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

Can my computer answer my routine email?
Can it book my next vacation to Fiji?
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.”

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)

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.

Language: still the ultimate UI

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?

---

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!

---

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

---

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
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

**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
- "Handouts": online

**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, etc. There are other classes for some of these topics too! 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

**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 PA/N 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
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...

The classic acid test for natural language processing.

Requires capabilities in both interpretation and generation.

About $33 billion spent annually on human translation!

Many slides from Kevin Knight (at ISI)

Machine Translation

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

Many slides from Kevin Knight (at ISI)

Statistical Solution

- Parallel Texts
- Rosetta Stone

Hieroglyphs
Demotic
Greek

Hmm, every time one sees “bank” or “bench”...

If it’s “banco de...”, it always becomes “bank” never “bench”...

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
Your assignment, translate this to Arcturan: farok crrr ok  hi hok  yor ok  cl ok  kan t ok  ok  -  yu rp