Natural Language Processing
CS224N/Ling284

Bill MacCartney
-> Gerald Penn <-
Winter 2011
Lecture 1

Course logistics in brief

- Instructors: Bill MacCartney and Gerald Penn
- TAs: Angel Chang, Shrey Gupta and Ritvik Mudur
- Time: MW 4:15–5:30.
  - Section: Fri ?? in ??
- Programming language: Java 1.6+
- Other information: see the webpage: http://cs224n.stanford.edu/
- Handouts vs.  ?

This class

- Assumes you come with some skills…
  - Some basic linear algebra, probability, and statistics; decent programming skills
  - But not everyone has the same skills
    - Assumes some ability to learn missing knowledge
- Teaches key theory and methods for statistical NLP: MT, information extraction, parsing, semantics, etc.
  - Learn techniques which can be used in practical, robust systems that can (partly) understand human language
- But it's something like an "AI Systems" class:
  - A lot of it is hands-on, problem-based learning
  - Often practical issues are as important as theoretical niceties
  - We often combine a bunch of ideas

How we will determine your grade

- 20% x 3 programming assignments
  - Assignments are due by 5pm on the respective due date.
- 0.33% x 18 quizzes (1 in each lecture)
  - Quiz answers must be received by 5pm on the Sunday following the lecture in which the quiz was posed.
- 34% x 1 final project on a topic of your choosing
  - Project proposals (unmarked) due on 9th February, 2011;
  - Projects due on 9th March, 2011;
  - Short project presentations will be made during the final examination period.
- Many more details about these can be found on the "Assignments" page of the class website.

Section timings – let's vote

- 9:00-9:50 (Skilling 193, Aud)
- 1:15-2:05 (Skilling 191)
- 2:15-3:05 (Skilling 191; Gates B03)
- 3:15-4:05 (Skilling 191, 193; Gates B01)
- 4:15-5:05 (Skilling 191, 193, Aud; Huang 018)

Natural language: the earliest UI

Star Trek: - universal translators;
- Data, the universe's only neural-network-powered robot, but no Bluetooth or 802.11

(cf. also false Maria in Metropolis – 1926)

Goals of the field of NLP

- Computers would be a lot more useful if they could handle our email, do our library research, chat to us …
- But they are fazed by natural human languages.
  - Or at least their programmers are … most people just avoid the problem and get into XML, or menus and drop boxes, or …
- But someone has to work on the hard problems!
  - How can we tell computers about language?
  - Or help them learn it as kids do?
- In this course we seek to identify many of the open research problems in natural language

What/where is NLP?

- Goals can be very far reaching …
  - True text understanding
  - Reasoning about texts
  - Real-time participation in spoken dialogs
  - IBM's QA system will be on Jeopardy! 14th-16th Feb.
- Or very down-to-earth …
  - Finding the price of products on the web
  - Analyzing reading level or authorship statistically
  - Sentiment detection about products or stocks
  - Extracting facts or relations from documents
- These days, the latter predominate (as NLP becomes increasingly practical, it is increasingly engineering-oriented – also related to changes in approach in AI/NLP)

Commercial world

The hidden structure of language

- We're going beneath the surface…
  - Not just string processing
  - Not just keyword matching in a search engine
    - Search Google on "tennis racquet" and "tennis racquets" or "laptop" and "notebook" and the results are quite different … though these days Google does lots of subtle stuff beyond keyword matching itself
    - Not just converting a sound stream to a string of words
      - Like Nuance/IBM/Dragon/Philips speech recognition
- We want to recover and manipulate at least *some* aspects of language structure and meaning

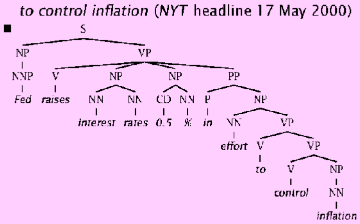
Is the problem just cycles?

- Bill Gates, Remarks to Gartner Symposium, October 6, 1997:
  - Applications always become more demanding. Until the computer can speak to you in perfect English and understand everything you say to it and learn in the same way that an assistant would learn -- until it has the power to do that -- we need all the cycles. We need to be optimized to do the best we can. Right now linguistics are right on the edge of what the processor can do. As we get another factor of two, then speech will start to be on the edge of what it can do.

The early history: 1950s

- Early NLP (Machine Translation) on machines less powerful than pocket calculators
- Foundational work on automata, formal languages, probabilities, and information theory
- First speech systems (Davis et al., Bell Labs)
- MT heavily funded by military – a lot of it was just word substitution programs but there were a few seeds of later successes, e.g., trigrams
- Little understanding of natural language syntax, semantics, pragmatics
- Problem soon appeared intractable

Why is NLU difficult? The hidden structure of language is hugely ambiguous

- Structures for: *Fed raises interest rates 0.5% in effort to control inflation* (*NYT* headline 17 May 2000)
- 

Where are the ambiguities?

Part of speech ambiguities

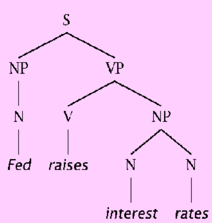
Syntactic attachment ambiguities

		VB					
	VBZ	VBP	VBZ				
NNP	NNS	NN	NNS	CD	NN		
Fed	*raises*	*interest*	*rates*	*0.5*	*%*	*in*	*effort*
						to	*control*
							inflation

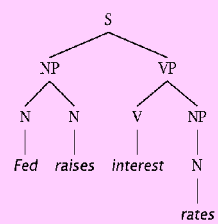
Word sense ambiguities: Fed → *"federal agent"*
interest → *a feeling of wanting to know or learn more*

Semantic interpretation ambiguities above the word level

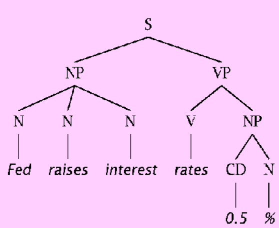
The bad effects of V/N ambiguities (1)



The bad effects of V/N ambiguities (2)



The bad effects of V/N ambiguities (3)



Why NLP is difficult: Newspaper headlines

- Minister Accused Of Having 8 Wives In Jail
- Juvenile Court to Try Shooting Defendant
- Teacher Strikes Idle Kids
- China to Orbit Human on Oct. 15
- Local High School Dropouts Cut in Half
- Red Tape Holds Up New Bridges
- Clinton Wins on Budget, bu
- Hospitals Are Sued by 7 Fo
- Police: Crack Found in Man

Reference Resolution

U: Where is The Green Hornet playing in Mountain View?
S: The Green Hornet is playing at the Century 16 theater.
U: When is it playing there?
S: It's playing at 2pm, 5pm, and 8pm.
U: I'd like 1 adult and 2 children for the first show.
How much would that cost?

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

Why is natural language computing hard?

- Natural language is:
  - highly ambiguous at all levels
  - complex and subtle use of context to convey meaning
  - fuzzy?, probabilistic
  - involves reasoning about the world
  - a key part of people interacting with other people (a social system):
    - persuading, insulting and amusing them
- But NLP can also be surprisingly easy sometimes:
  - rough text features can often do half the job

Making progress on this problem…

- The task is difficult! What tools do we need?
  - Knowledge about language
  - Knowledge about the world
  - A way to combine knowledge sources
- We used to write big honking grammars
- The answer that's been getting more traction:
  - probabilistic models built from language data
    - P("maison" → "house") high
    - P("L'avocat général" → "the general avocado") low
- Some computer scientists think this is a new "A.I." idea
  - But really it's an old idea that was stolen from the electrical engineers….

Where do we head?

Look at subproblems, approaches, and applications at different levels

- Statistical machine translation
- Statistical NLP: classification and sequence models (part-of-speech tagging, named entity recognition, information extraction)
- Syntactic (probabilistic) parsing
- Building semantic representations from text; QA.

- *(Unfortunately left out: natural language generation, phonology/morphology, speech dialogue systems, more on natural language understanding, …. There are other classes for some!)*

Daily Question!

- What is the ambiguity in this (authentic!) newspaper headline?

Ban on Nude Dancing
on Governor's Desk

Choose the intended reading of this headline:

a) [Ban [on Nude Dancing]] [on Governor's Desk]

b) [Ban on [[Nude Dancing] on Governor's Desk]]

c) [Ban on [Nude [Dancing on Governor's Desk]]]

d) [[Ban on Nude] Dancing [on Governor's Desk]]

Machine Translation

美国关岛国际机场及其办公室均接获一名自称沙地阿拉伯富商拉登等发出的电子邮件，威胁将会向机场等公众地方发动生化袭击後，关岛经保持高度戒备。

→

The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport .

The holy grail of natural language processing.

Requires capabilities in both interpretation and generation.

About $10 billion spent annually on human translation.

Scott Klemmer: I learned a surprising fact at our research group lunch today. Google Sketchup releases a version every 18 months, and the primary difficulty of releasing more often is not the difficulty of producing software, but the cost of internationalizing the user manuals!

Mainly slides from Kevin Knight (at ISI)

Translation (human and machine)

[redacted]

Ref: According to the data provided today by the Ministry of Foreign Trade and Economic Cooperation, as of November this year, China has actually utilized 46.959 billion US dollars of foreign capital, including 40.007 billion US dollars of direct investment from foreign businessmen.

IBM4:

the Ministry of Foreign Trade and Economic Cooperation, including foreign direct investment 40.007 billion US dollars today provide data include that year to November china actually using foreign 46.959 billion US dollars and

Yamada/Knight:

today's available data of the Ministry of Foreign Trade and Economic Cooperation shows that china's actual utilization of November this year will include 40.007 billion US dollars for the foreign direct investment among 46.959 billion US dollars in foreign capital

Machine Translation History

- 1950s: Intensive research activity in MT
- 1960s: Direct word-for-word replacement
- 1966 (ALPAC): NRC Report on MT
  - Conclusion: MT no longer worthy of serious scientific investigation.
- 1966-1975: 'Recovery period'
- 1975-1985: Resurgence (Europe, Japan)
  - Domain specific rule-based systems
- 1985-1995: Gradual Resurgence (US)
- 1995-2010: Statistical MT surges ahead

http://ourworld.compuserve.com/homepages/WJHutchins/MTS-93.htm

Warren Weaver

- "Also knowing nothing official about, but having guessed and inferred considerable about, the powerful new mechanized methods in cryptography—methods which I believe succeed even when one does not know what language has been coded—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 (1955:18, quoting a letter he wrote in 1947)

What happened between ALPAC and Now?

- Need for MT and other NLP applications confirmed
- Change in expectations
- Computers have become faster, more powerful
- WWW
- Political state of the world
- Maturation of Linguistics
- Hugely increased availability of data
- Development of statistical and hybrid statistical/symbolic approaches

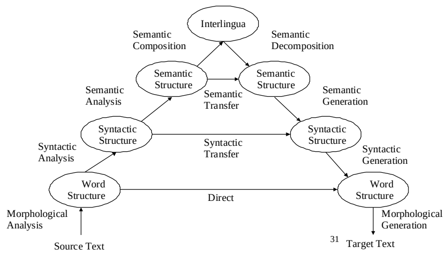
طالبت منظمة هيومن رايتس ووتش السلطات الإسرائيلية بأن ترفع فوراً القيود
التي تحرم تلاميذ المدارس العامة في قطاع غزة من الكتب واحتياجات المدارس
الأساسية مثل الأوراق والأقلام.

Called on organization Human Rights Watch the Israeli authorities to immediately lift restrictions that deny public school students in the Gaza Strip books

Google translate

Called Human Rights Watch, the Israeli authorities to immediately lift restrictions that prohibit public school students in the Gaza Strip of books and basic school needs such as paper and pens.

Three MT Approaches: Direct, Transfer, Interlingual (Vauquois triangle)



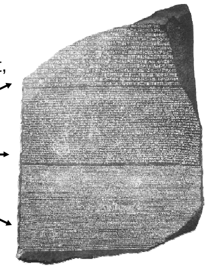
Statistical Solution

- Parallel Texts
  - Rosetta Stone (Egypt, 196 BCE)

Hieroglyphs

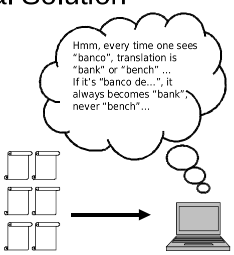
Enchorial Egyptian

Greek

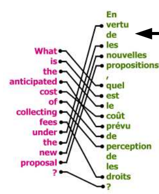


Statistical Solution

- Parallel Texts
  - Instruction Manuals
  - Hong Kong Legislation
  - Macao Legislation
  - Canadian Parliament Hansards
  - United Nations Reports
  - Official Journal of the European Communities

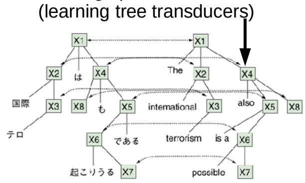


Alignment in Statistical MT



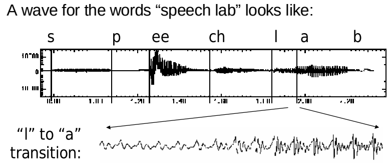
We either align words or phrases (learning *distortions* and *fertility*)...

...or align pieces of trees (learning tree transducers)



Speech Recognition: Acoustic Waves

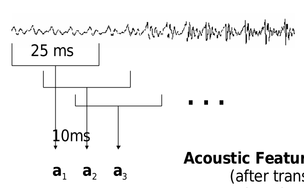
- Human speech generates a wave
  - like a loudspeaker moving
- A wave for the words "speech lab" looks like:



Graphs from Simon Arnfield's web tutorial on speech, Sheffield: http://www.psyc.leeds.ac.uk/research/cogn/speech/tutorial/

Acoustic Sampling

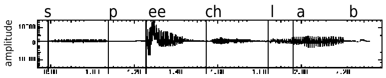
- 10 ms frame (ms = millisecond = 1/1000 second)
- ~25 ms window around frame [wide band] to allow/smooth signal processing – it let's you see formants



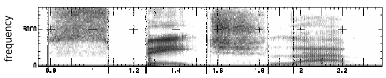
Result:
Acoustic Feature Vectors
(after transformation, numbers in roughly $\mathbf{R}^{14}$)

Spectral Analysis

- Frequency gives pitch (sort of); amplitude gives volume
  - sampling at ~8 kHz phone, ~16 kHz mic (kHz=1000 cycles/sec)



- Fourier transform of wave displayed as a spectrogram
  - darkness indicates energy at each frequency
  - hundreds to thousands of frequency samples



The Speech Recognition Problem

- The **Recognition Problem: Noisy channel model**
  - We started out with English words, they were encoded as an audio signal, and we now wish to decode.
  - Find most likely sequence **w** of "words" given the sequence of acoustic observation vectors **a**
  - Use Bayes' rule to create a **generative model** and then decode
  - $\text{ArgMax}_{\mathbf{w}}\ P(\mathbf{w}|\mathbf{a}) = \text{ArgMax}_{\mathbf{w}}\ P(\mathbf{a}|\mathbf{w})\ P(\mathbf{w}) / P(\mathbf{a})$

    $= \text{ArgMax}_{\mathbf{w}}\ P(\mathbf{a}|\mathbf{w})\ P(\mathbf{w})$

- **Acoustic Model:** $P(\mathbf{a}|\mathbf{w})$
- **Language Model:** $P(\mathbf{w})$

A probabilistic theory of a language

Probabilistic Language Models

- Assign probability $P(\mathbf{w})$ to word sequence $\mathbf{w} = w_1, w_2, \ldots, w_k$
- Can't directly compute probability of long sequence – one needs to decompose it
- Chain rule provides a **history-based** model:

  $P(w_1, w_2, \ldots, w_k) =$

  $P(w_1)\ P(w_2|w_1)\ P(w_3|w_1,w_2) \cdots P(w_k|w_1,\ldots,w_{k-1})$
- **Cluster** histories to reduce number of parameters
- E.g., just based on the last word (1st order Markov model):

  $P(w_1, w_2, \ldots, w_k) = P(w_1|\langle s\rangle)\ P(w_2|w_1)\ P(w_3|w_2) \cdots P(w_k|w_{k-1})$

- How do we estimate these probabilities?
  - We count word sequences in corpora
  - We "smooth" probabilities so as to allow unseen sequences

Speech: the most natural UI

Although some might disagree…