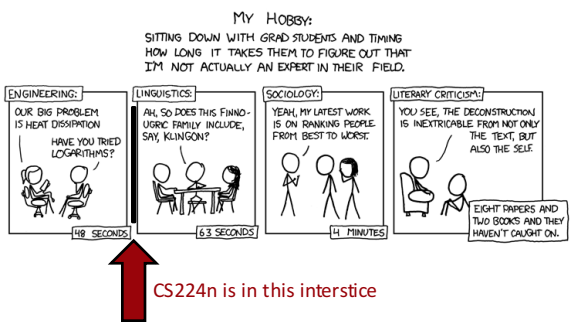


CS224n is in this interstice

cartoon from xkcd.com

Natural Language Processing CS224N/Ling284

Christopher Manning
Lecture 1

Lecture Plan

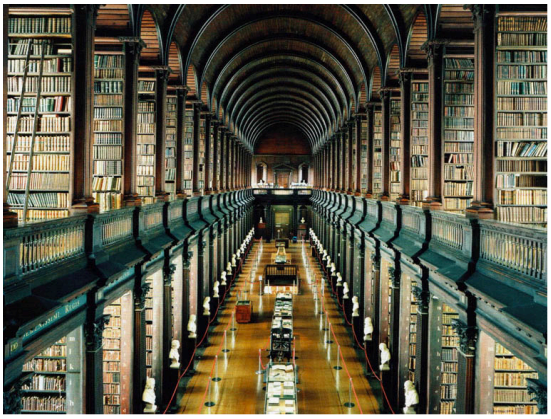
1. Human Language and Natural Language Processing: Their nature and goals (10 mins)
2. Why is language understanding hard? (15 mins)
3. Course logistics (5 mins)
4. Briefest of Introductions to Statistical NLP and Machine Translation (5 mins)
5. Translation Exercise: Learning to translate using parallel text (30 mins)

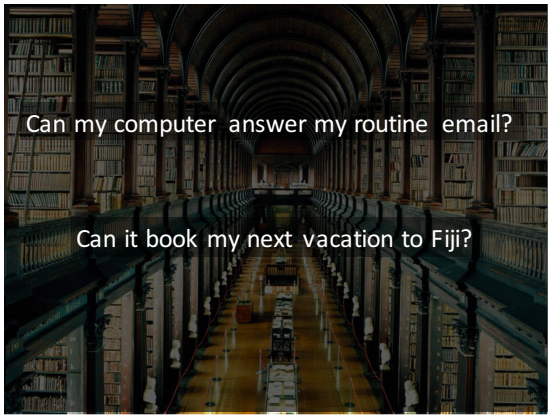
Emergency time reserves: 5 mins

Lecture Plan

1. **Human Language and Natural Language Processing: Their nature and goals (10 mins)**
2. Why is language understanding hard? (15 mins)
3. Course logistics (5 mins)
4. Briefest of Introductions to Statistical NLP and Machine Translation (5 mins)
5. Translation Exercise: Learning to translate using parallel text (30 mins)

Emergency time reserves: 5 mins





Siri

It knows what you mean.

Siri not only understands what you say, it's smart enough to know what you mean. So when you ask "Any good burger joints around here?" Siri will reply "I found a number of burger restaurants near you."

Siri

March 18, 2011 4:00 p.m.

Mentions of the Name 'Anne Hathaway' May Drive Berkshire Hathaway Stock

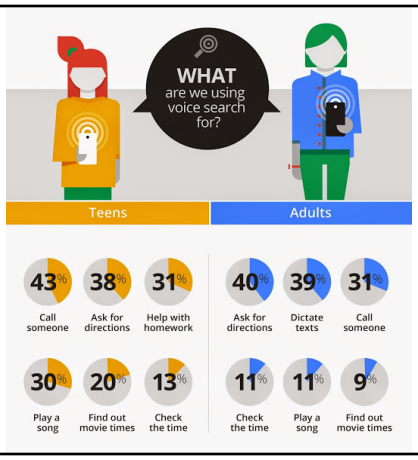
By **Patrick Huguenin**

The Huffington Post recently pointed out that whenever Anne Hathaway is in the news, the stock price for Warren Buffett's Berkshire Hathaway goes up. Really. When *Bride Wars* opened, the stock rose 2.61 percent.

Natural language: the earliest and still the best UI

Dave Bowman: Open the pod bay doors, HAL.
HAL: I'm sorry Dave. I'm afraid I can't do that.

(cf. also false Maria in Metropolis – 1926)



Source: Google study in October 2014.

http://googleblog.blogspot.com/2014/10/omg-mobile-voice-survey-reveals-teens.html

Language: still the ultimate UI

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

A Bug's Life is playing at the Century 16 Theater.

When is it playing there?

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

OK. I'd like 1 adult and 2 children for the first show. How much would that cost?

But we need domain knowledge, discourse knowledge, world knowledge, linguistic knowledge.

What's special about human language?

A human language is a system specifically **constructed to convey the speaker/writer's meaning**

- Which young kids can learn (**amazingly!**)

A human language is a **discrete/symbolic/categorical signaling system**

- rocket = 🚀 ; violin = 🎻
- With very minor exceptions for expressive signaling ("I loooove it." "Whoomppaaa")
- Symbols are not just an invention of logic / classical AI!

Why is it so?

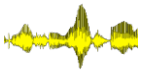
15

What's special about human language?

The categorical symbols of a language can be encoded as a signal for communication in several ways:

- Sound
- Gesture
- Images (writing)

The symbol is invariant is invariant across different encodings!



16

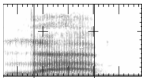
What's special about human language?

A human language is a **symbolic/categorical signaling system**

- Despite brain encoding as a continuous pattern of activation and transmission via continuous signals of sound/vision
- Presumably because of greater information-theoretic signaling reliability

Large vocabulary, symbolic encoding of words creates a problem for machine learning – sparsity!

lab



17

Lecture Plan

1. Human Language and Natural Language Processing: Their nature and goals (10 mins)
2. **Why is language understanding hard? (15 mins)**
3. Course logistics (5 mins)
4. Briefest of Introductions to Statistical NLP and Machine Translation (5 mins)
5. Translation Exercise: Learning to translate using parallel text (30 mins)

Emergency time reserves: 5 mins

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.

Why NLP is difficult: Newspaper headlines

1. Boy paralyzed after tumor fights back to gain black belt
2. San Jose cops kill man with knife

San Jose cops kill man with knife

Close

Text Paper Translate Listen

San Jose cops kill man with knife

Ex-college football player, 23, shot 9 times allegedly charged police at fiancee's home

By Hamed Aleaziz and Vivian Ho

A man fatally shot by San Jose police officers while allegedly charging at them with a knife was a 23-year-old former football player at De Anza College in Cupertino who was distraught and depressed, his family said Thursday.

Police officials said two officers opened fire Wednesday afternoon on Phillip Watkins outside his fiancee's home because they feared for their lives. The officers had been drawn to the home, officials said, by a 911 call reporting an armed home invasion that, it turned out, had been made by Watkins himself.

But the mother of Watkins' fiancee, who also lives in the home on the 1300 block of Sherman Street, said she witnessed the shooting and described it as excessive. Faye Buchanan said the confrontation happened shortly after she called a suicide intervention hotline in hopes of getting Watkins medical help.

Watkins' 911 call came in at 3:01 p.m., said Sgt. Heather Randol, a San Jose police spokeswoman. "The caller stated there was a male breaking into his home armed with a knife," Randol said. "The caller also stated he was locked in an upstairs bedroom with his children and requested help from police."

She said Watkins was on the sidewalk in front of the home when two officers got there. He was holding a knife with a 4-inch blade and ran toward the officers in a threatening manner, Randol said.

"Both officers ordered the suspect to stop and drop the knife," Randol said. "The caller also stated he was locked in an upstairs bedroom with his children and requesting for their safety and defense of their life, fired at the suspect."

On the police radio, one officer said, "We have a male with a knife. He's walking toward us."

"Shots fired! Shots fired!" an officer said moments later.

A short time later, an officer reported, "Male is down. Knife's still in hand."

Buchanan said she had been prompted to call the

Show continues on B8

Back Continue

Why is natural language understanding difficult

Fed raises interest rates 0.5% in effort to control inflation

- *NYT* headline, from better economic times (17 May 2000)

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

Lecture Plan

1. Human Language and Natural Language Processing: Their nature and goals (10 mins)
2. Why is language understanding hard? (15 mins)
3. **Course logistics (5 mins)**
4. Briefest of Introductions to Statistical NLP and Machine Translation (5 mins)
5. Translation Exercise: Learning to translate using parallel text (30 mins)

Emergency time reserves: 5 mins

Course logistics in brief

- Instructor: Christopher Manning
- TAs: Danqi Chen, Mihail Eric, Jade Huang, Neha Nayak, Ashwin Paranjape (maybe more)
- Time: TuTh 3:00–4:20, Skilling Aud

- The work is mainly big programming assignments
- Programming language: mainly or all Java
- Other information: see the class webpage
  - http://cs224n.stanford.edu/ a.k.a., http://www.stanford.edu/class/cs224n
- "Handouts": online

This class

- Assumes you come with some skills…
  - Some linear algebra, calculus, probability, and statistics; decent programming skills; know something about language
  - But not everyone has the same skills
    - Assumes some ability to learn missing knowledge, but beware PA4 NN
- 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
- 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

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.
- Deep Learning for NLP

- *(Unfortunately left out: natural language generation, phonology/morphology, speech dialogue systems, more on natural language understanding, …. There are other classes for some! cs224u/s)*

Lecture Plan

1. Human Language and Natural Language Processing: Their nature and goals (10 mins)
2. Why is language understanding hard? (15 mins)
3. Course logistics (5 mins)
4. **Briefest of Introductions to Statistical NLP and Machine Translation (5 mins)**
5. Translation Exercise: Learning to translate using parallel text (30 mins)

Emergency time reserves: 5 mins

Making progress on natural language processing …

- The task is difficult! What tools do we need?
  - Knowledge about language
  - Knowledge about the world
  - A way to combine knowledge sources
- The answer that got traction around 1990:
  - probabilistic models built from language data
    - P("maison" → "house") high
    - P("L'avocat général" → "the general avocado") low
- Many computer scientists think that this is applying "A.I." or "machine learning" to language … and it is
  - But it came from older ideas via electrical engineers….

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 classic acid test for natural language processing.

Requires capabilities in both interpretation and generation.

About $33 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!

Many slides from Kevin Knight (at ISI)

Statistical Solution

- Parallel Texts
  - Rosetta Stone

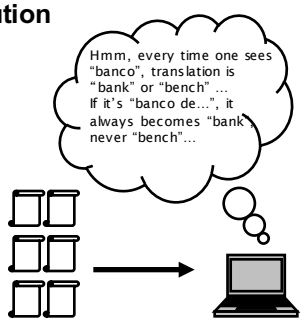
Hieroglyphs

Demotic

Greek

Statistical Solution

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



Lecture Plan

1. Human Language and Natural Language Processing: Their nature and goals (10 mins)
2. Why is language understanding hard? (15 mins)
3. Course logistics (5 mins)
4. Briefest of Introductions to Statistical NLP and Machine Translation (5 mins)
5. **Translation Exercise: Learning to translate using parallel text (30 mins)**

Emergency time reserves: 5 mins

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororoksprok . 1b. at-voon bichatdat.	7a. lalok farok ororok lalok sprok izok enemok . 7b. watjjat bichat wat datvat eneat .
2a. ok-drubelok-voon anok ploksprok . 2b. at-drubelat-voon pippatrratdat.	8a. lalok brok anok plok nok . 8b. iat lat pippatrratnnat.
3a. erok sprok izok hihok ghirok . 3b. totat dat arratvat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quatoloat at-yurp .
4a. ok-voon anok drokbrok jok . 4b. at-voon kratpippatsatlat .	10a. lalok mok nok yorok ghirokclok . 10b. watnnatgat mat bathilat .
5a. wiwok farok izok stok . 5b. totat jjat quatcat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. watnnatarrat matzanzanat .
6a. lalok sprok izok jok stok . 6b. watdat kratquatcat .	12a. lalok rarok nok izok hihok mok . 12b. watnnatforatarrat vatgat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororoksprok . 1b. at-voon bichatdat.	7a. lalok **farok** ororok lalok sprok izok enemok . 7b. watjjat bichat wat datvat eneat .
2a. ok-drubelok-voon anok ploksprok . 2b. at-drubelat-voon pippatrratdat.	8a. lalok brok anok plok nok . 8b. iat lat pippatrratnnat.
3a. erok sprok izok hihok ghirok . 3b. totat dat arratvat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quatoloat at-yurp .
4a. ok-voon anok drokbrok jok . 4b. at-voon kratpippatsatlat .	10a. lalok mok nok yorok ghirokclok . 10b. watnnatgat mat bathilat .
5a. wiwok **farok** izok stok . 5b. totat jjat quatcat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. watnnatarrat matzanzanat .
6a. lalok sprok izok jok stok . 6b. watdat kratquatcat .	12a. lalok rarok nok izok hihok mok . 12b. watnnatforatarrat vatgat .