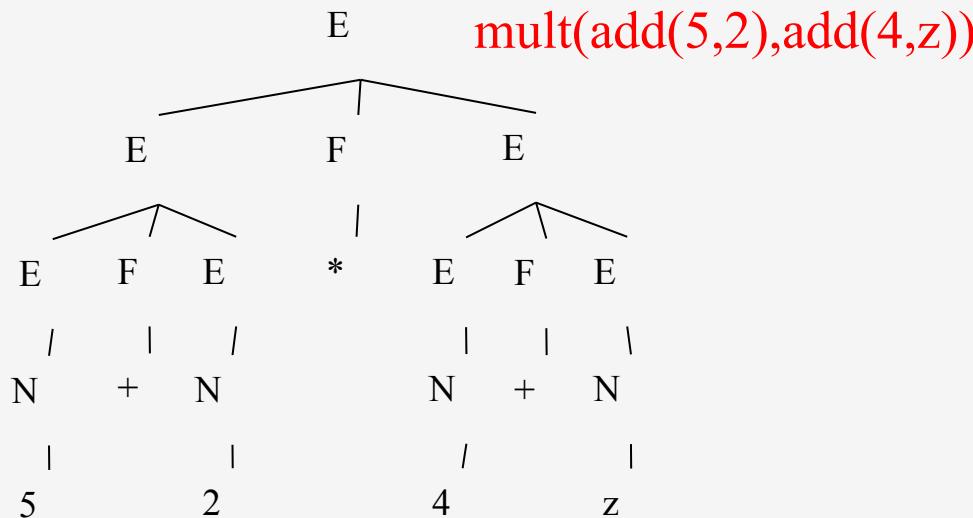
Semantics

- What is the meaning of: $(5+2)*(4+3)$?
- Parse tree



Semantics

- What if we had $(5+2)^*(4+z)$?



What About (English) Sentences?

- Every human is mortal.
- ??

Representing Meaning

- Capturing the meaning of linguistic utterances using formal notation
- Linguistic meaning
 - “It is 8 pm”
- Pragmatic meaning
 - “It is time to leave”
- Semantic analysis:
 - Assign each word a meaning
 - Combine the meanings of words into sentences
- *I bought a book:*
 $\exists x,y: \text{Buying}(x) \wedge \text{Buyer}(\text{speaker},x) \wedge \text{BoughtItem}(y,x) \wedge \text{Book}(y)$
Buying (Buyer=speaker, BoughtItem=book)

Entailment and Presupposition

- Entailment
 - One fact follows from another
 - “All cats have whiskers” and “Martin is a cat” entail the statement “Martin has whiskers”
 - “Martin has whiskers and a tail” entails “Martin has whiskers”
- Presupposition
 - “The Queen of Utopia is dead” presupposes that Utopia has a queen

Entailment and Presupposition

- NACLO problem from 2010
- Author: Aleka Blackwell
- <http://www.nacloweb.org/resources/problems/2010/M.pdf>
- <http://www.nacloweb.org/resources/problems/2010/MS.pdf>

Think about the meaning of the following sentence:

- (1) The 2010 Winter Olympics were in Canada.

Assuming that we only know sentence 1 to be true, is sentence 2 necessarily true?

- (2) The 2010 Winter Olympics were in Vancouver.

The answer is no. Assuming we only know sentence 1 to be true, the 2010 Winter Olympics could have taken place in any Canadian city, but not necessarily in Vancouver.

Now examine the relationship between sentences 3 and 4. Assuming sentence 3 is true, is sentence 4 now necessarily true?

- (3) The 2010 Winter Olympics were in Vancouver.
- (4) The 2010 Winter Olympics were in Canada.

Now the answer is yes. Since Vancouver is a Canadian city, any event which occurs in Vancouver necessarily occurs in Canada.

The logical relationship which holds between sentences 3 and 4 is called an entailment. In formal terms, sentence A entails sentence B if whenever A is true, B is necessarily true. The entailment relationship is typically represented graphically this way: A \Vdash B.

Here are some more examples of the entailment relationship between sentences:

- (5) Shaun White is a Winter Olympian \Vdash Shaun White is an Olympian
- (6) Shaun White is an Olympian \Vdash Shaun White is an athlete
- (7) Shaun White won a gold medal \Vdash Someone won a gold medal

Notice that the entailment relationship must hold in the specified direction but will not necessarily hold in both directions. So, sentence 3 entails sentence 4 even though sentence 4 does not entail sentence 3.

Entailment and Presupposition

Now examine the relationship between sentences 8 and 9.

- (8) I did not see Shaun White win the gold medal in the 2010 Winter Olympics.
- (9) Shaun White won the gold medal in the 2010 Winter Olympics.

Sentences 8 and 9 illustrate a relationship called presupposition. In this pair of sentences, the information presented in sentence 9 is what the speaker assumes (or presupposes) to be the case when uttering sentence 8. That is, to say "*I did not see Shaun White win the gold medal*" assumes the belief that Shaun White won a gold medal. In formal terms, sentence A presupposes sentence B if A not only implies B but also implies that the truth of B is somehow taken for granted. A presupposition of a sentence is thus part of the background against which its truth or falsity is judged. The presupposition relationship is typically represented graphically this way: A >> B

Here are some more examples of presuppositions (where the first sentence in each pair presupposes the second):

- (10) I regret not seeing Shaun White's gold medal run >> Shaun White had a gold medal run
- (11) Shaun White continues to rule the halfpipe >> Shaun White had been ruling the halfpipe
- (12) Snowboarding is now an Olympic sport >> Snowboarding was once not an Olympic sport

Entailment and Presupposition

For any given pair of sentences, the entailment and presupposition relationships may or may not hold, together or separately.

For each of the following possible combinations, your task is to provide one example of a pair of sentences with an explanation of your reasoning for proposing your pair of sentences as a valid and convincing example in each case.

- a. A pair of sentences in which sentence A **neither entails nor presupposes** sentence B.
- b. A pair of sentences in which sentence A **entails and presupposes** sentence B.
- c. A pair of sentences in which sentence A **presupposes but does not entail** sentence B.
- d. A pair of sentences in which sentence A **entails but does not presuppose** sentence B.

Answers

For any given pair of sentences, the entailment and presupposition relationships may or may not hold, together or separately.

- a. A pair of sentences in which sentence A **neither entails nor presupposes** sentence B.
- A. Shaun White is a Winter Olympian.
 - B. The 2010 Winter Olympics were in Vancouver.

Explanation: Sentences A and B are unrelated.

Entailment: Given that sentence A is true, there is no way to know whether sentence B is true or false. If Shaun White is a Winter Olympian, the 2010 Winter Olympics may or may not have taken place in Vancouver. Thus, there is no entailment relationship between these two sentences.

Presupposition: When uttering sentence A, a speaker would not take sentence B for granted (or assume that sentence B is background information against which the truth or falsity of sentence A would be judged). A speaker would not utter "Shaun White is a Winter Olympian" and assume the belief/take for granted that the 2010 Winter Olympics were in Vancouver.

- b. A pair of sentences in which sentence A **entails and presupposes** sentence B.
- A. Shaun White continues to rule the halfpipe
 - B. Shaun White had been ruling the halfpipe.

Entailment: If sentence A is true, sentence B is necessarily true. The entailment relationship between these sentences relies on the meaning of the verb *continue* – to *continue to rule* the halfpipe, Shaun White had to be ruling the halfpipe already. Thus, sentence A entails sentence B.

Presupposition: When uttering sentence A, a speaker would take sentence B for granted (or assume that sentence B is background information against which the truth or falsity of sentence A would be judged). A speaker who utters "Shaun White continues to rule the halfpipe" assumes the belief/takes for granted that Shaun White had been ruling the halfpipe. Thus, sentence A presupposes sentence B.

Entailment and Presupposition

c. A pair of sentences in which sentence A **presupposes but does not entail** sentence B.

- A. I did not see Shaun White win the gold medal in the 2010 Winter Olympics.
- B. Shaun White won the gold medal in the 2010 Winter Olympics.

Entailment: If sentence A is true, sentence B *may or may not* be true. The absence of an entailment relationship between these sentences relies on the words “did not see” – if it is true that I *did not* see Shaun White win the gold medal, then Shaun White *may or may not* have won the gold medal. Thus, sentence A does not entail sentence B.

Presupposition: When uttering sentence A, a speaker would take sentence B for granted (or assume that sentence B is background information against which the truth or falsity of sentence A would be judged). Specifically, a speaker who utters “I did not see Shaun White win the gold medal in the 2010 Winter Olympics” assumes the belief that Shaun White did actually win the gold medal in the 2010 Winter Olympics. Thus, sentence A presupposes sentence B.

Entailment and Presupposition

- d. A pair of sentences in which sentence A **entails but does not presuppose** sentence B.
- A. Shaun White did not win the gold medal in the 2010 Winter Olympics.
 - B. Shaun White did not both win the gold medal in the 2010 Winter Olympics and injure his ankle.

Entailment: If Shaun White did not win the gold medal in the 2010 Winter Olympics, then he necessarily did not *both* win that gold medal *and* injure his ankle, since he definitely did not win the gold medal. If one fact is not the case (the fact presented in sentence A), then both facts cannot be the case, either (the fact presented in sentence A + the new fact added to it in sentence B). Thus if sentence A is true, sentence B is *necessarily* true. Thus, sentence A entails sentence B.

Presupposition: When uttering sentence A, a speaker would not take sentence B for granted (or assume that sentence B is a background against which the truth or falsity of sentence A would be judged). Specifically, by uttering “Shaun White did not win the gold medal in the 2010 Winter Olympics” a speaker could not assume the belief that Shaun White did not both win the gold and injure his ankle, or that Shaun White either won a gold medal or injured his ankle. Whether Shaun White injured his ankle would not be information taken for granted when uttering “Shaun White did not win the gold medal in the 2010 Winter Olympics.” Thus, sentence A does not presuppose sentence B.

NLP

NLP

Introduction to NLP

*Representing and Understanding
Meaning*

Understanding Meaning

- If an agent hears a sentence and can act accordingly, the agent is said to understand it
- Example
 - Leave the book on the table
- Understanding may involve inference
 - Maybe the book is wrapped in paper?
- And pragmatics
 - Which book? Which table?
- So, understanding may involve a procedure

Properties

- **Verifiability**
 - Can a statement be verified against a knowledge base (KB)
 - Example: does my cat Martin have whiskers?
- **Unambiguousness**
 - Give me the book
 - Which book?
- **Canonical form**
- **Expressiveness**
 - Can the formalism express temporal relations, beliefs, ...?
 - Is it domain-independent?
- **Inference**

Representing Meaning

- One traditional approach is to use logic representations, e.g., FOL (first order logic)
- One can then use theorem proving (inference) to determine whether one statement entails another

Syntax of Propositional Logic

- The simplest type of logic
- The proposition symbols P_1, P_2, \dots are sentences
 - If S is a sentence, $\neg S$ is a sentence (negation)
 - If S_1 and S_2 are sentences, $S_1 \wedge S_2$ is a sentence (conjunction)
 - If S_1 and S_2 are sentences, $S_1 \vee S_2$ is a sentence (disjunction)
 - If S_1 and S_2 are sentences, $S_1 \Rightarrow S_2$ is a sentence (implication)
 - If S_1 and S_2 are sentences, $S_1 \Leftrightarrow S_2$ is a sentence (biconditional)

Propositional Logic in Backus Naur Form

- Sentence → AtomicSentence | ComplexSentence
- AtomicSentence → True | False | S | T | U ...
- ComplexSentence → (Sentence)
 - | \neg Sentence
 - | Sentence \wedge Sentence
 - | Sentence \vee Sentence
 - | Sentence \Rightarrow Sentence
 - | Sentence \Leftrightarrow Sentence

Operator Precedence

\neg (highest)

\wedge

\vee

\Rightarrow

\Leftrightarrow (lowest)

Translating Propositions to English

- $A = \text{Today is a holiday.}$
- $B = \text{We are going to the park.}$

- $A \Rightarrow B$
- $A \wedge \neg B$
- $\neg A \Rightarrow B$
- $\neg B \Rightarrow A$
- $B \Rightarrow A$

Translating Propositions to English

- $A = \text{Today is a holiday.}$
- $B = \text{We are going to the park.}$
- $A \Rightarrow B$
 $\text{If today is a holiday, we are going to the park.}$
- $A \wedge \neg B$
 $\text{Today is a holiday and we are not going to the park.}$
- $\neg A \Rightarrow \neg B$
 $\text{If today is not a holiday, then we are not going to the park.}$
- $\neg B \Rightarrow \neg A$
 $\text{If we are not going to the park, then today is not a holiday.}$
- $B \Rightarrow A$
 $\text{If we are going to the park, then today is a holiday.}$

Semantics of Propositional Logic

- $\neg S$ is true iff S is false
- $S_1 \wedge S_2$ is true iff S_1 is true and S_2 is true
- $S_1 \vee S_2$ is true iff S_1 is true or S_2 is true
- $S_1 \Rightarrow S_2$ is true iff S_1 is false or S_2 is true
- i.e., is false iff S_1 is true and S_2 is false
- $S_1 \Leftrightarrow S_2$ is true iff $S_1 \Rightarrow S_2$ is true and $S_2 \Rightarrow S_1$ is true
- Recursively, one can compute the truth value of longer formulas

Connectives

P	Q	$\neg P$	$P \wedge Q$	$P \vee Q$	$P \Rightarrow Q$	$P \Leftrightarrow Q$
F	F	T	F	F	T	T
F	T	T	F	T	T	F
T	F	F	F	T	F	F
T	T	F	T	T	T	T

Logical Equivalence

$(\alpha \wedge \beta) \equiv (\beta \wedge \alpha)$ commutativity of \wedge

$(\alpha \vee \beta) \equiv (\beta \vee \alpha)$ commutativity of \vee

$((\alpha \wedge \beta) \wedge \gamma) \equiv (\alpha \wedge (\beta \wedge \gamma))$ associativity of \wedge

$((\alpha \vee \beta) \vee \gamma) \equiv (\alpha \vee (\beta \vee \gamma))$ associativity of \vee

$\neg(\neg\alpha) \equiv \alpha$ double-negation elimination

$(\alpha \Rightarrow \beta) \equiv (\neg\beta \Rightarrow \neg\alpha)$ contraposition

$(\alpha \Rightarrow \beta) \equiv (\neg\alpha \vee \beta)$ implication elimination

$(\alpha \Leftrightarrow \beta) \equiv ((\alpha \Rightarrow \beta) \wedge (\beta \Rightarrow \alpha))$ biconditional elimination

$\neg(\alpha \wedge \beta) \equiv (\neg\alpha \vee \neg\beta)$ de Morgan

$\neg(\alpha \vee \beta) \equiv (\neg\alpha \wedge \neg\beta)$ de Morgan

$(\alpha \wedge (\beta \vee \gamma)) \equiv ((\alpha \wedge \beta) \vee (\alpha \wedge \gamma))$ distributivity of \wedge over \vee

$(\alpha \vee (\beta \wedge \gamma)) \equiv ((\alpha \vee \beta) \wedge (\alpha \vee \gamma))$ distributivity of \vee over \wedge

NLP

NLP

Introduction to NLP

First Order Logic

Properties of Propositional Logic

- Pros
 - Compositional
 - Declarative
- Cons
 - Limited expressive power
 - Represents facts

First Order Logic

- Used to represent
 - Objects – Martin the cat
 - Relations – Martin and Moses are brothers
 - Functions – Martin's age

First Order Logic

- $\text{Formula} \rightarrow \text{AtomicFormula} \mid \text{Formula Connective Formula}$
 $\mid \text{Quantifier Variable Formula} \mid \neg \text{Formula} \mid$
(Formula)
- $\text{AtomicFormula} \rightarrow \text{Predicate}(\text{Term}\dots)$
- $\text{Term} \rightarrow \text{Function}(\text{Term}\dots) \mid \text{Constant} \mid \text{Variable}$
- $\text{Connective} \rightarrow \wedge \mid \vee \mid \Rightarrow$
- $\text{Quantifier} \rightarrow \forall \mid \exists$
- $\text{Constant} \rightarrow \text{M} \mid \text{Martin}$
- $\text{Variable} \rightarrow x \mid y \mid \dots$
- $\text{Predicate} \rightarrow \text{Likes} \mid \text{Eats} \mid \dots$
- $\text{Function} \rightarrow \text{AgeOf} \mid \text{ColorOf} \mid \dots$

Common Mistake (1)

- \Rightarrow is the main connective with \forall
- Common mistake: using \wedge as the main connective with \forall :
$$\forall x \text{ Cat}(x) \wedge \text{EatsFish}(x)$$

means “Everyone is a cat and everyone eats fish”

Common Mistake (2)

- \wedge is the main connective with \exists
- Common mistake: using \Rightarrow as the main connective with \exists :
$$\exists x \text{ Cat}(x) \Rightarrow \text{EatsFish}(x)$$
is true if there is anyone who is not a cat!

First Order Logic

- NACLO problem from 2014
- Author: Ben King
- <http://www.nacloweb.org/resources/problems/2014/N2014-H.pdf>
- <http://www.nacloweb.org/resources/problems/2014/N2014-HS.pdf>

First Order Logic

(H) Bertrand and Russell (1/3) [10 points]

Teachers can be hard to understand sometimes. Case in point, the math teacher, Mr. Whitehead. Just this morning, he told the class, “It’s not the case that if at least one student studied for the test, then every student failed the test.” What does that even mean?

Well, the two new kids in the class, Bertrand and Russell, have come up with a plan to make sense of Mr. Whitehead’s statements. They call it first-order logic (FOL), a way to map these confusing statements into an unambiguous representation. Bertrand says the whole system is built the idea of propositions, a statement that is either true or false. Propositions can be statements about people or things like `studied_for(John, test)` or `is_hard(test)`. Propositions can also be combined to make more complex statements with the following symbols:

Symbol	Example statement	Interpretation	Explanation
\neg	$\neg \text{studied_for}(\text{John}, \text{test})$	John did <u>not</u> study for the test.	The statement is true if and only if John did not study for the test.
\wedge	$\text{is_hard}(\text{test}) \wedge \text{is_long}(\text{test})$	The test is long <u>and</u> hard.	This statement is true whenever the test is long and the test is hard.
\vee	$\text{is_hard}(\text{test}) \vee \text{is_long}(\text{test})$	The test is long <u>or</u> hard.	This statement is true if the test is long, or if the test is hard, or both.
\Rightarrow	$\text{studied_for}(\text{John}, \text{test}) \Rightarrow \text{aced}(\text{John}, \text{test})$	If John studied for the test, <u>then</u> he aced it.	This is true if the statement on the right side of the arrow is always true whenever the statement on the left side of the arrow is true. If the statement on the left is false, then the whole statement is true by default (if John didn't study, we don't know how he did on the test).

"But," says Russell, "the most important part of first-order logic is the quantifiers." Quantifiers allow you to make general statements like Mr. Whitehead loves to do.

Symbol	Example statement	Interpretation	Explanation
\forall	$[\forall_x : \text{student}(x) \Rightarrow \text{studied_for}(x, \text{test})]$	Every student studied for the test.	The \forall symbol makes a statement about every possible object (whether a student or not). It temporarily gives it the name x to make such a statement. We use the \Rightarrow symbol because we don't want to make any claims about whether non-students studied.
\exists	$[\exists_x : \text{student}(x) \wedge \text{aced}(x, \text{test})]$	There exists at least one student who aced the test.	The \exists symbol makes the claim that there is at least one (possibly more) object in the universe, temporarily called x , that satisfies the statement listed.

Bertrand and Russell also note that there are also a couple other things we can say about individuals (but not propositions or quantifiers). For example, if the names Jonathan and Jon both refer to the same person, we can say *Jon = Jonathan*. If we want to emphasize that John and Jon are different people, we can say *John ≠ Jon*.

H1. Translate Mr. Whitehead's statements into first-order logic by finding the proposition below that is equivalent to each statement and writing the letter of the proposition in the blank. Each statement has exactly one correct answer; not every proposition will be used.

	Everyone either passed or failed the test.
	Every student did not pass the test.
	Exactly one student passed the test.
	A student did not pass the test.
	It is not the case that if at least one student studied for the test, then every student failed the test.

A.	$[\exists_x : \text{student}(x) \wedge \neg \text{passed}(x, \text{test})]$
B.	$[\exists_x : \text{student}(x) \wedge \text{passed}(x, \text{test}) \wedge [\forall_y : \text{passed}(y, \text{test}) \implies x = y]]$
C.	$[\exists_x : \text{student}(x) \wedge \text{passed}(x, \text{test}) \wedge [\exists_y : \text{passed}(y, \text{test}) \wedge x = y]]$
D.	$[\forall_x : \text{passed}(x, \text{test}) \vee \text{failed}(x, \text{test})]$
E.	$\neg ([\exists_x : \text{student}(x) \wedge \text{studied_for}(x, \text{test})] \implies [\forall_x : \text{student}(x) \implies \text{failed}(x, \text{test})])$
F.	$[\exists_x : \text{passed}(x, \text{test}) \wedge \text{failed}(x, \text{test})]$
G.	$[\forall_x : \neg \text{student}(x) \implies \text{passed}(x, \text{test})]$
H.	$[\exists_x : \text{student}(x) \wedge \text{studied_for}(x, \text{test})] \implies \neg [\forall_x : \text{student}(x) \implies \text{failed}(x, \text{test})]$
I.	$\neg [\exists_x : \text{student}(x) \wedge \neg \text{passed}(x, \text{test})]$
J.	$[\forall_x : \text{student}(x) \implies \neg \text{passed}(x, \text{test})]$

H2. Translate first-order logic propositions into their equivalent English sentences by finding the statement below that is equivalent to each proposition and writing the letter of the statement in the blank. Each proposition has exactly one correct answer; not every statement will be used.

	$[\forall_x : \text{student}(x) \Rightarrow \text{studied_for}(x, \text{test})] \vee [\forall_y : \text{student}(y) \Rightarrow \text{passed}(y, \text{test})]$
	$[\forall_x : \text{student}(x) \Rightarrow [\text{studied_for}(x, \text{test}) \vee \text{passed}(x, \text{test})]]$
	$[\forall_x : (\text{test}(x) \wedge \text{long}(x)) \Rightarrow \text{hard}(x)]$
	$[\exists_x : \text{test}(x) \wedge (\text{long}(x) \vee \text{hard}(x))]$
	$[\forall_x : \text{test}(x) \wedge \neg (\text{long}(x) \wedge \text{hard}(x)) \Rightarrow \neg [\forall_y : \text{student}(y) \Rightarrow \text{failed}(y, x)]]$

A.	There is a test that is long or hard.
B.	If a test is not long and not hard, then every student did not fail it.
C.	Every student studied for or passed the test.
D.	Every test that is long is also hard.
E.	Every student studied for the test or every student passed the test.
F.	If there is a test that is hard or not long, then at least one student failed it.
G.	Every test is long and hard.
H.	If a test is not both long and hard, then not every student failed it.

Solutions

1. D
J
B
A
E

2. E
C
D
A
H

Lambda Expressions

- Example
 - $\text{inc}(x) = \lambda x \ x+1$
 - then $\text{inc}(4) = (\lambda x \ x+1)(4) = 5$
- Example
 - $\text{add}(x,y) = \lambda x, \lambda y (x+y)$
 - then $\text{add}(3,4) = (\lambda x, \lambda y (x+y))(3)(4) = (\lambda y \ 3+y)(4) = 3+4 = 12$
- Useful for semantic parsing (see later)

NLP

NLP

Introduction to NLP

Knowledge Representation

Knowledge Representation

- Ontologies
- Categories and objects
- Events
- Times
- Beliefs

Knowledge Representation

- Object
 - Martin the cat
- Categories
 - Cat
- Ontology
 - Mammal includes Cat, Dog, Whale
 - Cat includes PersianCat, ManxCat
- ISA relation
 - ISA (Martin,Cat)
- AKO relation
 - AKO (PersianCat,Cat)
- HASA relation
 - HASA (Cat, tail)

Semantics of FOL

- FOL sentences can be assigned a value of *true* or *false*.

$ISA(Milo, Cat) = true$

- *Milo is younger than Martin*

$<(AgeOf(Milo), AgeOf(Martin)) = true$

$=(AgeOf(Milo), AgeOf(Martin)) = false$

Examples with Quantifiers

- All cats eat fish
$$\forall x : \text{ISA}(x, \text{Cat}) \Rightarrow \text{EatFish}(x)$$

Representing Events

- Martin ate
- Martin ate in the morning
- Martin ate fish
- Martin ate fish in the morning

One Possible Representation

- FOL representations
 - Eating1(Martin)
 - Eating2(Martin,Morning)
 - Eating3(Martin,Fish)
 - Eating4(Martin,Fish,Morning)
- Meaning postulates
 - $\text{Eating4}(x,y,z) \rightarrow \text{Eating3}(x,y)$
 - $\text{Eating4}(x,y,z) \rightarrow \text{Eating2}(x,z)$
 - $\text{Eating4}(x,y,z) \rightarrow \text{Eating1}(x)$

Example from Jurafsky and Martin

Second Possible Representation

- $\text{Eating4}(x,y,z)$
 - With some arguments unspecified
- Problems
 - Too many commitments
 - Hard to combine $\text{Eating4}(\text{Martin}, \text{Fish}, z)$ with $\text{Eating4}(\text{Martin}, y, \text{Morning})$

Third Possible Representation

- Reification
 - $\exists e: \text{ISA}(e, \text{Eating}) \wedge \text{Eater}(e, \text{Martin}) \wedge \text{Eaten}(e, \text{Fish})$

Representing Time

- Example
 - Martin went from the kitchen to the yard
 - ISA(e,Going) \wedge Goer(e,Martin) \wedge Origin (e,kitchen) \wedge Target (e,yard)
- Issue
 - no tense information: past? present? future?
- Fluents
 - A predicate that is true at a given time: $T(f,t)$

Representing Time

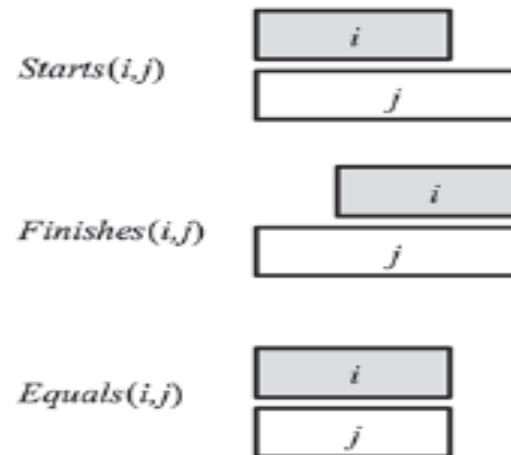
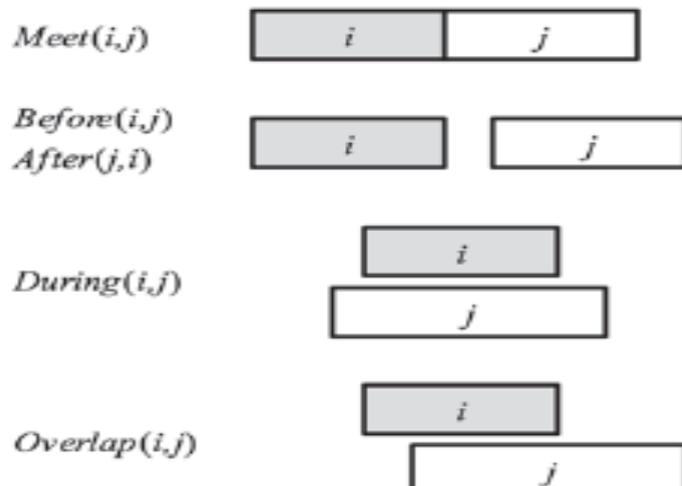
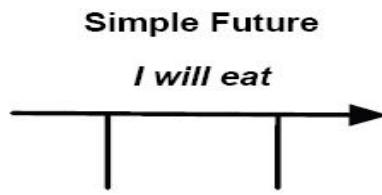
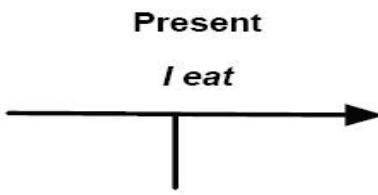
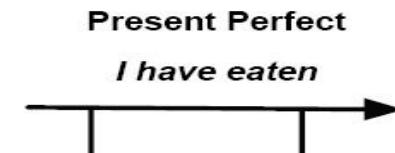
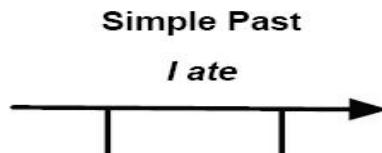


Figure 12.2 Predicates on time intervals.

Example from Russell and Norvig

Representing Time



Example from Jurafsky and Martin

Representing Beliefs

- Example
 - Milo believes that Martin ate fish
- One possible representation
 - $\exists e,b: \text{ISA}(e,\text{Eating}) \wedge \text{Eater}(e,\text{Martin}) \wedge \text{Eaten}(e,\text{Fish}) \wedge \text{ISA}(b,\text{Believing}) \wedge \text{Believer}(b,\text{Milo}) \wedge \text{Believed}(b,e)$
- However this implies (by dropping some of the terms) that “Martin ate fish” (without the belief event)
- Modal logic
 - Possibility, Temporal Logic, Belief Logic

NLP

NLP

Introduction to NLP

Inference

Modus Ponens

- Modus ponens:

 α $\alpha \Rightarrow \beta$

 β

- Example:

Cat(Martin) $\forall x: Cat(x) \Rightarrow EatsFish(x)$ *EatsFish(Martin)*

Inference

- Forward chaining
 - as individual facts are added to the database, all derived inferences are generated
- Backward chaining
 - starts from queries
 - Example: the Prolog programming language
- Prolog example
 - ```
father(X, Y) :- parent(X, Y), male(X).
parent(john, bill).
parent(jane, bill).
female(jane).
male(john).
?- father(M, bill).
```

# Examples

The kinship domain:

- Brothers are siblings

$$\forall x,y \ Brother(x,y) \Rightarrow Sibling(x,y)$$

- One's mother is one's female parent

$$\forall m,c \ Mother(c) = m \Leftrightarrow (Female(m) \wedge Parent(m,c))$$

- “Sibling” is symmetric

$$\forall x,y \ Sibling(x,y) \Leftrightarrow Sibling(y,x)$$

# Universal Instantiation

- Every instantiation of a universally quantified sentence is entailed by it:

$$\forall v \alpha$$

---

$$\text{Subst}(\{v/g\}, \alpha)$$

for any variable  $v$  and ground term  $g$

- E.g.,  $\forall x \text{ Cat}(x) \wedge \text{Fish}(y) \Rightarrow \text{Eats}(x,y)$  yields:  
 $\text{Cat}(\text{Martin}) \wedge \text{Fish}(\text{Blub})$

# Existential Instantiation

- For any sentence  $\alpha$ , variable  $v$ , and constant symbol  $k$  that does not appear elsewhere in the knowledge base:

$$\frac{\exists v \alpha}{\text{Subst}(\{v/k\}, \alpha)}$$

- E.g.,  $\exists x \text{Cat}(x) \wedge \text{EatsFish}(x)$  yields:

$$\text{Cat}(C_1) \wedge \text{EatsFish}(C_1)$$

provided  $C_1$  is a new constant symbol, called a Skolem constant

# Unification

- If a substitution  $\theta$  is available, unification is possible
- Examples:
  - $p = \text{Eats}(x,y)$ ,  $q = \text{Eats}(x,\text{Blub})$ , possible if  $\theta = \{y/\text{Blub}\}$
  - $p = \text{Eats}(\text{Martin},y)$ ,  $q = \text{Eats}(x,\text{Blub})$ , possible if  $\theta = \{x,\text{Martin},y/\text{Blub}\}$
  - $p = \text{Eats}(\text{Martin},y)$ ,  $q = \text{Eats}(y,\text{Blub})$ , fails because  $\text{Martin} \neq \text{Blub}$
- Subsumption
  - Unification works not only when two things are the same but also when one of them subsumes the other one
  - Example: All cats eat fish, Martin is a cat, Blub is a fish

NLP

NLP

# Introduction to NLP

*Semantic Parsing*

# Semantic Parsing

- Converting natural language to a logical form
  - e.g., executable code for a specific application
- Example:
  - Airline reservations
  - Geographical query systems

# Stages of Semantic Parsing

- Input
  - Sentence
- Syntactic Analysis
  - Syntactic structure
- Semantic Analysis
  - Semantic representation

# Compositional Semantics

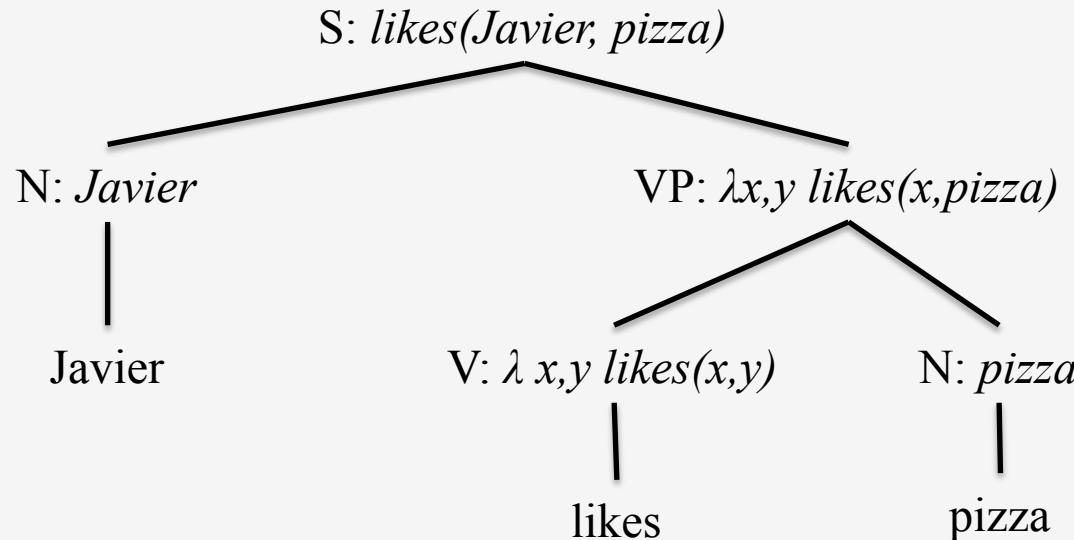
- Add semantic attachments to CFG rules
- Compositional semantics
  - Parse the sentence syntactically
  - Associate some semantics to each word
  - Combine the semantics of words and non-terminals recursively
  - Until the root of the sentence

# Example

- Input
  - Javier likes pizza
- Output
  - *like(Javier, pizza)*

# Semantic Parsing

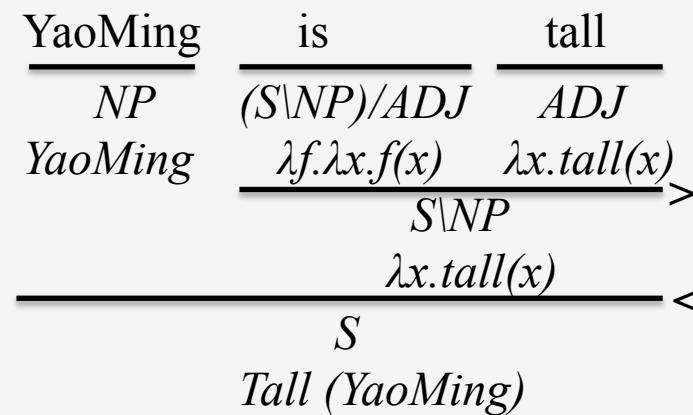
- Associate a semantic expression with each node



# Using CCG (Steedman 1996)

- CCG representations for semantics

- $ADJ: \lambda x. tall(x)$
  - $(S \setminus NP)/ADJ : \lambda f. \lambda x. f(x)$
  - $NP: YaoMing$



# CCG

- NACLO problem from 2014
- Authors: Jonathan Kummerfeld, Aleka Blackwell, and Patrick Littell
- <http://www.nacloweb.org/resources/problems/2014/N2014-O.pdf>
- <http://www.nacloweb.org/resources/problems/2014/N2014-OS.pdf>
- <http://www.nacloweb.org/resources/problems/2014/N2014-P.pdf>
- <http://www.nacloweb.org/resources/problems/2014/N2014-PS.pdf>

# CCG

One way for computers to understand language is by forming a structure that represents the relationships between words using a technique called Combinatorial Categorial Grammar (CCG). Computer scientists and linguists can use CCG to parse sentences (that is, try to figure out their structure) and then extract meaning from the structure.

As the name suggests, Combinatorial Categorial Grammar parses sentences by combining categories. Each word in a sentence is assigned a particular category; note that / and \ are two different symbols:

|       |               |
|-------|---------------|
| I     | NP            |
| books | NP            |
| sleep | S \ NP        |
| enjoy | (S \ NP) / NP |

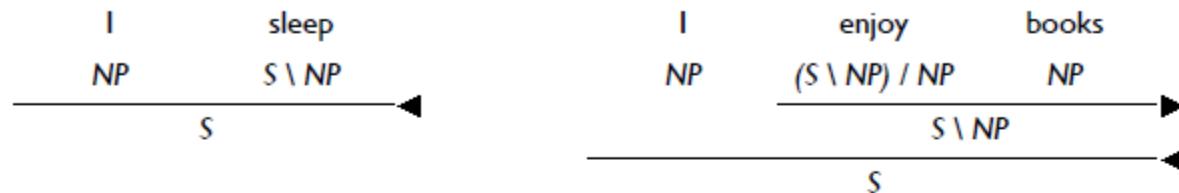
# CCG

One way for computers to understand language is by forming a structure that represents the relationships between words using a technique called Combinatorial Categorial Grammar (CCG). Computer scientists and linguists can use CCG to parse sentences (that is, try to figure out their structure) and then extract meaning from the structure.

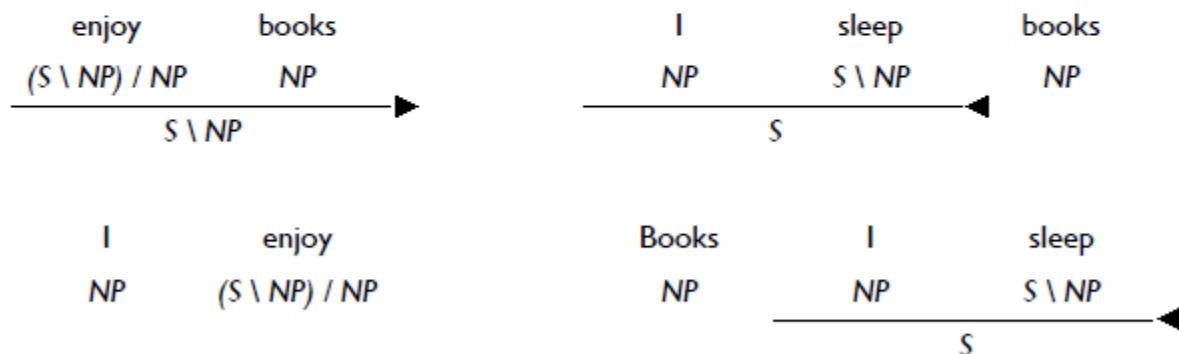
As the name suggests, Combinatorial Categorial Grammar parses sentences by combining categories. Each word in a sentence is assigned a particular category; note that / and \ are two different symbols:

|       |               |
|-------|---------------|
| I     | NP            |
| books | NP            |
| sleep | S \ NP        |
| enjoy | (S \ NP) / NP |

These categories are then combined in systematic ways. We will not explain how, but we will give you two successful parses...



...and four unsuccessful parses...



If a parse is successful, the sentence is declared “grammatical”; if not, the sentence is declared “ungrammatical”.

# CCG

- O1. Using the above examples as evidence, figure out how CCG parses sentences, and describe it briefly here:
- O2. In the sentence “I enjoy long books”, list all of the categories that, if assigned to “long”, make the sentence have a successful parse.
- O3. Not every grammatical sentence of English will be declared “grammatical” by the process above. Using only the words “I”, “books”, “sleep”, and “enjoy”, form a grammatically correct English sentence that will fail to parse given the categories above. You don’t have to use all four of the words.

# Answer

O1. CCG assigns a category to each word and constructs a parse by combining pairs of categories to form an S. Not all pairs of categories can combine. A pair is allowed to combine if one category (e.g. A) is contained within the category next to it (e.g. B / A) and lies on the side indicated by the slash (\ for left, / for right). When two categories combine, the result is a new category, taken from the left of the slash (B in this example).

O2. There are four categories that ‘long’ could have that would create a successful parse of ‘I enjoy long books’:

1. NP / NP
2. (( S \ NP ) \ (( S \ NP ) / NP )) / NP
3. (( S \ NP ) / NP) \ (( S \ NP ) / NP)
4. (( S / NP ) \ NP) \ (( S \ NP ) / NP)

The first of these is probably the most appropriate. Some possible reasons:

- It is by far the simplest. (After all, all our other categories are relatively simple.)
- It keeps the existing structure of the sentence (where “enjoy” combines with what follows it and then with what precedes it).
- “Long” describes “books” and not “enjoy”, so it might make sense to keep them together.
- The first would be the only one to work if “long books” were in any other position.

O3. Possible answers: “I enjoy sleep”, topicalized object sentences like “Books I enjoy” and “Sleep I enjoy”.

This problem is a follow-up to problem O and has to be solved after that problem. Tok Pisin (also referred to as New Guinea Pidgin or Melanesian Pidgin) is a creole language spoken in the northern mainland of Papua New Guinea and surrounding islands. It is an official language and the mostly widely used language in the country, spoken by over 5 million people.

Many Tok Pisin words come originally from English – its name comes from “talk” and “pidgin”! -- but Tok Pisin isn’t just English. It has a distinct grammar and uses these words in different (but systematic!) ways.

P1. Below are sentences in Tok Pisin with a scrambled list of English translations. Match each sentence to its English equivalent.

|    |                               |  |
|----|-------------------------------|--|
| I. | Brata bilong em i stap rit.   |  |
| 2. | Ol i stap dringim wara.       |  |
| 3. | Ol i ken ritim buk bilong mi. |  |
| 4. | Em i ritim buk pinis.         |  |
| 5. | Em i laik rit.                |  |
| 6. | Susa bilong em i ken rait.    |  |
| 7. | Susa bilong mi i boilim wara. |  |
| 8. | Wara i boil pinis.            |  |

|    |                            |
|----|----------------------------|
| A. | He has read the book.      |
| B. | My sister boils the water. |
| C. | They can read my book.     |
| D. | His sister can write.      |
| E. | His brother is reading.    |
| F. | The water has boiled.      |
| G. | He wants to read.          |
| H. | They are drinking water.   |

# CCG

P2. Translate the following Tok Pisin sentence into English:

Brata bilong mi i stap ritim buk bilong susa bilong mi.

---

P3. Translate the following English sentence into Tok Pisin:

Their sister wants to write a book.

---

# Top Bar Reserved for U-M Branding and Course Information

P4. Describing these words in terms of their CCG categories (introduced in Problem O) highlights that these aren't English words combined according to English rules, but are Tok Pisin words combined according to Tok Pisin rules.

Match each Tok Pisin word to its CCG category. Some categories will be used more than once. The symbol  $S_b$  is short for 'Bare Clause'.

|     |         |  |
|-----|---------|--|
| 1.  | bilong  |  |
| 2.  | brata   |  |
| 3.  | boil    |  |
| 4.  | boilim  |  |
| 5.  | buk     |  |
| 6.  | dringim |  |
| 7.  | em      |  |
| 8.  | i       |  |
| 9.  | ken     |  |
| 10. | laik    |  |

|     |        |  |
|-----|--------|--|
| 11. | mi     |  |
| 12. | ol     |  |
| 13. | pinis  |  |
| 14. | stap   |  |
| 15. | raitim |  |
| 16. | rit    |  |
| 17. | ritim  |  |
| 18. | susa   |  |
| 19. | wara   |  |

|    |                                               |
|----|-----------------------------------------------|
| A. | NP                                            |
| B. | (NP \ NP) / NP                                |
| C. | (\\$ \ NP) / (S <sub>b</sub> \ NP)            |
| D. | (S <sub>b</sub> \ NP)                         |
| E. | (S <sub>b</sub> \ NP) / NP                    |
| F. | (S <sub>b</sub> \ NP) \ (S <sub>b</sub> \ NP) |
| G. | (S <sub>b</sub> \ NP) / (S <sub>b</sub> \ NP) |

P5. Explain your answer.

# CCG

P1.

|    |                                      |   |
|----|--------------------------------------|---|
| 1. | <i>Brata bilong em i stap rit.</i>   | E |
| 2. | <i>Ol i stap dringim wara.</i>       | H |
| 3. | <i>Ol i ken ritim buk bilong mi.</i> | C |
| 4. | <i>Em i ritim buk pinis.</i>         | A |
| 5. | <i>Em i laik rit.</i>                | G |
| 6. | <i>Susa bilong em i ken rait.</i>    | D |
| 7. | <i>Susa bilong mi i boilim wara.</i> | B |
| 8. | <i>Wara i boil pinis.</i>            | F |

P2. My brother is reading my sister's book.

P3. Susa bilong ol i laik raitim buk.

|    |                            |
|----|----------------------------|
| A. | He has read the book.      |
| B. | My sister boils the water. |
| C. | They can read my book.     |
| D. | His sister can write.      |
| E. | His brother is reading.    |
| F. | The water has boiled.      |
| G. | He wants to read.          |
| H. | They are drinking water.   |

# CCG

P4.

|     |         |   |
|-----|---------|---|
| 1.  | bilong  | B |
| 2.  | brata   | A |
| 3.  | boil    | D |
| 4.  | boilim  | E |
| 5.  | buk     | A |
| 6.  | dringim | E |
| 7.  | em      | A |
| 8.  | i       | C |
| 9.  | ken     | G |
| 10. | laik    | G |

|     |        |   |
|-----|--------|---|
| 11. | mi     | A |
| 12. | ol     | A |
| 13. | pinis  | F |
| 14. | stap   | G |
| 15. | raitim | E |
| 16. | rit    | D |
| 17. | ritim  | E |
| 18. | susa   | A |
| 19. | wara   | A |

|    |                                               |
|----|-----------------------------------------------|
| A. | NP                                            |
| B. | (NP \ NP) / NP                                |
| C. | (S \ NP) / (S <sub>b</sub> \ NP)              |
| D. | (S <sub>b</sub> \ NP)                         |
| E. | (S <sub>b</sub> \ NP) / NP                    |
| F. | (S <sub>b</sub> \ NP) \ (S <sub>b</sub> \ NP) |
| G. | (S <sub>b</sub> \ NP) / (S <sub>b</sub> \ NP) |

# CCG

- P5.
- A. Any noun or pronoun is category A (NP) because they can be used as a noun.
  - B. The word "bilong" shows possession of the preceding NP by the following NP; therefore, it is (NP\NP)/NP. Also, the phrase [NP bilong NP] yields a noun phrase (NP).
  - C. The word "i" is necessary for a grammatical sentence, so it is (S\NP)/(S<sub>b</sub>\NP). It wants a following verb phrase (indicated by (S<sub>b</sub>\NP)) and a preceding noun phrase (NP). NP+i+(S<sub>b</sub>\NP) forms a sentence.
  - D. Each intransitive verb (boil and rit) can stand on its own as S<sub>b</sub>\NP, forming the verb phrase.
  - E. Transitive verbs (boilim, dringim, raitim, ritim; the ones ending in -im), need a following NP, so they are categorized as (S<sub>b</sub>\NP)/NP, a verb phrase followed by a noun phrase.
  - F. The verbs "stap," "ken," and "laik" precede the primary verb phrase and need another verb phrase to create an S<sub>b</sub>\NP, so they are the category (S<sub>b</sub>\NP)/(S<sub>b</sub>\NP).
  - G. The verb "pinis" comes after the main verb, so it is of the category (S<sub>b</sub>\NP)\(S<sub>b</sub>\NP) which requires a (S<sub>b</sub>\NP) to precede it.

# GeoQuery (Zelle and Mooney 1996)

What is the capital of the state with the largest population?  
 answer(C, (capital(S,C), largest(P, (state(S),  
 population(S,P))))).

What are the major cities in Kansas?  
 answer(C, (major(C), city(C), loc(C,S),  
 equal(S, stateid(kansas)))).

| Form            | Predicate                            |
|-----------------|--------------------------------------|
| capital(C)      | C is a capital (city).               |
| city(C)         | C is a city.                         |
| major(X)        | X is major.                          |
| place(P)        | P is a place.                        |
| river(R)        | R is a river.                        |
| state(S)        | S is a state.                        |
| capital(C)      | C is a capital (city).               |
| area(S,A)       | The area of S is A.                  |
| capital(S,C)    | The capital of S is C.               |
| equal(V,C)      | variable V is ground term C.         |
| density(S,D)    | The (population) density of S is D.  |
| elevation(P,E)  | The elevation of P is E.             |
| high_point(S,P) | The highest point of S is P.         |
| higher(P1,P2)   | P1's elevation is greater than P2's. |
| loc(X,Y)        | X is located in Y.                   |
| low_point(S,P)  | The lowest point of S is P.          |
| len(R,L)        | The length of R is L.                |
| next_to(S1,S2)  | S1 is next to S2.                    |
| size(X,Y)       | The size of X is Y.                  |
| traverse(R,S)   | R traverses S.                       |

| Type    | Form                | Example           |
|---------|---------------------|-------------------|
| country | countryid(Name)     | countryid(usa)    |
| city    | cityid(Name, State) | cityid(austin,tx) |
| state   | stateid(Name)       | stateid(texas)    |
| river   | riverid(Name)       | riverid(colorado) |
| place   | placeid(Name)       | placeid(pacific)  |

# Zettlemoyer and Collins (2005)

Utah :=  $NP$   
 Idaho :=  $NP$   
 borders :=  $(S \setminus NP)/NP$

a) What states border Texas  
 $\lambda x.state(x) \wedge borders(x, \text{texas})$

b) What is the largest state  
 $\arg \max(\lambda x.state(x), \lambda x.size(x))$

c) What states border the state that borders the most states  
 $\lambda x.state(x) \wedge borders(x, \arg \max(\lambda y.state(y), \lambda y.\text{count}(\lambda z.state(z) \wedge borders(y, z))))$

a)

$$\begin{array}{c}
 \frac{\begin{array}{ccc} \text{Utah} & \text{borders} & \text{Idaho} \\ NP & \frac{(S \setminus NP)/NP}{\lambda x. \lambda y. borders(y, x)} & NP \\ utah & \xrightarrow{\lambda f. \lambda g. \lambda x. f(x) \wedge g(x)} & idaho \end{array}}{\xrightarrow{\lambda y. borders(y, idaho)}}
 \\ \xleftarrow{S} \\
 borders(utah, idaho)
 \end{array}$$

b)

$$\begin{array}{ccccc}
 \frac{}{(S/(S \setminus NP))/N} & \frac{N}{\lambda x. state(x)} & \frac{(S \setminus NP)/NP}{\lambda x. \lambda y. borders(y, x)} & \frac{NP}{\lambda y. borders(y, \text{texas})} & \frac{}{Texas} \\
 \lambda f. \lambda g. \lambda x. f(x) \wedge g(x) & \xrightarrow{\lambda x. state(x)} & \lambda x. \lambda y. borders(y, x) & \xrightarrow{\lambda y. borders(y, \text{texas})} & \text{texas} \\
 \frac{S/(S \setminus NP)}{\lambda g. \lambda x. state(x) \wedge g(x)} & & \frac{(S \setminus NP)}{\lambda y. borders(y, \text{texas})} & & \\
 & & \xrightarrow{S} & & \\
 & & \lambda x. state(x) \wedge borders(x, \text{texas}) & &
 \end{array}$$

Utah :=  $NP : utah$   
 Idaho :=  $NP : idaho$   
 borders :=  $(S \setminus NP)/NP : \lambda x. \lambda y. borders(y, x)$

# Zettlemoyer and Collins (2005)

|             |                                                                 |
|-------------|-----------------------------------------------------------------|
| states      | $:= N : \lambda x. state(x)$                                    |
| major       | $:= N/N : \lambda f. \lambda x. major(x) \wedge f(x)$           |
| population  | $:= N : \lambda x. population(x)$                               |
| cities      | $:= N : \lambda x. city(x)$                                     |
| rivers      | $:= N : \lambda x. river(x)$                                    |
| run through | $:= (S \setminus NP)/NP : \lambda x. \lambda y. traverse(y, x)$ |
| the largest | $:= NP/N : \lambda f. \text{arg max}(f, \lambda x. size(x))$    |
| river       | $:= N : \lambda x. river(x)$                                    |
| the highest | $:= NP/N : \lambda f. \text{arg max}(f, \lambda x. elev(x))$    |
| the longest | $:= NP/N : \lambda f. \text{arg max}(f, \lambda x. len(x))$     |

Figure 6: Ten learned lexical items that had highest associated parameter values from a randomly chosen development run in the Geo880 domain.

|             |                                                                 |
|-------------|-----------------------------------------------------------------|
| states      | $:= N : \lambda x. state(x)$                                    |
| major       | $:= N/N : \lambda f. \lambda x. major(x) \wedge f(x)$           |
| population  | $:= N : \lambda x. population(x)$                               |
| cities      | $:= N : \lambda x. city(x)$                                     |
| rivers      | $:= N : \lambda x. river(x)$                                    |
| run through | $:= (S \setminus NP)/NP : \lambda x. \lambda y. traverse(y, x)$ |
| the largest | $:= NP/N : \lambda f. \text{arg max}(f, \lambda x. size(x))$    |
| river       | $:= N : \lambda x. river(x)$                                    |
| the highest | $:= NP/N : \lambda f. \text{arg max}(f, \lambda x. elev(x))$    |
| the longest | $:= NP/N : \lambda f. \text{arg max}(f, \lambda x. len(x))$     |

NLP

NLP

# Introduction to NLP

*Discourse Analysis*

# Issues with Discourse

- Anaphora
  - I went to see my grandfather at the hospital. *The old man* has been there for weeks. *He* had surgery a few days ago.
- Referring expressions and antecedents
- Issues with both single sentences and multi-sentential text
- Needed
  - Models of discourse

# Coreference

- Sample use of anaphora
  - John saw Mary in the park. As every morning, she was walking her dog.
- What does “she” refer to?
- Candidate referents
  - John
  - Mary
  - The park
  - Every morning
  - Her dog

# MUC-7 Coreference Task

<COREF ID="6" TYPE="IDENT" REF="5" MIN="Aeroflot">The Russian airline Aeroflot</COREF> has been hit with <COREF ID="19">a writ</COREF> for loss and damages, filed in <COREF ID="15">Hong Kong</COREF> by <COREF ID="7" TYPE="IDENT" REF="4" MIN="families">the families of <COREF ID="22" MIN="passengers">seven passengers killed in <COREF ID="8" TYPE="IDENT" REF="9" MIN="crash">an air crash</COREF></COREF></COREF>.

All 75 people on board <COREF ID="12" MIN="Airbus">the <COREF ID="10" TYPE="IDENT" REF="6">Aeroflot</COREF> <COREF ID="25">Airbus</COREF> </COREF> died when <COREF ID="11" TYPE="IDENT" REF="12">it</COREF> ploughed into a Siberian mountain in March 1994.

# MUC-7 Coreference Task

<COREF ID="6" TYPE="IDENT" REF="5" MIN="Aeroflot">The Russian airline Aeroflot</COREF> has been hit with <COREF ID="19">a writ</COREF> for loss and damages, filed in <COREF ID="15">Hong Kong</COREF> by <COREF ID="7" TYPE="IDENT" REF="4" MIN="families">the families of <COREF ID="22" MIN="passengers">seven passengers killed in <COREF ID="8" TYPE="IDENT" REF="9" MIN="crash">an air crash</COREF></COREF></COREF>.

All 75 people on board <COREF ID="12" MIN="Airbus">the <COREF ID="10" TYPE="IDENT" REF="6">Aeroflot</COREF> <COREF ID="25">Airbus</COREF> </COREF> died when <COREF ID="11" TYPE="IDENT" REF="12">it</COREF> ploughed into a Siberian mountain in March 1994.

# Screwdriver on Wikipedia

- A screwdriver is a tool, manual or powered, for turning (driving or removing) screws. A typical simple screwdriver has a handle and a shaft, and a tip that the user inserts into the screw head to turn it. The shaft is usually made of tough steel to resist bending or twisting. The tip may be hardened to resist wear, treated with a dark tip coating for improved visual contrast between tip and screw—or ridged or treated for additional 'grip'. Handles are typically wood, metal, or plastic and usually hexagonal, square, or oval in cross-section to improve grip and prevent the tool from rolling when set down. Some manual screwdrivers have interchangeable tips that fit into a socket on the end of the shaft and are held in mechanically or magnetically. These often have a hollow handle that contains various types and sizes of tips, and a reversible ratchet action that allows multiple full turns without repositioning the tip or the user's hand.

# Screwdriver on Wikipedia

- A screwdriver is a tool, manual or powered, for turning (driving or removing) screws. A typical simple screwdriver has a handle and a shaft, and a tip that the user inserts into the screw head to turn it. **The shaft** is usually made of tough steel to resist bending or twisting. **The tip** may be hardened to resist wear, treated with a dark tip coating for improved visual contrast between tip and screw—or ridged or treated for additional 'grip'. **Handles** are typically wood, metal, or plastic and usually hexagonal, square, or oval in cross-section to improve grip and prevent the tool from rolling when set down. Some manual screwdrivers have interchangeable tips that fit into a socket on the end of the shaft and are held in mechanically or magnetically. **These** often have a hollow handle that contains various types and sizes of tips, and a reversible ratchet action that allows multiple full turns without repositioning the tip or the user's hand.

# Coreference Resolution

- Agreement constraints
  - gender, number, animacy
- Syntactic constraints
  - e.g., parallelism
- Sentence ordering
  - recency

# Saliency Weights

|                                       |     |
|---------------------------------------|-----|
| Sentence recency                      | 100 |
| Subject emphasis                      | 80  |
| Existential emphasis (“there is ...”) | 70  |
| Accusative (direct object) emphasis   | 50  |
| Indirect object emphasis              | 40  |
| Non-adverbial emphasis                | 50  |
| Head noun emphasis                    | 80  |

[Lappin and Leass 1994]

# Lappin and Leass (cont'd)

- Recency handling
  - weights are cut in half after each sentence is processed.
- Examples:
  - An Acura Integra is parked in the lot. (subject)
  - There is an Acura Integra parked in the lot. (existential predicate nominal)
  - John parked an Acura Integra in the lot. (object)
  - John gave Susan an Acura Integra. (indirect object)
  - In his Acura Integra, John showed Susan his new CD player. (demarcated adverbial PP)

# Resolution of Anaphora Procedure (RAP)

- Collect the potential referents (up to four sentences back).
- Remove potential referents that do not agree in number or gender with the pronoun.
- Remove potential referents that do not pass intrasentential syntactic coreference constraints.
- Compute the total salience value of the referent by adding any applicable values for role parallelism (+35) or cataphora (-175).
- Select the referent with the highest salience value. In case of a tie, select the closest referent in terms of string position.
- When moving to a new sentence, halve all scores for the existing entities on the list.

# Example

- John saw a beautiful Acura Integra at the dealership last week. He showed it to Bill. He bought it.

|            | Rec | Subj | Exist | Obj | Ind Obj | Non Adv | Head N | Total |
|------------|-----|------|-------|-----|---------|---------|--------|-------|
| John       | 100 | 80   |       |     |         | 50      | 80     | 310   |
| Integra    | 100 |      |       | 50  |         | 50      | 80     | 280   |
| dealership | 100 |      |       |     |         | 50      | 80     | 230   |

# Example (cont'd)

| Referent   | Phrases                     | Value |
|------------|-----------------------------|-------|
| John       | {John}                      | 155   |
| Integra    | {a beautiful Acura Integra} | 140   |
| dealership | {the dealership}            | 115   |

# Example (cont'd)

| Referent   | Phrases                     | Value |
|------------|-----------------------------|-------|
| John       | {John, he <sup>1</sup> }    | 465   |
| Integra    | {a beautiful Acura Integra} | 140   |
| dealership | {the dealership}            | 115   |

# Example (cont'd)

| Referent   | Phrases                         | Value |
|------------|---------------------------------|-------|
| John       | {John, he <sup>1</sup> }        | 465   |
| Integra    | {a beautiful Acura Integra, it} | 420   |
| dealership | {the dealership}                | 115   |

# Example (cont'd)

| Referent   | Phrases                         | Value |
|------------|---------------------------------|-------|
| John       | {John, he <sup>1</sup> }        | 465   |
| Integra    | {a beautiful Acura Integra, it} | 420   |
| Bill       | {Bill}                          | 270   |
| dealership | {the dealership}                | 115   |

# Example (cont'd)

| Referent   | Phrases                                       | Value |
|------------|-----------------------------------------------|-------|
| John       | {John, he <sup>1</sup> }                      | 232.5 |
| Integra    | {a beautiful Acura Integra, it <sup>1</sup> } | 210   |
| Bill       | {Bill}                                        | 135   |
| dealership | {the dealership}                              | 57.5  |

NLP

NLP

# Discourse Analysis

Coherence

# Coherence

- Examples

I saw Mary in the street. She was looking for a bookstore.

? I saw Mary in the street. She has a cat.

?? I saw Mary in the street. The Pistons won.

- Rhetorical Structure Theory (Mann and Thompson 1988)

# Nucleus and Satellite



The carpenter was tired.

He had been working all day.

# Nucleus and Satellite

- The satellite increases the belief in the relation described in the nucleus
- Some relations have only a nucleus, others have two nuclei, yet others have one nucleus and one satellite

# Coherence Relations

- **Result**
  - The carpenter worked all day. The new cabinet was ready in the evening.
- **Explanation**
  - The carpenter was tired. He had spent the entire day building a new cabinet.
- **Parallel**
  - The carpenter worked all day. The upholsterer took the day off.
- **Elaboration**
  - The carpenter built a cabinet. The cabinet had four drawers and an oversized rear panel.
- **Other relations**
  - Nucleus+satellite: circumstance, volitional cause, purpose, interpretation, restatement, summary
  - Multi-nuclear: sequence, contrast, joint

[Mann and Thompson 1988]

# Sample Rhetorical Relations

| Relation    | Nucleus                                       | Satellite                                                              |
|-------------|-----------------------------------------------|------------------------------------------------------------------------|
| Antithesis  | ideas favored by the author                   | ideas disfavored by the author                                         |
| Background  | text whose understanding is being facilitated | text for facilitating understanding                                    |
| Concession  | situation affirmed by author                  | situation which is apparently inconsistent but also affirmed by author |
| Elaboration | basic information                             | additional information                                                 |
| Purpose     | an intended situation                         | the intent behind the situation                                        |
| Restatement | a situation                                   | a reexpression of the situation                                        |
| Summary     | text                                          | a short summary of that text                                           |

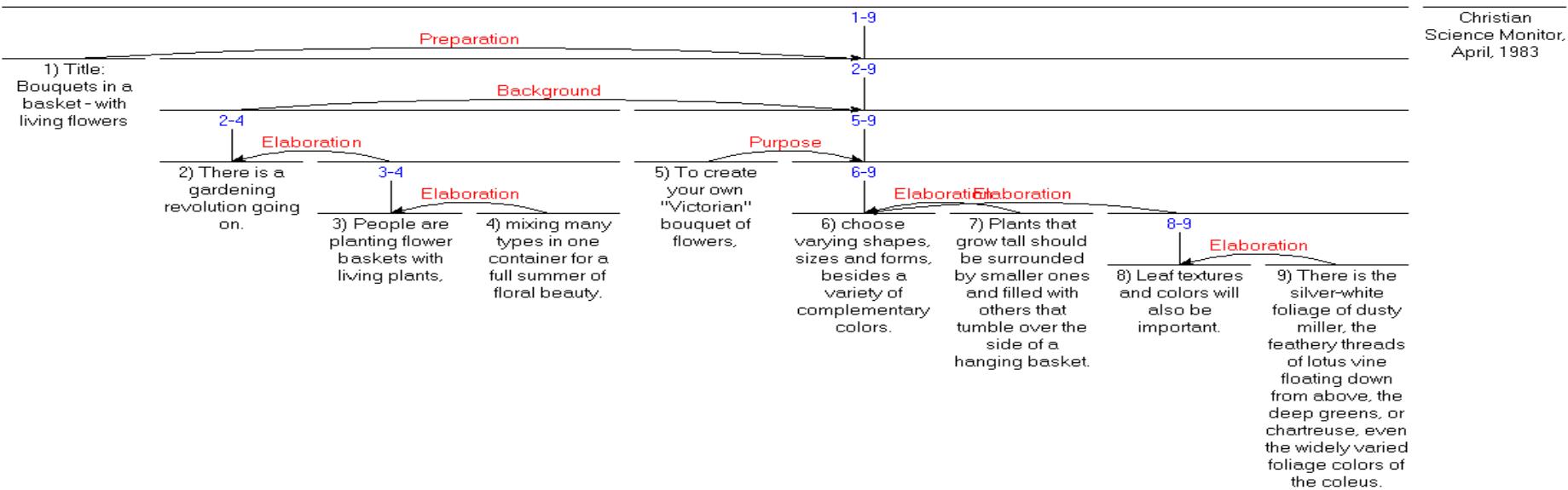
# Example

- 1) Title: Bouquets in a basket – with living flowers
- 2) There is a gardening revolution going on.
- 3) People are planting flower baskets with living plants,
- 4) mixing many types in one container for a full summer of floral beauty.
- 5) To create your own "Victorian" bouquet of flowers,
- 6) choose varying shapes, sizes and forms, besides a variety of complementary colors.
- 7) Plants that grow tall should be surrounded by smaller ones and filled with others that tumble over the side of a hanging basket.
- 8) Leaf textures and colors will also be important.
- 9) There is the silver-white foliage of dusty miller, the feathery threads of lotus vine floating down from above, the deep greens, or chartreuse, even the widely varied foliage colors of the coleus.

Christian Science Monitor, April, 1983

from Mann/Matthiessen/Thompson

# Example (cont'd)



# Discourse Parsing

- Four RST relations: contrast, cause-explanation-evidence, condition, elaboration + non-relation
- Up to 4M automatically labeled examples per relation
- Naïve Bayes
- Word co-occurrence features

[Marcu and Echihabi 2002]

# Centering

- Goal: understand the local coherence of discourse
- Why some texts are considered more coherent
- Inference load associated with badly chosen referring expressions
- Too much focus shift makes the text hard to understand.

# Centering

- Every utterance  $U_n$  has a backwards looking center  $C_b$ , which connects  $U_n$  with the previous utterance  $U_{n-1}$ .
- Every utterance also has a partially ordered set of forward looking centers  $C_f$  related to the next utterance  $U_{n+1}$ . The order depends on syntax (e.g., subject>object)
- The preferred center  $C_p$  is the highest ranking element of  $C_f$ .

# Cross-document Structure (CST)

| Number | Relationship type          | Level | Description                                                                 |
|--------|----------------------------|-------|-----------------------------------------------------------------------------|
| 1      | Identity                   | Any   | The same text appears in more than one location                             |
| 2      | Equivalence (paraphrasing) | S, D  | Two text spans have the same information content                            |
| 3      | Translation                | P, S  | Same information content in different languages                             |
| 4      | Subsumption                | S, D  | One sentence contains more information than another                         |
| 5      | Contradiction              | S, D  | Conflicting information                                                     |
| 6      | Historical background      | S     | Information that puts current information in context                        |
| 7      | Cross-reference            | P     | The same entity is mentioned                                                |
| 8      | Citation                   | S, D  | One sentence cites another document                                         |
| 9      | Modality                   | S     | Qualified version of a sentence                                             |
| 10     | Attribution                | S     | One sentence repeats the information of another while adding an attribution |
| 11     | Summary                    | S, D  | Similar to Summary in RST: one sentence summarizes another                  |

S=Sentence, P=Paragraph, D=document

# Cross-document Structure (CST)

| Number | Relationship type     | Level | Description                                                                               |
|--------|-----------------------|-------|-------------------------------------------------------------------------------------------|
| 12     | Follow-up             | S     | Additional information which reflects facts that have happened since the previous account |
| 13     | Elaboration           | S     | Additional information that wasn't included in the last account                           |
| 14     | Indirect speech       | S     | Shift from direct to indirect speech or vice-versa                                        |
| 15     | Refinement            | S     | Additional information that is                                                            |
| 16     | Agreement             | S     | One source expresses agreement with another                                               |
| 17     | Judgement             | S     | A qualified account of a fact                                                             |
| 18     | Fulfilment            | S     | A prediction turned true                                                                  |
| 19     | Description           | S     | Insertion of a description                                                                |
| 20     | Reader profile        | S     | Style and background-specific change                                                      |
| 21     | Contrast              | S     | Contrasting two accounts of facts                                                         |
| 22     | Parallel              | S     | Comparing two accounts of facts                                                           |
| 23     | Generalization        | S     | Generalization                                                                            |
| 24     | Change of perspective | S,D   | The same source presents a fact in a different light                                      |

# Argumentative Zoning

- **Aim**
  - research goal of the paper
- **Textual**
  - statements about section structure
- **Own**
  - description of the authors' work (methodology, results, discussion)
- **Background**
  - generally accepted scientific background
- **Contrast**
  - comparison with other work
- **Basis**
  - statements of agreement with other work
- **Other**
  - description of other researchers' work

# Local Entity Coherence

**Table 2**

Summary augmented with syntactic annotations for grid computation.

- 1 [The Justice Department]<sub>s</sub> is conducting an [anti-trust trial]<sub>o</sub> against [Microsoft Corp.]<sub>x</sub> with [evidence]<sub>x</sub> that [the company]<sub>s</sub> is increasingly attempting to crush [competitors]<sub>o</sub>.
- 2 [Microsoft]<sub>o</sub> is accused of trying to forcefully buy into [markets]<sub>x</sub> where [its own products]<sub>s</sub> are not competitive enough to unseat [established brands]<sub>o</sub>.
- 3 [The case]<sub>s</sub> revolves around [evidence]<sub>o</sub> of [Microsoft]<sub>s</sub> aggressively pressuring [Netscape]<sub>o</sub> into merging [browser software]<sub>o</sub>.
- 4 [Microsoft]<sub>s</sub> claims [its tactics]<sub>s</sub> are commonplace and good economically.
- 5 [The government]<sub>s</sub> may file [a civil suit]<sub>o</sub> ruling that [conspiracy]<sub>s</sub> to curb [competition]<sub>o</sub> through [collusion]<sub>x</sub> is [a violation of the Sherman Act]<sub>o</sub>.
- 6 [Microsoft]<sub>s</sub> continues to show [increased earnings]<sub>o</sub> despite [the trial]<sub>x</sub>.

# Local Entity Coherence

**Table 1**

A fragment of the entity grid. Noun phrases are represented by their head nouns. Grid cells correspond to grammatical roles: subjects (S), objects (O), or neither (X).

|   | Department | Trial | Microsoft | Evidence | Competitors | Markets | Products | Brands | Case | Netscape | Software | Tactics | Government | Suit | Earnings |   |
|---|------------|-------|-----------|----------|-------------|---------|----------|--------|------|----------|----------|---------|------------|------|----------|---|
| 1 | S          | O     | S         | X        | O           | -       | -        | -      | -    | -        | -        | -       | -          | -    | -        | 1 |
| 2 | -          | -     | O         | -        | -           | X       | S        | O      | -    | -        | -        | -       | -          | -    | -        | 2 |
| 3 | -          | -     | S         | O        | -           | -       | -        | S      | O    | O        | -        | -       | -          | -    | -        | 3 |
| 4 | -          | -     | S         | -        | -           | -       | -        | -      | -    | S        | -        | -       | -          | -    | -        | 4 |
| 5 | -          | -     | -         | -        | -           | -       | -        | -      | -    | -        | S        | O       | -          | -    | -        | 5 |
| 6 | -          | X     | S         | -        | -           | -       | -        | -      | -    | -        | -        | -       | O          | -    | -        | 6 |

- 6 sentences
- S=subject, O=object, X=neither

[Barzilay and Lapata 2008]

# Local Entity Coherence

**Table 3**

Example of a feature-vector document representation using all transitions of length two given syntactic categories S, O, X, and –.

|       | S S | S O | S X | S – | O S | O O | O X | O – | X S | X O | X X | X – | – S | – O | – X | --  |
|-------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| $d_1$ | .01 | .01 | 0   | .08 | .01 | 0   | 0   | .09 | 0   | 0   | 0   | .03 | .05 | .07 | .03 | .59 |
| $d_2$ | .02 | .01 | .01 | .02 | 0   | .07 | 0   | .02 | .14 | .14 | .06 | .04 | .03 | .07 | 0.1 | .36 |
| $d_3$ | .02 | 0   | 0   | .03 | .09 | 0   | .09 | .06 | 0   | 0   | .05 | .03 | .07 | .17 | .39 |     |

NLP

NLP

# Natural Language Processing

Dialogue Systems

# “Who’s on first” 1/2

**Abbott**

You know, strange as it may seem, they give ball players nowadays very peculiar names...Now, on the Cooperstown team we have Who's on first, What's on second, I Don't Know is on third-

**Costello**

That's what I want to find out. I want you to tell me the names of the fellows on the Cooperstown team.

**Abbott**

I'm telling you. Who's on first, What's on second, I Don't Know is on third.

**Costello**

You know the fellows' names?

**Abbott**

Yes.

**Costello**

Well, then, who's playin' first?

**Abbott**

Yes.

**Costello**

I mean the fellow's name on first base.

**Abbott**

Who.

**Costello**

The fellow's name on first base for Cooperstown.

**Abbott**

Who.

**Costello**

The guy on first base.

**Abbott**

Who is on first base.

**Costello**

Well, what are you asking me for?

**Abbott**

I'm not asking you--I'm telling you. Who is on first.

<https://www.youtube.com/watch?v=kTcRRaXV-fg>

# “Who’s on first” 2/2

**Costello**

I'm asking you--who's on first?

**Abbott**

That's the man's name.

**Costello**

That's who's name?

**Abbott**

Yes.

**Costello**

Well, go ahead, tell me!

**Abbott**

Who.

**Costello**

The guy on first.

**Abbott**

Who.

**Costello**

The first baseman.

**Abbott**

Who is on first.

**Costello**

Have you got a first baseman on first?

**Abbott**

Certainly.

**Costello**

Well, all I'm trying to find out is what's the guy's name on first base.

**Abbott**

Oh, no, no. What is on second base.

**Costello**

I'm not asking you who's on second.

# What Makes Dialogue Different

- Turn-taking
- Default turn-taking rule
  - Only take a turn at a relevant place (e.g., a pause, after a question)
- Barge-in

# Conversational Implicature

- Example:
  - How can I help you?
  - I am looking for a Thai restaurant.
- Implicature
  - Meaningful inferences that the listener can make

# Grice's Maxims

- Maxim of quantity
  - make your contribution informative
  - but not more than needed
- Maxim of quality
  - do not say what you believe is false
  - do not say that for which you lack evidence
- Maxim of relevance
- Maxim of manner
  - avoid ambiguity
  - avoid obscurity
  - be brief
  - be orderly

# Grice's Grifter Gadgets

- <http://www.naclo.cs.cmu.edu/problems2013/N2013-Q.pdf>
- <http://www.naclo.cs.cmu.edu/problems2013/N2013-QS.pdf>
- NACLO problem by Jordan Boyd-Graber

You are employed by a company that makes Grice's Grifter Gader (GGG), a small flying robot that helps you cheat at card games. The robot flies above your opponent's shoulder, looks at their cards, and then telepathically sends a message into your brain. (It's not the most ethical job in the world, but you took it because, hey, you get to work with **flying telepathic robots** — nobody could say no to that.)

These gadgets have to abide by the following maxims:

- **Relevance (R)** What GGG says should be relevant to the player's needs (winning at the card game); it should give the minimum number of facts necessary for the player to make the best play possible (telepathic communication isn't cheap!)
- **Manner (M)** In addition to giving the minimum number of facts necessary, those facts should be expressed as simply as possible
- **Quantity (N)** GGG should give all needed information, i.e. it should not leave anything out
- **Quality (L)** GGG shouldn't say things that are wrong (otherwise, what's the point of cheating)

Linguists believe that humans follow similar rules<sup>1</sup>. For example, when you ask a friend what the weather is like, he would violate the maxim of quantity if he recited the hourly barometric pressure over the previous three days. Because the GGG communicates through telepathic natural language, it should also obey these maxims.

Here's the game GGG is trying to help a player win. Before each round, the dealer shuffles a deck with forty cards, where each card has one of four suits (club ♣, heart ♥, spade ♠, diamond ♦) and a number from 1 to 10. The player and her opponent each get three cards. The player picks one of her three cards and gives it to the opponent. The opponent gets points equal to the product of the two highest numbers in the same suit (if there are no cards of the same suit, the hand is worth one point). For example:

| Opponent's Hand | Player Card | Points                        |
|-----------------|-------------|-------------------------------|
| 4♥ 3♥ 2♥        | 1♣          | $4 \times 3 = 12$             |
| 4♣ 5♥ 9♦        | 6♣          | $6 \times 4 = 24$             |
| 4♣ 5♥ 9♦        | 10♣         | 1 (no cards of the same suit) |

The GGG can't see the player's cards (it hovers above the opponent's shoulder), so it needs to give the player enough information for her to play the best card no matter what cards she has. For example, if the GGG sees that the opponent has a 4♣ 3♣ 2♦, it can't just say "play a heart", because the player might not have that in her hand.

<sup>1</sup> Language is ambiguous. In addition to the ambiguity of syntax and semantics, there are often social conventions that both speakers and listeners assume in a conversation. This was described by the linguist H. Paul Grice in the early 1960s. He proposed that speakers and listeners assume the maxims described in this problem.

Because of these maxims, conversation participants are able to make Gricean implicatures. These allow us to extrapolate from incomplete information. For example, if A asks B 'Where's Lisa?' and B replies 'Lisa got the flu,' the maxim of relevance allows A to assume that Lisa is staying at home because she is sick, even though this was never explicitly stated. Identifying and constructing these logical leaps in this restricted environment is the goal of this problem.

**Q1. What's wrong with my GGG?**

You have to debug some defective units. Given an opponent's hand and the output of a GGG, give the maxims violated (use R, N, L, or M). Each example will violate one maxim.

| Opponent's Hand | Output                                                                               | Maxim Violated |
|-----------------|--------------------------------------------------------------------------------------|----------------|
| 4♥ 3♠ 2♦        | He has a four of hearts, a three of spades, and a two of clubs.                      | _____          |
| 4♥ 3♥ 2♥        | He has a four of hearts, a three of hearts, and a two of hearts.                     | _____          |
| 4♥ 3♦ 2♠        | He has hearts, diamonds, and spades.                                                 | _____          |
| 6♥ 7♠ 3♦        | He has a six of hearts, a seven of spades, a three of diamonds, and the sky is blue. | _____          |
| 2♠ 1♠ 3♣        | He has an even prime number of spades, and the smallest odd prime number of clubs.   | _____          |

**Q2. Correcting the GGG**

Given an opponent's hand, a maxim violated, and the output of a GGG, replace the underlined portion of the output with text that would fix the violation of the maxim (without violating any others!).

|           |                                                                           |           |       |
|-----------|---------------------------------------------------------------------------|-----------|-------|
| 4♥ 2♦ 3♥  | He has a four of hearts, a two of diamonds, and <u>a three of hearts.</u> | Relevance | _____ |
| 8♦ 2♦ 10♣ | He has a ten of clubs and <u>an eight of spades.</u>                      | Quality   | _____ |
| 8♦ 2♥ 10♣ | He has an eight of diamonds <u>and a two of hearts.</u>                   | Quantity  | _____ |

**Q3. Playing the Game**

Given the following statements by a (fully functional) GGG, give a configuration of the opponent's cards that is consistent with the statement and **all** the maxims (if there's more than one possible configuration, just give one).

- A. Don't play a heart. \_\_\_\_\_
- B. He has no hearts. \_\_\_\_\_
- C. He has clubs and hearts. \_\_\_\_\_
- D. He has a three of clubs and a two of spades. \_\_\_\_\_

# Solution

**Q1.** Quality

Relevance

Quantity

Relevance

Manner

**Q2. NOTHING**

Eight of diamonds

A ten of spades and a two of hearts

**Q3. 1♥ 3♥ 8♥ — anything with all hearts**

3♦ 3♣ 3♠ — anything with identical values in different suits

Any hand where  $\max(\spadesuit) = \max(\heartsuit)$ , and any other club or heart

3♣ 2♣ and one of: 1♣/2♣/1♣

# Speech Acts

- **Assertives**
  - suggesting, putting forward, swearing, boasting, concluding
- **Directives**
  - asking, ordering, requesting, inviting, advising, begging
- **Commissives**
  - promising, planning, vowing, betting, opposing
- **Expressives**
  - thanking, apologizing, welcoming, deplored
- **Declarations**
  - I resign, you're fired.

Example from Jurafsky and Martin

# Dialogue Systems

- Architecture
  - Understanding
  - Dialogue manager
  - Task manager
  - Generation

# Prosody

- Properties of speech
  - Rhythm
  - Intonation
  - Stress
- Used to express emotions, emphasis, etc.

# Emphasis Example

- Try saying this sentence seven times, each time with an emphasis on a different word:
  - I never said she stole my money.

NLP

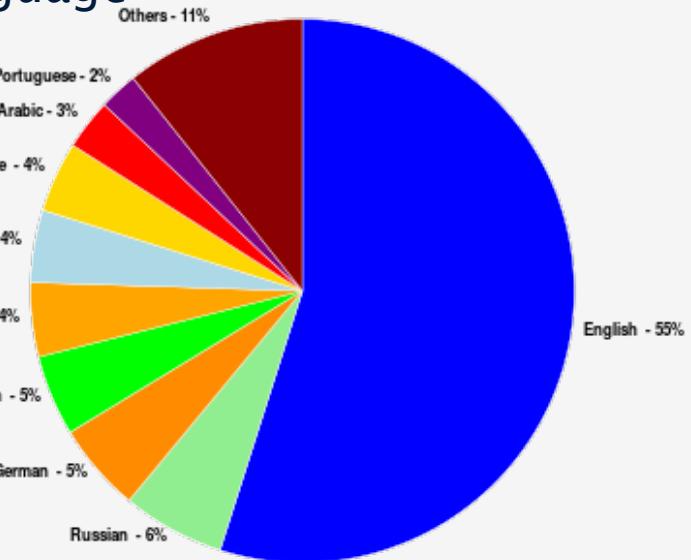
NLP

# Introduction to NLP

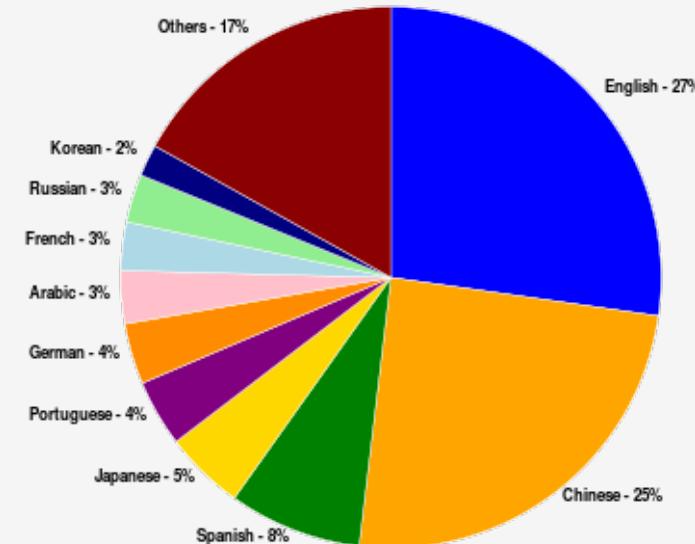
*Machine Translation*

# Multilingual Users

- Content languages for websites by language



- Percentage of Internet users by language



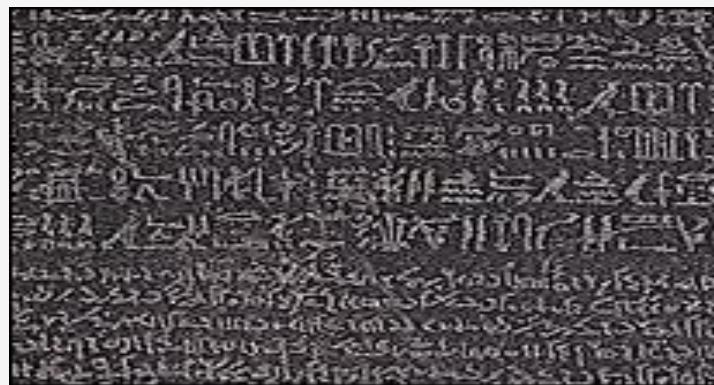
April 2013

[http://en.wikipedia.org/wiki/Global\\_Internet\\_usage](http://en.wikipedia.org/wiki/Global_Internet_usage)



[The Tower of Babel, by Pieter Bruegel the Elder, 1563]

# The Rosetta Stone



Carved in 196 BC in Egypt  
Deciphered by Champollion in 1822  
Mixture of Egyptian (hieroglyphs and Demotic) and Greek

<http://www.ancientegypt.co.uk/writing/rosetta.html>

# NACLO Problem

- [http://nacloweb.org/resources/problems/2012/  
N2012-C.pdf](http://nacloweb.org/resources/problems/2012/N2012-C.pdf)
- [http://nacloweb.org/resources/problems/2012/  
N2012-CS.pdf](http://nacloweb.org/resources/problems/2012/N2012-CS.pdf)
- Problem by Simon Zwarts, based on work by Kevin Knight

# Arcturan Problem – 1/4

*It's hard enough to translate between languages when you understand both languages. It's harder still when you only understand one. But what do computers do? They don't truly understand either language. To illustrate the challenge that computers face, Kevin Knight posed this classic puzzle (Knight 1997): given two equivalent texts in two unknown alien languages, how would you go about translating one to another?*

It is the year 2354 AD. Our scientists have been eavesdropping on messages between two alien civilizations for a very long time, but we have never met either. The closest aliens, the Centauri, have finally begun to communicate with us. Their first message was a message of peace, “Farok crrok hihok yorok clok kantok ok -yurp.”

Now, we know that the Centauri have been in contact for some time with the Arcturan race, who live in another solar system. We have never had contact with the Arcturans, but newly developed technology makes it possible for us to send them a message. We would like to send them, first, a message of peace, but because we do not understand their language, this is not an easy task.

Luckily, we have intercepted communications from the Centauri that include both languages. Here are 12 sentences in Centauri and their 12 translations in Arcturan. Unfortunately, because we have only been eavesdropping, their meaning is unknown. However, we do know that the sentence pairs on each line are translations of each other. We want to use this information to translate the original peace message from the Centauri and then send this to the Arcturans. Your assignment will be to do this translation.

# Arcturan Problem – 2/4

## CENTAURI

ok-voon ororok sprok.

ok-drubel ok-voon anok plok sprok.

erok sprok izok hihok ghirok.

ok-voon anok drok brok jok.

wiwok farok izok stok.

lalok sprok izok jok stok.

lalok farok ororok lalok sprok izok enimok.

lalok brok anok plok nok.

wiwok nok izok kantok ok-yurp.

lalok mok nok yorok ghirok clok.

lalok nok crrok hihok yorok zanzanok.

lalok rarok nok izok hihok mok.

## ARCTURAN

at-voon bichat dat.

at-drubel at-voon pippat rrat dat.

totat dat arrat vat hilat.

at-voon krat pippat sat lat.

totat jjat quat cat.

wat dat krat quat cat.

wat jjat bichat wat dat vat eneat.

iat lat pippat rrat nnat.

totat nnat quat oloat at-yurp.

wat nnat gat mat bat hilat.

wat nnat arrat mat zanzanat.

wat nnat forat arrat vat gat.

# Arcturan Problem – 3/4

**C-1** Let's start with the first Centauri word: "farok". This word occurs in two of our Centauri sentences. Given that these sentences' Arcturan translations only have one word in common with each other, we can assume that this word is the translation for "farok". Which word it is?

farok

|  |  |  |  |  |  |  |  |  |
|--|--|--|--|--|--|--|--|--|
|  |  |  |  |  |  |  |  |  |
|--|--|--|--|--|--|--|--|--|

**C-2** Do the same thing for "hihok" and "yorok". For "yorok" you will need to make some assumptions about word ordering.

hihok

|  |  |  |  |  |  |  |  |  |
|--|--|--|--|--|--|--|--|--|
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |

yorok

**C-3** The Centauri word "clok" only occurs once. However, you can figure out its Arcturan translation in another way.

clok

|  |  |  |  |  |  |  |  |  |
|--|--|--|--|--|--|--|--|--|
|  |  |  |  |  |  |  |  |  |
|--|--|--|--|--|--|--|--|--|

**C-4** Try to use the processes from the previous assignments to complete as much as possible of the following table.

crrrok

|  |  |  |  |  |  |  |  |  |
|--|--|--|--|--|--|--|--|--|
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |

kantok

ok-yurp



# Arcturan Problem – 4/4

**C-5** Complete the translation of “farok crrok hihok yorok clok kantok ok-yurp.” Keep in mind that Centauri and Arcturan sentences can have a different word order. There may be more than one correct reply.

**C-6** After some years a reply message is received in Arcturan. It reads, "Totat nnat forat arrat mat bat." Translate this message into Centauri. There may be more than one correct reply.

# Arcturan Solution – 1/3

C-I The questions in this assignment are based on examples in Knight (1997). In fact, both Centauri and Arcturan have underlying real world languages, as it turns out Centauri is English and Arcturan is Spanish. The languages are obfuscated to Centauri and Arcturan in order to illustrate how a Statistical Machine Translation (SMT) system must start from scratch, since it has no prior knowledge of how the languages work.

## CENTAURI

Ok-voon ororok sprok.  
Garcia and associates.

Ok-drubel ok-voon anok plok sprok.  
Carlos Garcia has three associates.

Erok sprok izok hihok ghirok.  
His associates are not strong.

Ok-voon anok drok brok jok.  
Garcia has a company also.

Wiwok farok izok stok.  
Its clients are angry.

Lalok sprok izok jok stok.  
The associates are also angry.

## ARCTURAN

At-voon bichat dat.  
Garcia y asociados.

At-drubel at-voon pippat rrat dat.  
Carlos Garcia tiene tres asociados.

Totat dat arrat vat hilat.  
Sus asociados no son fuertes.

At-voon krat pippat sat lat.  
Garcia tambien tiene una empresa.

Totat jjat quat cat.  
Sus clientes estan enfadados.

Wat dat krat quat cat.  
Los asociados tambien estan enfadados.

# Arcturan Solution – 2/3

Lalok farok ororok lalok sprok izok enimok.  
The clients and the associates are enemies.

Lalok brok anok plok nok.  
The company has three groups.

Wiwok nok izok kantok ok-yurp.  
Its groups are in Europe.

Lalok mok nok yorok ghirok clok.  
The modern groups sell strong pharmaceuticals.

Lalok nok crrrok hihok yorok zanzanok.  
The groups do not sell zanzanine.

Lalok rarok nok izok hihok mok.  
The small groups are not modern.

Wat jjat bichat wat dat vat eneat.  
Los clientes y los asociados son enemigos.

lat lat pippat rrat nnat.  
La empresa tiene tres grupos.

Totat nnat quat oloat at-yurp.  
Sus grupos están en Europa.

Wat nnat gat mat bat hilat.  
Los grupos modernos venden medicinas fuertes.

Wat nnat arrat mat zanzanat.  
Los grupos no venden zanzania.

Wat nnat forat arrat vat gat.  
Los grupos pequeños no son modernos.

# Arcturan Solution – 3/3

The novel sentence which was offered for translation in English is: "clients do not sell pharmaceuticals in Europe."

## Answers

C-1 jjat

C-2 hihok = arrat, yorok = mat

C-3 We need to use the process of elimination, when mapping all the words between the two sentences two words are unaligned, we assume these are translations of each other. Thus, clok = bat.

C-4 Here are the new matches:

|         |         |
|---------|---------|
| crrok   | (empty) |
| kantok  | oloat   |
| ok-yurp | at-yurp |

"crrok" does not seem to have a Arcturan equivalent, like in English the word "do" is not translated in "do not sell" which simply becomes "not sells" in Spanish. (Or to put it another way, the Centauri word *crrok* has a translation, but it's the "empty" word.)

C-5 jjat arrat mat bat oloat at-yurp

Since you cannot deduce with certainty the exact order of the Arcturan sentence, various orders of these words will be accepted.

C-6 Immediately, you are faced with a dilemma: should you translate *totat* as *erok* or *wiwok*? Because *wiwok* occurs more frequently and because you've never seen *erok* followed by any of the other words you're considering, *wiwok* seems more likely. (However, admittedly, this is only a best guess, and *erok* will also be accepted.) Next, you consider various word orders. There appears to be no grammatical path through these words. Suddenly, you remember that curious Centauri word *crrok*, which had no translation. *Crrok*, however, turns out to be a natural bridge between *nok* and *hihok*, giving you the translation:

wiwok rarok nok crrok hihok yorok clok.

# Parallel Corpora

- The Rosetta Stone
- The Hansards Corpus
- The Bible

# Hansards Example

- English
  - <s id=960001> I would like the government and the Postmaster General to agree that we place the union and the Postmaster General under trusteeship so that we can look at his books and records, including those of his management people and all the memos he has received from them, some of which must have shocked him rigid.
  - <s id=960002> If the minister would like to propose that, I for one would be prepared to support him.
- French
  - <s id=960001> Je voudrais que le gouvernement et le ministre des Postes conviennent de placer le syndicat et le ministre des Postes sous tutelle afin que nous puissions examiner ses livres et ses dossiers, y compris ceux de ses collaborateurs, et tous les mémoires qu'il a reçus d'eux, dont certains l'ont sidéré.
  - <s id=960002> Si le ministre voulait proposer cela, je serais pour ma part disposé à l'appuyer.

# English-Cebuano Bible Example

In the beginning God created the heaven and the earth.

Sa sinugdan gibuhat sa Dios ang mga langit ug ang yuta.

And God called the firmament Heaven.

Ug gihinganlan sa Dios ang hawan nga Langit.

And God called the dry land Earth

Ug ang mamala nga dapit gihinganlan sa Dios nga Yuta

- use: co-occurrence, word order, cognates
- corpora are needed
- sentence alignment needs to be done first

# NACLO Problem

- <http://nacloweb.org/resources/problems/2012/N2012-D.pdf>
- <http://nacloweb.org/resources/problems/2012/N2012-DS.pdf>
- Problem by Dragomir Radev

Many languages are related to each other for historical reasons. They may have a common ancestor or they may have borrowed words from each other. Linguists group languages into families and branches, based on their common ancestry.

Here is a list of translations of the first article of the Universal Declaration of Human Rights in 17 languages:

Your task is to identify similarities among these languages and group them into seven clusters (groups) of related languages as sketched in the diagram below:

Here is a list of translations of the first article of the Universal Declaration of Human Rights in 17 languages:

A. (English) All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood.

B. (Latin) Omnes homines dignitate et iure liberi et pares nascuntur, rationis et conscientiae participes sunt, quibus inter se concordiae studio est agendum.

C. Vsi ljudje se rodijo svobodni in imajo enako dostojanstvo in enake pravice. Obdarjeni so z razumom in vestjo in bi morali ravnati drug z drugim kakor bratje.

D. Dieub ha par en o dellezegezh hag o gwirioù eo ganet an holl dud. Poell ha skiant zo dezho ha dleout a reont bevañ an eil gant egile en ur spered a genvreudeuriezh.

E. Tuots umans naschan libers ed eguals in dignità e drets. Els sun dotats cun intellet e consciencia e desan agir tanter per in uin spiert da fraternità.

F. Toate ființele umane se nasc libere și egale în demnitate și în drepturi. Ele sunt înzestrațe cu rațiune și conștiință și trebuie să se comporte unii față de altele în spiritul fraternității.

G. Genir pawb yn rhydd ac yn gydradd â'i gilydd mewn urddas a hawliau. Fe'u cynysgaeddir â rheswm a chydwybod, a dylai pawb ymddwyn y naill at y llall mewn ysbryd cymodlon.

H. Visi žmonės gimsta laisvi ir lygūs savo orumu ir teisėmis. Jiems suteiktas protas ir sąžinė ir jie turi elgtis vienas kito atžvilgiu kaip broliai.

I. Totu sos èsseres umanos naschint liberos e eguales in dinnidade e in deretos. Issos tenent sa resone e sa cussèntzia e depent operare s'unu cun s'àteru cun ispìritu de fraternidade.

J. Gizon-emakume guztiak aske jaiotzen dira, duintasun eta eskubide berberak dituztela; eta ezaguera eta kontzientzia dutenez gero, elkarren artean senide legez jokatu beharra dute.

K. Kai rahvas roittahes vällinny da taza-arvozinnu omas arvos da oigevuksis. Jogahizele heis on annettu mieli da omatundo da heil vältämättäh pidäy olla keskenäh, kui vellil.

L. Všetci ľudia sa rodia slobodní a sebe rovní , čo sa týka ich dostôjnosti a práv. Sú obdarení rozumom a majú navzájom jednat' v bratskom duchu.

M. Nascinu tutti l'omi libari è pari di dignità è di diritti. Pussediu a raghjoni è a cuscenza è li tocca ad agis-  
cia trà elli di modu fraternu.

N. Saoláitear na daoine uile saor agus comhionann ina ndínit agus ina gcearta. Tá bauidh an réasúin agus  
an choinsiasa acu agus dlíd iad féin d'iompar de mheon bhrithreachais i leith a chéile.

O. Visi cilvēki piedzimst brīvi un vienlīdzīgi savā pašcienā un tiesībās. Viņi ir apveltīti ar saprātu un  
sirdsapziņu, un viņiem jāizturas citam pret citu brālības garā.

P. Kaikki ihmiset syntyvät vapaina ja tasavertaisina arvoltaan ja oikeuksiltaan. Heille on annettu järki ja  
omatunto, ja heidän on toimittava toisiaan kohtaan veljeyden hengessä.

Q. Wszyscy ludzie rodzą się wolni i równi pod względem swej godności i swych praw. Są oni obdarzeni  
rozumem i sumieniem i powinni postępować wobec innych w duchu braterstwa.

# Solution

1. CLQ Slavic
2. BEFIM Romance
3. J Basque
4. HO Baltic
5. DGN Celtic
6. KP Finno-Ugric
7. A English

NLP

NLP

# Machine Translation

*Basic Techniques*

# Translation as Decoding

- “One naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: 'This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.' ”
- Warren Weaver, “Translation (1955)”

# Question For The Audience

- If you know (some) French
- Look at the next two slides
- They contain a recipe in English
  - and its automatic translation to French by Google
- Can you identify the problems in the translation?

# Automatic Translation

- Systran
- Google Translate

## Ingredients

Original recipe makes 4 - 6 servings

2 cups heavy whipping cream

1 tablespoon chopped fresh basil

1 tablespoon chopped fresh thyme

2 teaspoons salt

2 teaspoons ground black pepper

1 1/2 teaspoons crushed red pepper flakes

1 teaspoon ground white pepper

1 cup chopped green onions

1 cup chopped parsley

1/2 pound shrimp, peeled and deveined

1/2 pound scallops

1/2 cup shredded Swiss cheese

1/2 cup grated Parmesan cheese

1 pound dry fettuccine pasta

## Directions

Cook pasta in a large pot of boiling salted water until al dente. Meanwhile, pour cream into large skillet. Cook over medium heat, stirring constantly, until just about boiling. Reduce heat, and add herbs, salt, peppers, onions, and parsley. Simmer 7 to 8 minutes, or until thickened. Stir in seafood, cooking until shrimp is no longer transparent. Stir in cheeses, blending well. Drain pasta. Serve sauce over noodles.

<http://translate.google.com/>

## ingrédients

Recette originale fait 4 - 6 portions

2 tasses de crème épaisse à fouetter

1 cuillère à soupe de basilic frais haché

De thym frais 1 cuillère à soupe

2 cuillères à café de sel

2 cuillères à café de poivre noir moulu

1 1/2 cuillères à café écrasés de flocons de piment rouge

Sol 1 cuillère à café de poivre blanc

1 tasse oignons verts hachés

1 tasse de persil haché

1/2 livre de crevettes, décortiquées et déveinées

1/2 escalopes *de livres*

1/2 tasse de fromage râpé suisse

1/2 tasse parmesan râpé

£ 1 sec pâtes fettucine

## instructions

Cuire les pâtes dans une grande casserole d'eau bouillante salée jusqu'à ce qu'elles soient al dente. Pendant ce temps, versez la crème dans une grande poêle. Cuire à feu moyen, en remuant constamment, jusqu'à ce que presque ébullition. Réduire le feu et ajoutez les herbes, sel, poivrons, oignons et persil. Laisser mijoter 7 à 8 minutes, ou jusqu'à ce que le mélange épaisse. Incorporer les fruits de mer, cuisson jusqu'à ce que les crevettes ne soit plus transparente. Incorporer le fromage en mélangeant bien. Égoutter les pâtes. Servir la sauce sur les nouilles.

# Answer

Cuire les pâtes dans une grande casserole d'eau bouillante salée jusqu'à ce qu'elles soient al dente. Pendant ce temps, versez la crème dans une grande poêle. Cuire à feu moyen, en remuant constamment, *jusqu'à ce que presque ébullition. Réduire le feu et ajoutez* les herbes, sel, poivrons, oignons et persil. Laisser mijoter 7 à 8 minutes, ou jusqu'à ce que le mélange épaisse. *Incorporer* les fruits de mer, *cuisson* jusqu'à ce que les crevettes ne soit plus *transparente*. Incorporer le fromage en mélangeant bien. Égoutter les pâtes. Servir la sauce sur les nouilles.

# Language Differences (1/2)

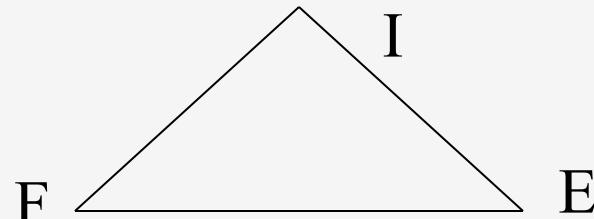
- Word order
  - SVO: English, Mandarin
  - VSO: Irish, Classical Arabic
  - SOV: Hindi, Japanese
- Prepositions (Jap.)
  - to Mariko, Mariko-ni
- Inflection (Sp.)
  - have: tengo, tienes, tenemos, tienen, tener
- Lexical distinctions (Sp.):
  - the bottle floated out – la botella salió flotando
- Brother (Jap.)
  - otooto (younger), oniisan (older)
- They (Fr.)
  - elles (feminine), ils (masculine)

# Language Differences (2/2)

- Word order in phrases (Fr.)
  - la maison bleue, the blue house
- Word order in sentences (Jap.)
  - I like to drink coffee
  - watashi wa kohii o nomu no ga suki desu
  - I-subj coffee-obj drink-dat-rheme like
- vocabulary (Sp.)
  - wall
  - pared, muro
- phrases (Fr.)
  - play
  - pièce de théâtre

# Basic Strategies of MT

- Direct Approach
  - 50's, 60's
  - naïve
  - the flesh is weak, but the spirit is strong
  - out of sight, out of mind
- Indirect: Transfer
- Indirect: Interlingua
  - No looking back
  - Language-neutral
  - No influence on the target language



# Basic Strategies of MT

- This is a blue house.
- Direct Approach
  - translate each word separately
- Transfer
  - Eng (adj noun) -> Fr (noun adj)
- Interlingua
  - House(H)  $\wedge$  Blue(H)

NLP

NLP

# Machine Translation

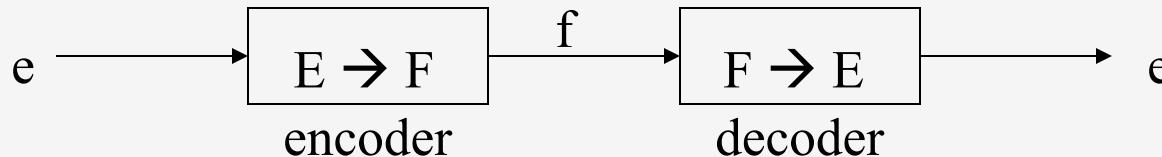
*Noisy Channel Methods*

# The Noisy Channel Model

- Source-channel model of communication
- Parametric probabilistic models of language and translation

# Statistics

- Given  $f$ , guess  $e$



$$e' = \operatorname{argmax}_e P(e|f) = \operatorname{argmax}_e P(f|e) P(e)$$

translation model      language model

# Statistical MT

Translate from French: “une fleur rouge”?

|                      | $p(e)$ | $p(f e)$ | $p(e)*p(f e)$ |
|----------------------|--------|----------|---------------|
| <i>a flower red</i>  |        |          |               |
| <i>red flower a</i>  |        |          |               |
| <i>flower red a</i>  |        |          |               |
| <i>a red dog</i>     |        |          |               |
| <i>dog cat mouse</i> |        |          |               |
| <i>a red flower</i>  |        |          |               |

# Statistical MT

|                      | $p(e)$ | $p(f e)$ | $p(e)*p(f e)$ |
|----------------------|--------|----------|---------------|
| <i>a flower red</i>  | low    | high     | low           |
| <i>red flower a</i>  |        |          |               |
| <i>flower red a</i>  |        |          |               |
| <i>a red dog</i>     |        |          |               |
| <i>dog cat mouse</i> |        |          |               |
| <i>a red flower</i>  |        |          |               |

# Statistical MT

|                      | $p(e)$ | $p(f e)$ | $p(e)*p(f e)$ |
|----------------------|--------|----------|---------------|
| <i>a flower red</i>  | low    | high     | low           |
| <i>red flower a</i>  | low    | high     | low           |
| <i>flower red a</i>  |        |          |               |
| <i>a red dog</i>     |        |          |               |
| <i>dog cat mouse</i> |        |          |               |
| <i>a red flower</i>  |        |          |               |

# Statistical MT

|                      | $p(e)$ | $p(f e)$ | $p(e)*p(f e)$ |
|----------------------|--------|----------|---------------|
| <i>a flower red</i>  | low    | high     | low           |
| <i>red flower a</i>  | low    | high     | low           |
| <i>flower red a</i>  | low    | high     | low           |
| <i>a red dog</i>     |        |          |               |
| <i>dog cat mouse</i> |        |          |               |
| <i>a red flower</i>  |        |          |               |

# Statistical MT

|                      | $p(e)$ | $p(f e)$ | $p(e)*p(f e)$ |
|----------------------|--------|----------|---------------|
| <i>a flower red</i>  | low    | high     | low           |
| <i>red flower a</i>  | low    | high     | low           |
| <i>flower red a</i>  | low    | high     | low           |
| <i>a red dog</i>     | high   | low      | low           |
| <i>dog cat mouse</i> | low    | low      | low           |
| <i>a red flower</i>  | high   | high     | high          |

# Example

$$\begin{aligned} & p(\text{Chinese}|\text{English}) \\ & \times p(\text{English}) \\ & \sim p(\text{English}|\text{Chinese}) \end{aligned}$$

# MT/Noisy Channel Models

- Text-to-text (summarization)
  - also text-to-signal, speech recognition, OCR, spelling correction
- Example (OCR)
  - $P(\text{text}|\text{pixels}) = P(\text{text}) P(\text{pixels}|\text{text})$

# Generative Story (almost IBM)

- I watched an interesting play
- I watched watched an interesting play play play
- I watched watched an play play play interesting
- J' ai vu une pièce de théâtre intéressante

# IBM's EM Trained Models (1-5)

- Word translation
- Local alignment
- Fertilities
- Class-based alignment
- Non-deficient algorithm (avoid overlaps, overflow)

# Steps

- Tokenization
- Sentence alignment (1-1, 2-2, 2-1 mappings)
  - Church and Gale (based on sentence length)
  - Church (sequences of 4-grams) - based on cognates

# Sentence Alignment

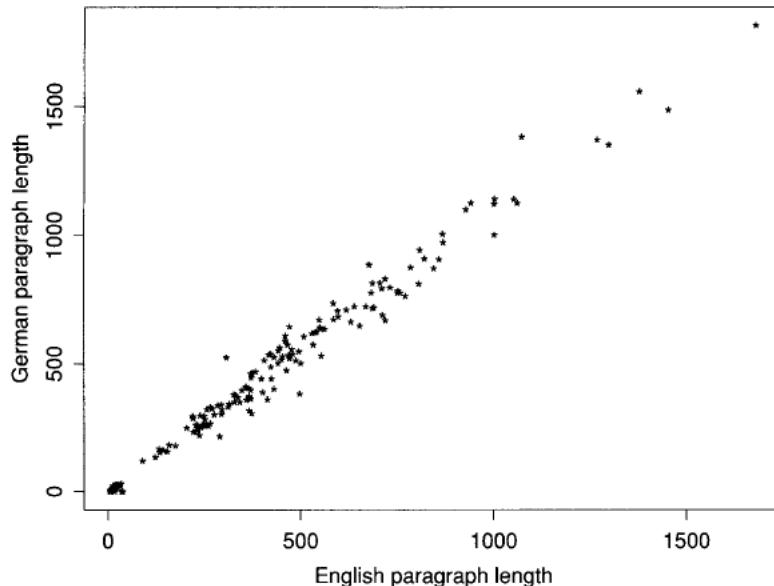


Figure 1

Paragraph lengths are highly correlated. The horizontal axis shows the length of English paragraphs, while the vertical scale shows the lengths of the corresponding German paragraphs. Note that the correlation is quite large (.991).

Table 5  
Prob(match)

| Category   | Frequency | Prob(match) |
|------------|-----------|-------------|
| 1-1        | 1167      | 0.89        |
| 1-0 or 0-1 | 13        | 0.0099      |
| 2-1 or 1-2 | 117       | 0.089       |
| 2-2        | 15        | 0.011       |
|            | 1312      | 1.00        |

[Church/Gale 1993]

# Model 1

- Alignments
  - La maison bleue
  - The blue house
  - Alignments: {1,2,3}, {1,3,2}, {1,3,3}, {1,1,1}
  - All are equally likely
- Conditional probabilities
  - $P(f|A,e) = ?$

# Model 1 (cont'd)

- Algorithm
  - Pick length of translation
  - Choose an alignment
  - Pick the French words
  - That gives you  $P(f, A|e)$
  - We need  $P(f|A, e)$
  - Use EM (expectation–maximisation) to find the hidden variables

# Model 1

- We need  $p(f|e)$  but we don't know the word alignments (which are assumed to be equally likely)

$$p(f, A | e) = p(A | e)^* p(f | A, e) = \frac{c}{(l+1)^m} \prod_{j=1}^m p(f_j | e_{a_j})$$

## Model 2

- Distortion parameters  $D(i|j,l,m)$ 
  - i and j are words in the two sentences
  - l and m are the lengths of these sentences.

# Model 3

- Fertility
- $P(\phi_i | e)$
- Examples
  - (a) play = pièce de théâtre
  - (to) place = mettre en place
- $p_1$  is an extra parameter that defines  $\phi_0$

# References

- [http://www.isi.edu/natural-language/mt/  
wkbk.rtf](http://www.isi.edu/natural-language/mt/wkbk.rtf)  
(an awesome tutorial by Kevin Knight)
  - <http://www.statmt.org/>  
(a comprehensive site, including references to  
the old IBM papers, pointers to Moses, for  
hw5, etc.)

NLP

NLP

# Machine Translation

*Advanced Methods*

# Other Techniques

- Tree-to-tree – Yamada and Knight
- Phrase-based – Och and Ney
- Syntax-based – Och et al.

# Clause Restructuring (Collins et al)

- Ich werde Ihnen den Report aushaendigen ... damit Sie den eventuell uebernehmten koennen.
- I will pass\_on to\_you the report, so\_that you can adopt that perhaps
- verb initial: that perhaps adopt can -> adopt that perhaps can
- verb second: so that you adopt...can -> so that you can adopt
- move subject: so that can you adopt -> so that you can adopt
- particles: we accept the presidency \*Particle\* -> we accept the presidency

(in German, split-prefix phrasal verbs are very common “anrufen” -> “rufen sie bitte noch einmal an” – call right back please)

# Synchronous Grammars

- Generate parse trees in parallel in two languages using different rules
- E.g.,
  - $\text{NP} \rightarrow \text{ADJ N}$  (in English)
  - $\text{NP} \rightarrow \text{N ADJ}$  (in Spanish)

# Machine Translation

*Evaluation Methods*

# Evaluation

- Human judgments
  - adequacy
  - grammaticality
  - [expensive]
- Automatic methods
  - Edit cost (at the word, character, or minute level)
  - BLEU

# BLEU (Papineni et al. 2002)

- Simple n-gram precision
- Multiple human references
- Brevity penalty
- Correlates with human assessments of automatic systems
- Doesn't correlate well when comparing human and automatic translations

# Example from MTC

- <http://www.eecs.umich.edu/~radev/nlp/mtc/>
- Chinese:
  - Napster执行长希尔伯斯辞职
- English:
  - Napster CEO Hilbers Resigns
  - Napster CEO Hilbers resigned
  - Napster Chief Executive Hilbers Resigns
  - Napster CEO Konrad Hilbers resigns

# Machine Translation

*Decoding*

# Decoding

- Find a translation that maximizes  $P(F|E)P(E)$
- NP-complete for IBM model 1
- Use a phrase translation table (e.g., Koehn's Pharaoh system, 2004)
- Use A\* search to find the subset of phrase translations that covers the source sentence
- Combine with beam search

# Tools for Machine Translation

- Language modeling toolkits
  - SRILM, CMULM
- Translation systems
  - Giza++, Moses
- Decoders
  - Pharaoh

NLP

NLP

# Introduction to NLP

*Text Generation*

# Basic NLP Pipeline

- (U)nderstanding and (G)eneration



# Definition

- Natural language generation is the process of deliberately constructing a natural language text in order to meet specified communicative goals.

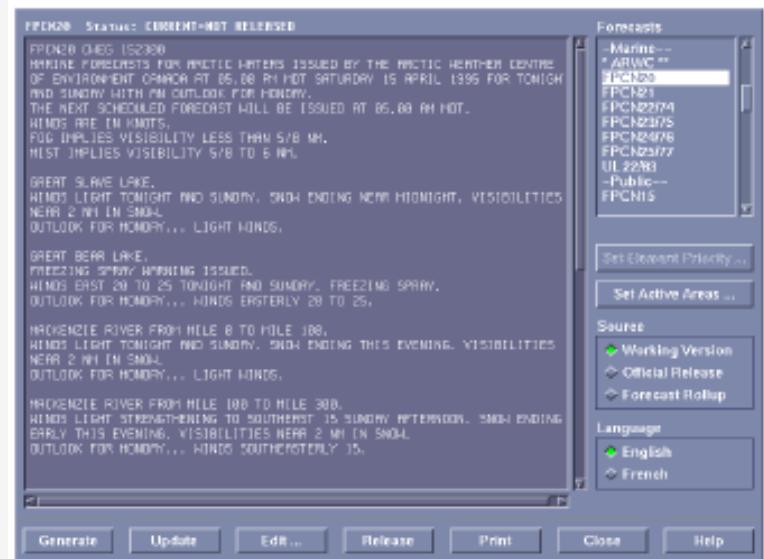
[McDonald 1992]

# What is NLG?

- Mapping meaning to text
- Stages:
  - Content selection
  - Lexical choice
  - Sentence structure: aggregation, referring expressions
  - Discourse structure

# Example of an NLG System

- FOG (Goldberg et al. 1994)
- Weather forecast reports for the Canadian Weather Service
- Input
  - Numerical simulation data annotated by humans



# Plandoc

- Function:
  - Produces a report describing the simulation options that an engineer has explored
- Input
  - A simulation log file
- Developer
  - Bellcore and Columbia University

# Input for Plandoc

- RUNID fiberall FIBER 6/19/93 act yes
- FA 1301 2 1995
- FA 1201 2 1995
- FA 1401 2 1995
- FA 1501 2 1995
- ANF co 1103 2 1995 48
- ANF 1201 1301 2 1995 24
- ANF 1401 1501 2 1995 24
- END. 856.0 670.2

# Output

- This saved fiber refinement includes all DLC changes in Run-ID ALLDLC. RUN-ID FIBERALL demanded that PLAN activate fiber for CSAs 1201, 1301, 1401 and 1501 in 1995 Q2. It requested the placement of a 48-fiber cable from the CO to section 1103 and the placement of 24-fiber cables from section 1201 to section 1301 and from section 1401 to section 1501 in the second quarter of 1995. For this refinement, the resulting 20 year route PWE was \$856.00K, a \$64.11K savings over the BASE plan and the resulting 5 year IFC was \$670.20K, a \$60.55K savings over the BASE plan.

# Considerations

- NLG is about choices
  - Content
  - Coherence
  - Style
  - Media
  - Syntax
  - Aggregation
  - Referring expressions
  - Lexical choice

NLP