

# Words & Pictures

Tamara Berg  
NLP Overview

Many slides from:  
Raymond J. Mooney, Dan Klein,  
Claire Cardie & Yejin Choi

# The Dream

- It'd be great if machines could
  - Process our email (usefully)
  - Translate languages accurately
  - Help us manage, summarize, and aggregate information
  - Use speech as a UI (when needed)
  - Talk to us / listen to us
- But they can't:
  - Language is complex, ambiguous, flexible, and subtle
  - Good solutions need linguistics and machine learning knowledge

- So:



# What is NLP?

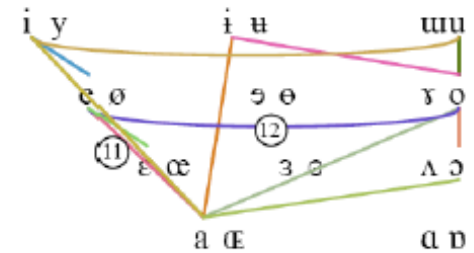


- **Fundamental goal: *deep* understand of *broad* language**
  - Not just string processing or keyword matching!
- **End systems that we want to build:**
  - Ambitious: speech recognition, machine translation, information extraction, dialog interfaces, question answering...
  - Modest: spelling correction, text categorization...

# What is Nearby NLP?

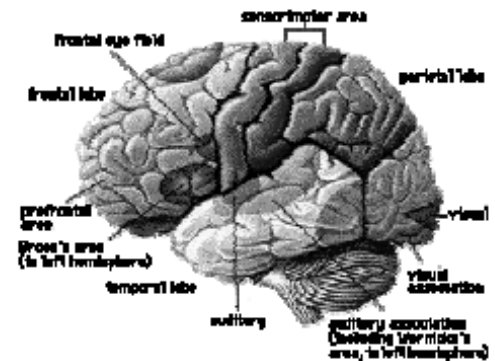
## ■ Computational Linguistics

- Using computational methods to learn more about how language works
- We end up doing this and using it



## ■ Cognitive Science

- Figuring out how the human brain works
- Includes the bits that do language
- Humans: the only working NLP prototype!



## ■ Speech?

- Mapping audio signals to text
- Traditionally separate from NLP, converging?
- Two components: acoustic models and language models
- Language models in the domain of stat NLP



# Why is NLP hard?

Reason (1) – human language is ambiguous.

- Task: Pronoun Resolution

- Jack drank the wine on the table. **It** was red and round.
- Jack saw Sam at the party. **He** went back to the bar to get another drink.
- Jack saw Sam at the party. **He** clearly had drunk too much.

[Adapted from Wilks (1975)]

# Why is NLP hard?

Reason (1) – human language is ambiguous

- Task: Preposition Attachment (aka PP-attachment)

– I ate the bread with pecans.



– I ate the bread with fingers.



# Why is NLP hard?

Reason (2) – requires reasoning beyond what is explicitly mentioned **(A,B)** , and some of the reasoning requires world knowledge **(C)**

*I couldn't submit my homework because my horse ate it.*

Implies that...

- A. *I have a horse.*
- B. *I did my homework.*
- C. *My homework was done on a soft object (such as papers) as opposed to a hard/heavy object (such as a computer). – it's more likely that my horse ate papers than a computer.*

# Why is NLP hard?

Reason (3) – Language is difficult even for human.

- Learning mother tongue (native language)
  - you might think it's easy, but...
    - compare 5 year old V.S. 10 year old V.S. 20 year old
- Learning foreign languages
  - even harder

# Is NLP really that hard?

In the back of your mind, if you're still thinking...

*“My native language is so easy. How hard can it be to type all the grammar rules, and idioms, etc into a software program? Sure it might take a while, but with enough people and money, it should be doable!”*

You are not alone!

# Brief History of NLP

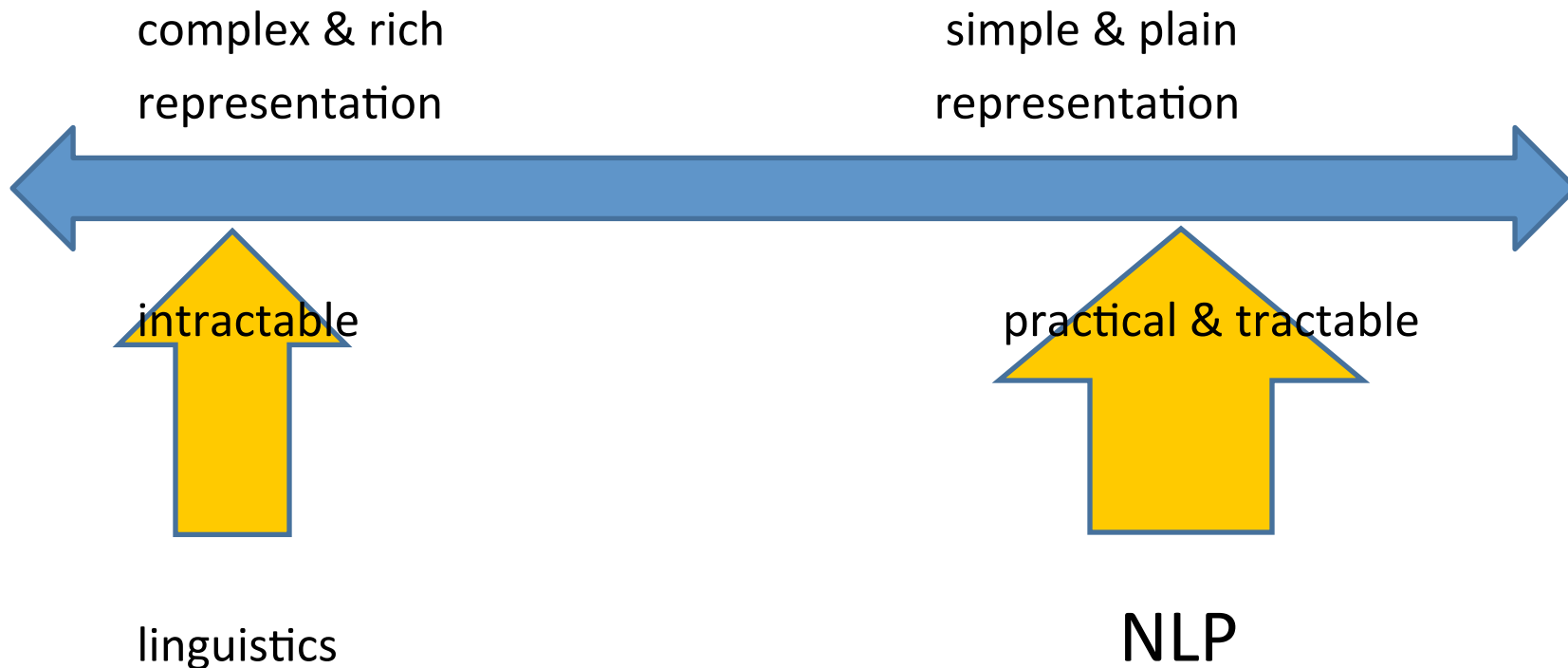
- Mid 1950's – mid 1960's: Birth of NLP and Linguistics
  - At first, people thought NLP is easy! Researchers predicted that “machine translation” can be solved in 3 years or so.
  - Mostly hand-coded rules / linguistics-oriented approaches
  - The 3 year project continued for 10 years, but still no good result, despite the significant amount of expenditure.
- Mid 1960's – Mid 1970's: A Dark Era
  - After the initial hype, a dark era follows -- people started believing that machine translation is impossible, and most abandoned research for NLP.

# Brief History of NLP

- 1970's and early 1980's – Slow Revival of NLP
  - Some research activities revived, but the emphasis is still on linguistically oriented, working on small toy problems with weak empirical evaluation
- Late 1980's and 1990's – Statistical Revolution!
  - By this time, the computing power increased substantially .
  - Data-driven, statistical approaches with simple representation win over complex hand-coded linguistic rules.
  - *“Whenever I fire a linguist our machine translation performance improves.” (Jelinek, 1988)*
- 2000's – Statistics Powered by Linguistic Insights
  - With more sophistication with the statistical models, richer linguistic representation starts finding a new value.

# Why is NLP hard?

Reason (4) – representation v.s. computability



# Why learn NLP?

- Because it's fun.
  - It's a field that is relatively young and growing rapidly
  - => a lot of opportunities for being creative and making contributions.

# Why learn NLP?

- Because you can make the world better.
  - Computer system that can help with your writing/ composition
    - beyond spell checker or grammar checker
  - Computer system that reads all the important blogs and news and provides you the summary
    - Product review analysis

# Why learn NLP?

- Because your future employer will love it.



**IBM Research**



**Powerset**  
NATURAL LANGUAGE SEARCH



**YAHOO!**  
LABS



**Microsoft**

# Natural Language

A language that is spoken, signed, or written by humans for general-purpose communication, as distinguished from formal languages (such as computer programming languages or the "languages" used in the study of formal logic) and from constructed languages (esperanto).

## Top 10 Languages used on the web

1	English	30.40%	427,436,880	7	Arabic	4.20%	59,810,400
2	Chinese	16.60%	233,216,713	8	Portuguese	4.10%	58,180,960
3	Spanish	8.70%	122,349,144	9	Korean	2.50%	34,820,000
4	Japanese	6.70%	94,000,000	10	Italian	2.40%	33,712,383
5	French	4.80%	67,315,894	11	Rest	15.20%	213,270,757
6	German	4.50%	63,611,789				

# Communication

- The goal in the production and comprehension of natural language is communication.
- Communication for the speaker:
  - **Intention**: Decide when and what information should be transmitted (a.k.a. *strategic generation*). May require planning and reasoning about agents' goals and beliefs.
  - **Generation**: Translate the information to be communicated (in internal logical representation or “language of thought”) into string of words in desired natural language (a.k.a. *tactical generation*).
  - **Synthesis**: Output the string in desired modality, text or speech.

# Communication (cont)

- Communication for the hearer:
  - **Perception**: Map input modality to a string of words, e.g. *optical character recognition (OCR)* or *speech recognition*.
  - **Analysis**: Determine the information content of the string.
    - **Syntactic interpretation (parsing)**: Find the correct parse tree showing the phrase structure of the string.
    - **Semantic Interpretation**: Extract the (literal) meaning of the string (*logical form*).
    - **Pragmatic Interpretation**: Consider effect of the overall context on altering the literal meaning of a sentence.
  - **Incorporation**: Decide whether or not to believe the content of the string and add it to the KB.

# Natural language on the web

## Regular Free text.

Graphics (from Greek γραφικός; see -graphy) are visual presentations on some surface, such as a wall, canvas, computer screen, paper, or stone to brand, inform, illustrate, or entertain. Examples are photographs, drawings, Line Art, graphs, diagrams, typography, numbers, symbols, geometric designs, maps, engineering drawings, or other images. Graphics often combine text, illustration, and color. Graphic design may consist of the deliberate selection, creation, or arrangement of typography alone, as in a brochure, flier, poster, web site, or book without any other element. Clarity or effective communication may be the objective, association with other cultural elements may be sought, or merely, the creation of a distinctive style.

Graphics can be functional or artistic. The latter can be a recorded version, such as a photograph, or an interpretation by a scientist to highlight essential features, or an artist, in which case the distinction with imaginary graphics may become blurred.

# Natural language on the web

Captions – natural language, but highly stylized & directly associated with pictures.

Doctor Nikola shows a fork that was removed from an Israeli woman who swallowed it while trying to catch a bug that flew in to her mouth, in Poriah Hospital northern Israel July 10, 2003. Doctors performed emergency surgery and removed the fork. (Reuters)

# Natural language on the web

Captions – natural language, but highly stylized & directly associated with pictures.



Doctor Nikola shows a fork that was removed from an Israeli woman who swallowed it while trying to catch a bug that flew in to her mouth, in Poriah Hospital northern Israel July 10, 2003. Doctors performed emergency surgery and removed the fork. (Reuters)

# Natural language on the web

Speech - with the explosion of video on the web the amount of speech is also growing quickly.

Need automatic speech->text translation for extracting information to associate with videos.

Total Internet	12,677,063	100.0
Google Sites	5,107,302	40.3
Fox Interactive	439,091	3.5
Viacom Digital	324,903	2.6
Yahoo! Sites	304,331	2.4
Microsoft Sites	296,285	2.3
Hulu	226,540	1.8
Turner Network	214,709	1.7
Disney Online	137,165	1.1
AOL LLC	115,306	0.9
ESPN	95,622	0.8

Number of videos on the internet, Nov 2008

# Natural language on the web

## Tags/keywords

- Folksonomy is the practice and method of collaboratively creating and managing tags to annotate and categorize content.
- Usually, freely chosen keywords are used instead of a controlled vocabulary.
- Became popular on the Web around 2004 as part of social software applications including social bookmarking and annotating photographs. Tagging allows non-expert users to collectively classify and find information.



Tag cloud showing Web 2.0 themes. Size indicates frequency of tag

# NLP 101: Syntax, Semantics, Pragmatics

- **Syntax** – grammatical ordering of words
- **Semantics** – meaning of words, phrases, sentences
- **Pragmatics** – meaning of words, phrases, sentences based on situational and social context

# Syntax V.S. Semantics

know-bodies, devoted we to under-do for you  
every Sunday-day of dressy morning, black pond,  
sky's germs, chairs' ponds - prove it, stain!  
us, rain-free & orphaned, we're living laboratories

Poem by Jeff Harrison

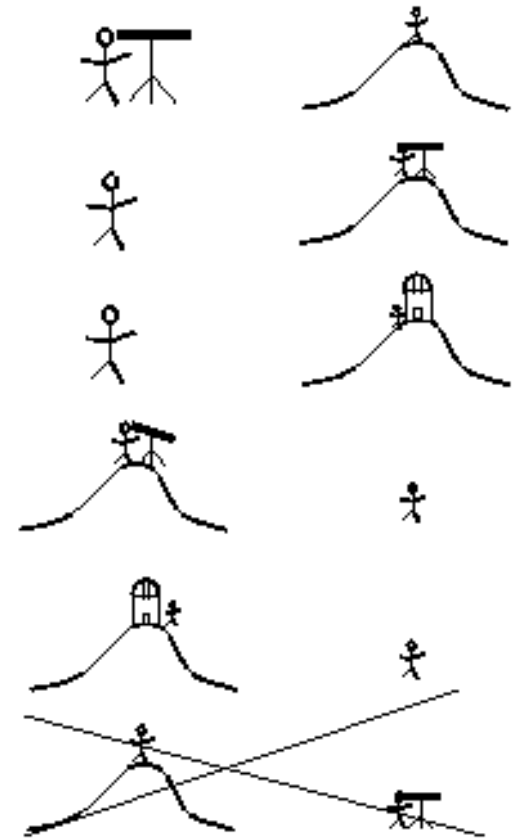
# Semantics v.s. Pragmatics

What does "You have a green light" mean?

- You are holding a green light bulb?
- You have a green light to cross the street?
- You can go ahead with your plan?

# Ambiguity

- Natural language is highly ambiguous and must be *disambiguated*.
  - I saw the man on the hill with a telescope.
  - I saw the Grand Canyon flying to LA.
  - Time flies like an arrow.
  - Horse flies like a sugar cube.
  - Time runners like a coach.
  - Time cars like a Porsche.



# Ambiguity is Ubiquitous

- Speech Recognition
  - “recognize speech” vs. “wreck a nice beach”
  - “youth in Asia” vs. “euthanasia”
- Syntactic Analysis
  - “I ate spaghetti **with** chopsticks” vs. “I ate spaghetti **with** meatballs.”
- Semantic Analysis
  - “The dog is in the **pen**.” vs. “The ink is in the **pen**.”
  - “I put the **plant** in the window” vs. “Ford put the **plant** in Mexico”
- Pragmatic Analysis
  - From “The Pink Panther Strikes Again”:
  - Clouseau: Does your dog bite?  
Hotel Clerk: No.  
Clouseau: [*bowing down to pet the dog*] Nice doggie.  
[*Dog barks and bites Clouseau in the hand*]  
Clouseau: I thought you said your dog did not bite!  
Hotel Clerk: That is not my dog.

# Ambiguity is Explosive

- Ambiguities compound to generate enormous numbers of possible interpretations.
- In English, a sentence ending in  $n$  prepositional phrases has *over*  $2^n$  syntactic interpretations (cf. Catalan numbers).
  - “I saw the man with the telescope”: **2 parses**
  - “I saw the man on the hill with the telescope.”: **5 parses**
  - “I saw the man on the hill in Texas with the telescope”:  
**14 parses**
  - “I saw the man on the hill in Texas with the telescope at noon.”: **42 parses**
  - “I saw the man on the hill in Texas with the telescope at noon on Monday” **132 parses**

# Humor and Ambiguity

- Many jokes rely on the ambiguity of language:
  - Groucho Marx: One morning I shot an elephant in my pajamas. How he got into my pajamas, I'll never know.
  - She criticized my apartment, so I knocked her flat.
  - Noah took all of the animals on the ark in pairs. Except the worms, they came in apples.
  - Policeman to little boy: "We are looking for a thief with a bicycle." Little boy: "Wouldn't you be better using your eyes."
  - Why is the teacher wearing sun-glasses. Because the class is so bright.

# Why is Language Ambiguous?

# Why is Language Ambiguous?

- Having a unique linguistic expression for every possible conceptualization that could be conveyed would make language overly complex and linguistic expressions unnecessarily long.
- Allowing resolvable ambiguity permits shorter linguistic expressions, i.e. data compression.
- Language relies on people's ability to use their knowledge and inference abilities to properly resolve ambiguities.
- Infrequently, disambiguation fails, i.e. the compression is lossy.

# Natural Languages vs. Computer Languages

- Ambiguity is the primary difference between natural and computer languages.
- Formal programming languages are designed to be unambiguous, i.e. they can be defined by a grammar that produces a unique parse for each sentence in the language.
- Programming languages are also designed for efficient (deterministic) parsing.

# Natural Language Tasks

- Processing natural language text involves many various syntactic, semantic and pragmatic tasks in addition to other problems.

# Syntactic Tasks

# Word Segmentation

- Breaking a string of characters (graphemes) into a sequence of words.
- In some written languages (e.g. Chinese) words are not separated by spaces.
- Even in English, characters other than white-space can be used to separate words [e.g. , ; . - : ( ) ]
- Examples from English URLs:
  - jumptheshark.com ⇒ jump the shark .com
  - myspace.com/pluckerswingbar
    - ⇒ myspace .com pluckers wing bar
    - ⇒ myspace .com plucker swing bar

# Morphological Analysis

- **Morphology** is the field of linguistics that studies the internal structure of words. (Wikipedia)
- A **morpheme** is the smallest linguistic unit that has semantic meaning (Wikipedia)
  - e.g. “carry”, “pre”, “ed”, “ly”, “s”
- Morphological analysis is the task of segmenting a word into its morphemes:
  - carried  $\Rightarrow$  carry + ed (past tense)
  - independently  $\Rightarrow$  in + (depend + ent) + ly
  - Googlers  $\Rightarrow$  (Google + er) + s (plural)
  - unlockable  $\Rightarrow$  un + (lock + able) ?  
 $\Rightarrow$  (un + lock) + able ?

# Part Of Speech (POS) Tagging

- Annotate each word in a sentence with a part-of-speech.

I ate the spaghetti with meatballs.

Pro V Det N Prep N

John saw the saw and decided to take it to the table.

PN V Det N Con V Part V Pro Prep Det N

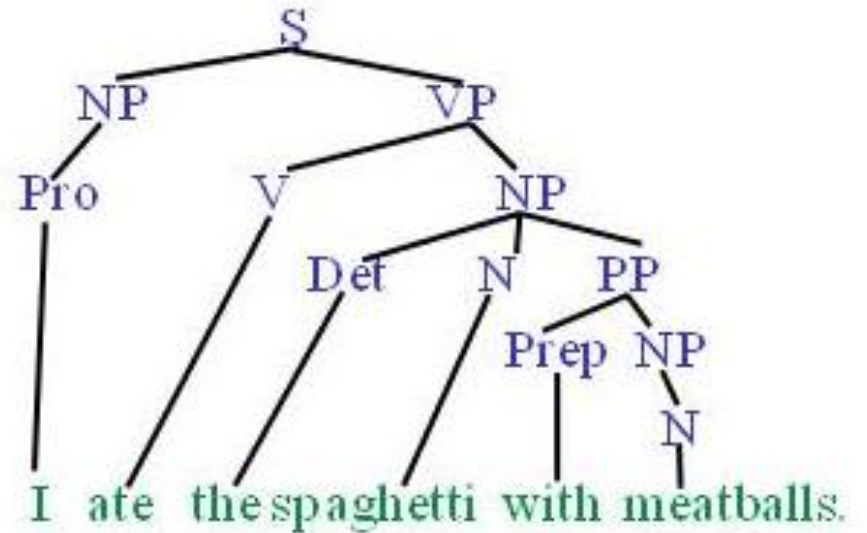
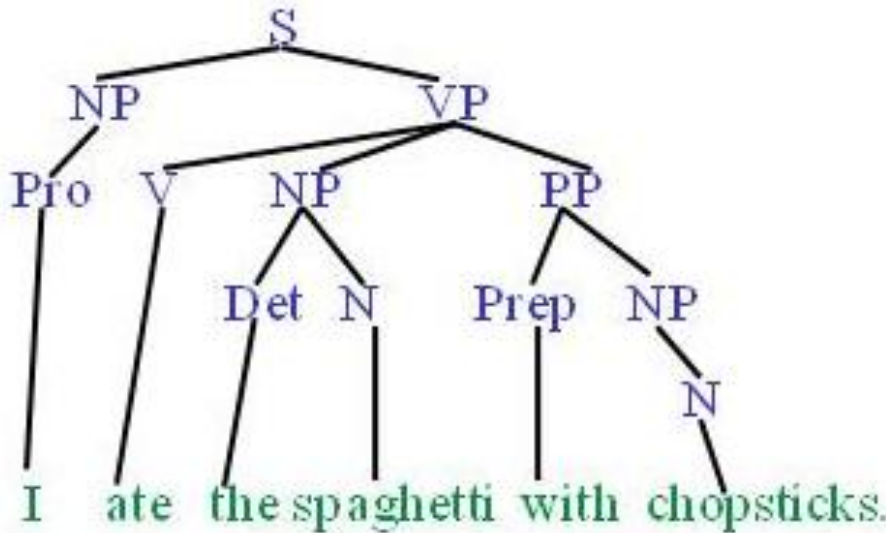
- Useful for subsequent syntactic parsing and word sense disambiguation.

# Phrase Chunking

- Find all noun phrases (NPs) and verb phrases (VPs) in a sentence.
  - [NP I] [VP ate] [NP the spaghetti] [PP with] [NP meatballs].
  - [NP He ] [VP reckons ] [NP the current account deficit ] [VP will narrow ] [PP to ] [NP only # 1.8 billion ] [PP in ] [NP September ]

# Syntactic Parsing

- Produce the correct syntactic parse tree for a sentence.



# Semantic Tasks

# Word Sense Disambiguation (WSD)

- Words in natural language usually have a fair number of different possible meanings.
  - Ellen has a strong **interest** in computational linguistics.
  - Ellen pays a large amount of **interest** on her credit card.
- For many tasks (question answering, translation), the proper sense of each ambiguous word in a sentence must be determined.

# Semantic Role Labeling (SRL)

- For each clause, determine the semantic role played by each noun phrase that is an argument to the verb.

agent patient source destination instrument

– John drove Mary from Austin to Dallas in his Toyota Prius.

– The hammer broke the window.

- Also referred to a “case role analysis,” “thematic analysis,” and “shallow semantic parsing”



# Textual Entailment



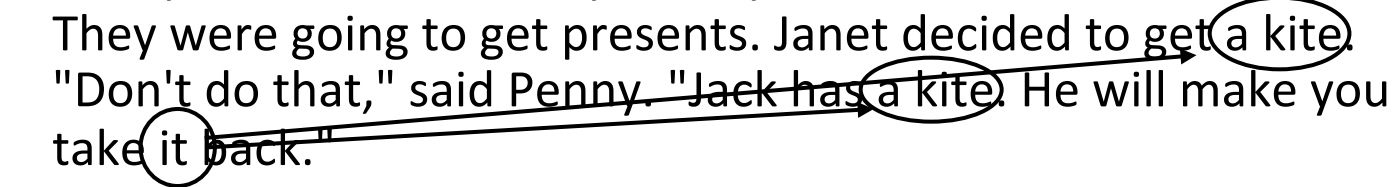
- Determine whether one natural language sentence entails (implies) another under an ordinary interpretation.

# Textual Entailment Problems from PASCAL Challenge

TEXT	HYPOTHESIS	ENTAILMENT
<i>Eyeing the huge market potential, currently led by Google, Yahoo took over search company Overture Services Inc last year.</i>	<i>Yahoo bought Overture.</i>	TRUE
<i>Microsoft's rival Sun Microsystems Inc. bought Star Office last month and plans to boost its development as a Web-based device running over the Net on personal computers and Internet appliances.</i>	<i>Microsoft bought Star Office.</i>	FALSE
<i>The National Institute for Psychobiology in Israel was established in May 1971 as the Israel Center for Psychobiology by Prof. Joel.</i>	<i>Israel was established in May 1971.</i>	FALSE
<i>Since its formation in 1948, Israel fought many wars with neighboring Arab countries.</i>	<i>Israel was established in 1948.</i>	TRUE

# Pragmatics/Discourse Tasks

## Anaphora Resolution/ Co-Reference

- Determine which phrases in a document refer to the same underlying entity.
  - John put the carrot on the plate and ate it.
  - Bush started the war in Iraq. But the president needed the consent of Congress.
- Some cases require difficult reasoning.
  - Today was Jack's birthday. Penny and Janet went to the store. They were going to get presents. Janet decided to get a kite. "Don't do that," said Penny. "Jack has a kite. He will make you take it back."

# Other Tasks

# Other Tasks

---

- Useful applications...
  - E.g. information retrieval

Topic: Advantages and disadvantages of using potassium hydroxide in any aspect of organic farming, especially...



# Other Tasks

---

- Useful applications...

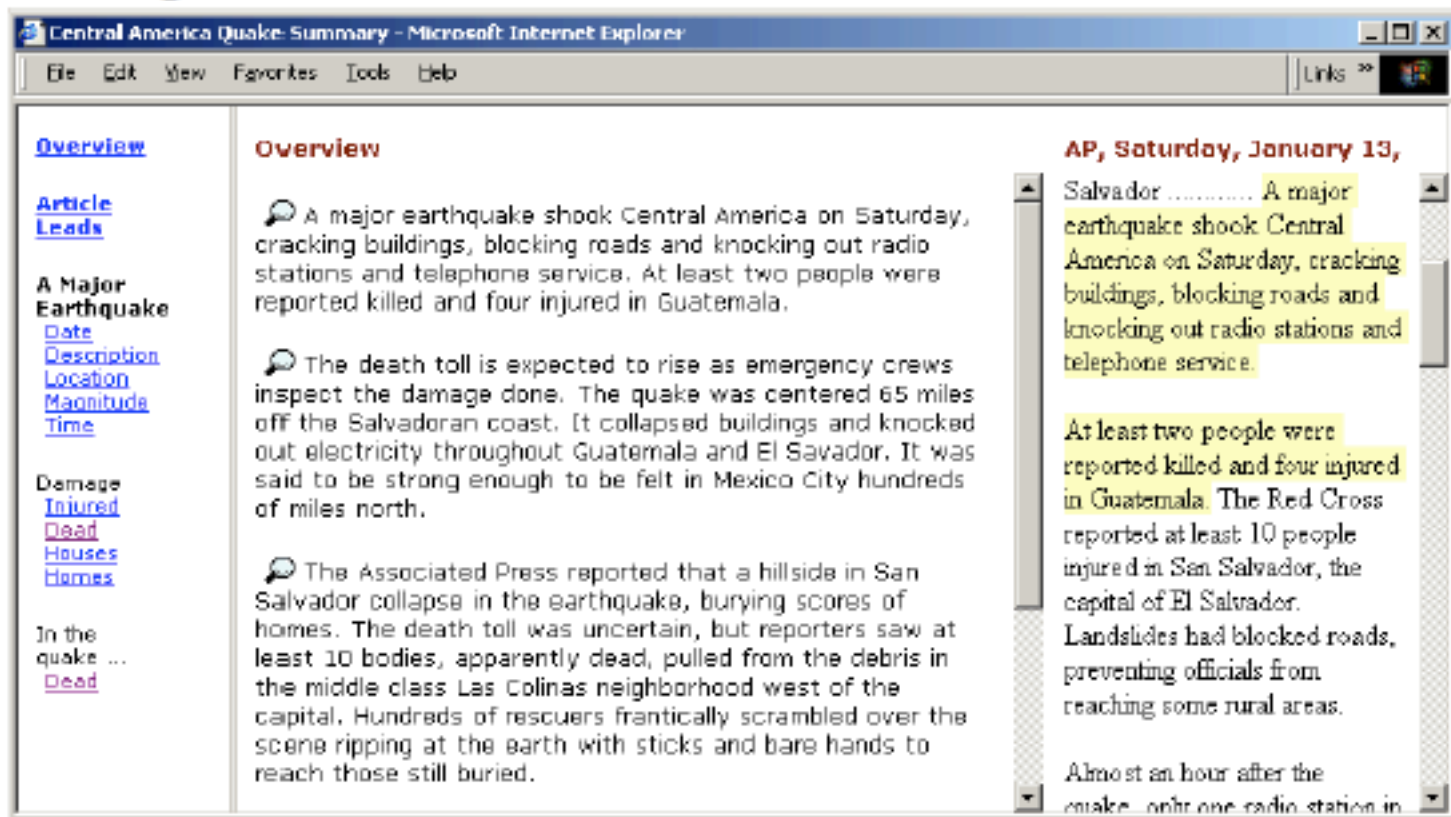
- E.g. question answering systems
  - » How many calories are there in a Big Mac?
  - » Who is the voice of Miss Piggy?
  - » Who was the first American in space?
- Retrieve not just relevant documents, but return the answer



Start System

# Other Tasks

- Useful applications...
  - E.g. summarization



[White et al., 2002]

Google search

Slide from Claire Cardie

# Machine Translation

---

## Original Text

新华网石家庄11月16日电（记者 张涛）11月15日是河北省沧州市的“供暖日”，当地大风、阴雨天，最低气温降至1℃。然而，至少上千户市民家里的暖气仍是冰凉的。原来，这个市今年实施有史以来最大规模的集中供暖“扩面”工程，许多居民小区过去的小锅炉关停、拆除了，而集中供暖却因工程量太大要推迟半个月。

## Translated Text

-- Shijiazhuang, November 16 (Xinhua Zhang Tao) November 15 is the city of Cangzhou, Hebei Province "heating Day," local windy, rainy days, the minimum temperature dropped to 1 °C. However, at least 1,000 members of the public on home heating is still cool. Originally, the city implemented this year's biggest ever focus on heating "expansion of" works, many small residential area in the past a small boiler shutdown, demolition, and the central heating because of too much work should be delayed two weeks.

- SOTA: much better than nothing, but more an understanding aid than a replacement for human translators
- New, better methods e.g. [babelfish](#), [translation party](#)

## Ambiguity Resolution is Required for Translation

- Syntactic and semantic ambiguities must be properly resolved for correct translation:
  - “John plays the guitar.” → “John toca la guitarra.”
  - “John plays soccer.” → “John juega el fútbol.”
- An apocryphal story is that an early MT system gave the following results when translating from English to Russian and then back to English:
  - “The spirit is willing but the flesh is weak.” ⇒ “The liquor is good but the meat is spoiled.”
  - “Out of sight, out of mind.” ⇒ “Invisible idiot.”

# Resolving Ambiguity

- Choosing the correct interpretation of linguistic utterances requires knowledge of:
  - Syntax
    - An agent is typically the subject of the verb
  - Semantics
    - Michael and Ellen are names of people
    - Austin is the name of a city (and of a person)
    - Toyota is a car company and Prius is a brand of car
  - Pragmatics
  - World knowledge
    - Credit cards require users to pay financial interest
    - Agents must be animate and a hammer is not animate

# Manual Knowledge Acquisition

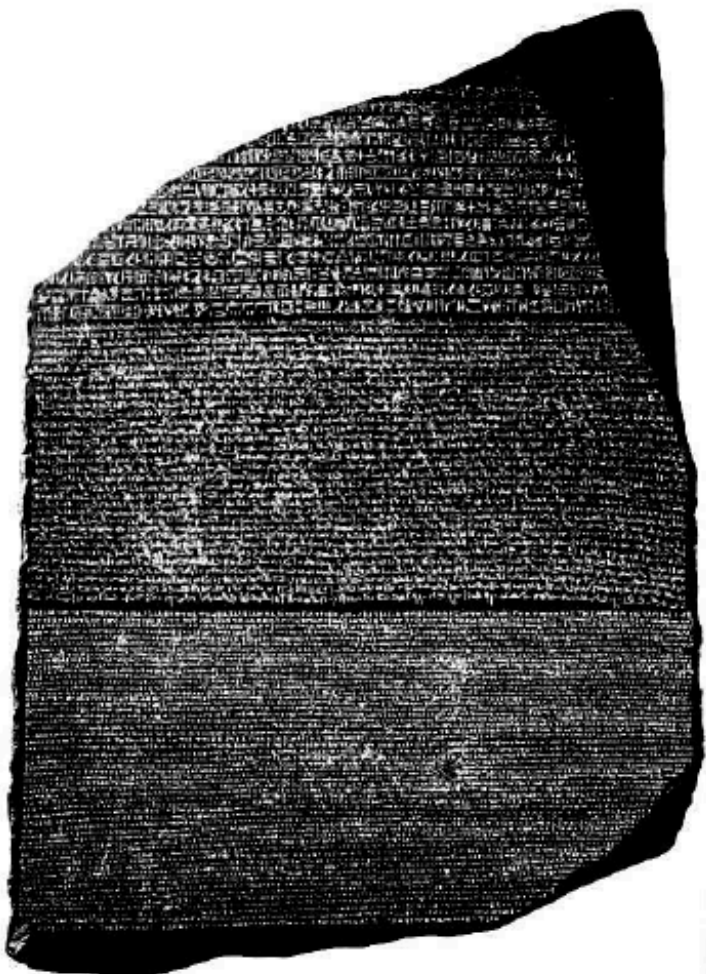
- Traditional, “rationalist,” approaches to language processing require human specialists to specify and formalize the required knowledge.
- Manual knowledge engineering, is difficult, time-consuming, and error prone.
- “Rules” in language have numerous exceptions and irregularities.
  - “All grammars leak.”: Edward Sapir (1921)
- Manually developed systems were expensive to develop and their abilities were limited and “brittle” (not robust).

# Automatic Learning Approach

- Use machine learning methods to automatically acquire the required knowledge from appropriately annotated text corpora.
- Various referred to as the “corpus based,” “statistical,” or “empirical” approach.
- Statistical learning methods were first applied to speech recognition in the late 1970’s and became the dominant approach in the 1980’s.
- During the 1990’s, the statistical training approach expanded and came to dominate almost all areas of NLP.

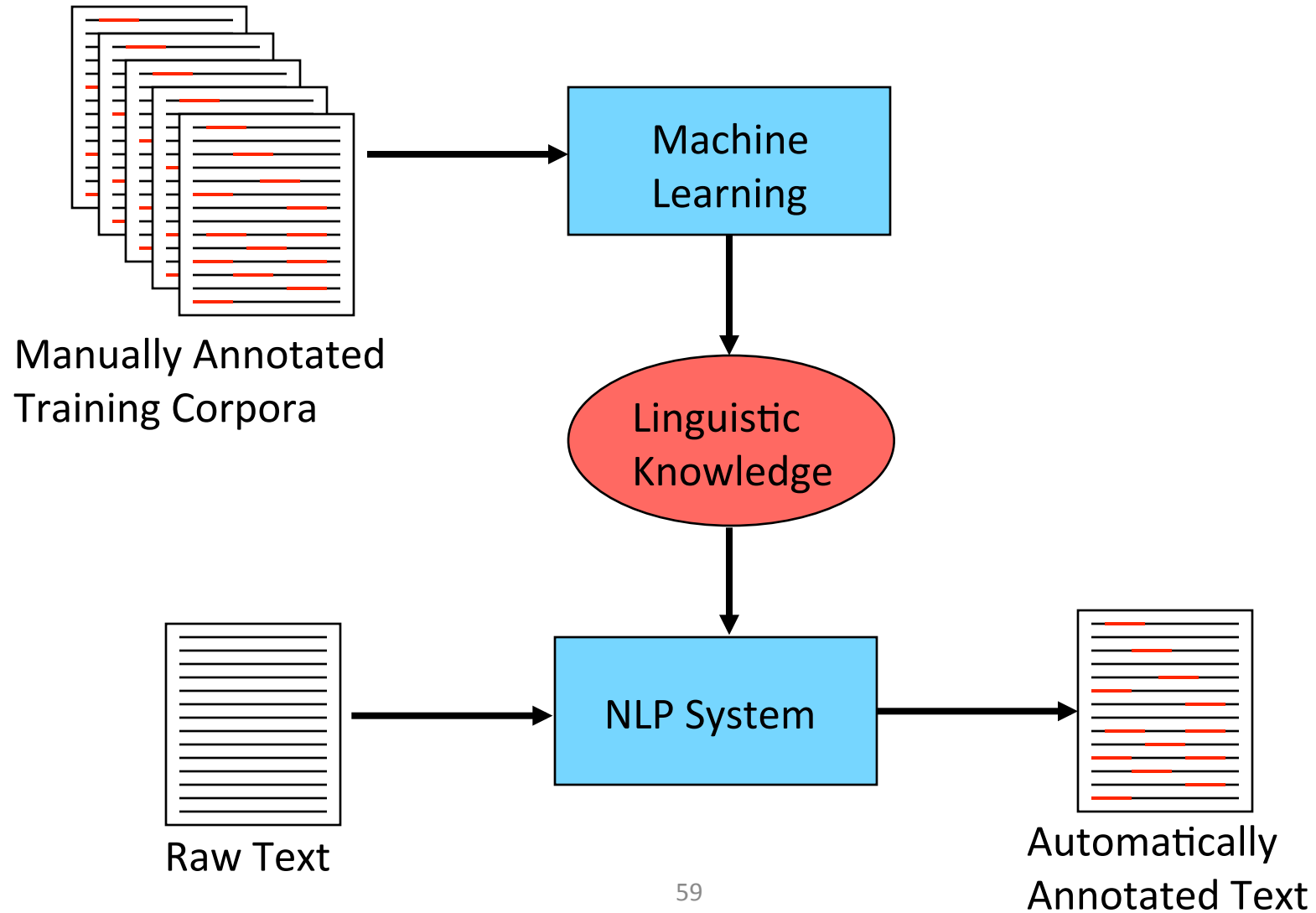
# Corpora

---



- A corpus is a collection of text
  - Often annotated in some way
  - Sometimes just lots of text
  - Balanced vs. uniform corpora
- Examples
  - Newswire collections: 500M+ words
  - Brown corpus: 1M words of tagged “balanced” text
  - Penn Treebank: 1M words of parsed WSJ
  - Canadian Hansards: 10M+ words of aligned French / English sentences
  - The Web: billions of words of who knows what

# Learning Approach



# Advantages of the Learning Approach

- Large amounts of electronic text are now available.
- Annotating corpora is easier and requires less expertise than manual knowledge engineering.
- Learning algorithms have progressed to be able to handle large amounts of data and produce accurate probabilistic knowledge.
- The probabilistic knowledge acquired allows robust processing that handles linguistic regularities as well as exceptions.

# The Importance of Probability

- Unlikely interpretations of words can combine to generate spurious ambiguity:
  - “The a are of l” is a valid English noun phrase (Abney, 1996)
    - “a” is an adjective for the letter A
    - “are” is a noun for an area of land (as in hectare)
    - “l” is a noun for the letter l
  - “Time flies like an arrow” has 4 parses, including those meaning:
    - Insects of a variety called “time flies” are fond of a particular arrow.
    - A command to record insects’ speed in the manner that an arrow would.
- Some combinations of words are more likely than others:
  - “vice president Gore” vs. “dice precedent core”
- Statistical methods allow computing the most likely interpretation by combining probabilistic evidence from a variety of uncertain knowledge sources.

# Problem: Sparsity

- However: sparsity is always a problem
  - New unigram (word), bigram (word pair), and rule rates in newswire

