Administrivia

- Status report 1 due tomorrow
- Look at written problems 3 by Monday
  - Recall we do have quizzes for this class
- Exam 2 will be available early next week
- Upcoming schedule
  - Today: NLP
  - Tuesday: Robotics
  - Thursday: Computer Vision
  - Last week: Philosophy and ethics of AI

What is NLP?

Natural language processing (NLP) is a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages.

- Wikipedia
What is NLP?

The goal of this new field is to get computers to perform useful tasks involving human language...

- Dan Jurafsky

Key: Natural text

“A growing number of businesses are making Facebook an indispensable part of hanging out their shingles. Small businesses are using ...”

Natural text is written by people, generally for people

Why do we even care about natural text in computer science?

Why do we need computers for dealing with natural text?

Why do we need computers?
Web is just the start...

- e-mail
  - 247 billion e-mails a day
- corporate databases
- Twitter
  - 27 million tweets a day
- Blogs: 126 million different blogs

Why is NLP hard?

- Iraqi Head Seeks Arms
- Juvenile Court to Try Shooting Defendant
- Stolen Painting Found by Tree
- Kids Make Nutritious Snacks
- Local HS Dropouts Cut in Half
- Obesity Study Looks for Larger Test Group
- British Left Waffles on Falkland Islands
- Red Tape Holds Up New Bridges
- Hospitals Are Sued by 7 Foot Doctors

Why is NLP hard?

User: Where is Contagion playing in the Middlebury Area?
System: Contagion is playing at the Marquis Theatre.
User: When is it playing there?
System: It’s playing at 2pm, 5pm and 8pm.
User: I’d like 1 adult and 2 children for the first show. How much would that cost?

Why is NLP hard?

I saw her duck
Why is NLP hard?

I saw her duck

Natural language:
- is highly ambiguous at many different levels
- is complex and contains subtle use of context to convey meaning
- is probabilistic?
- involves reasoning about the world
- is highly social
- is a key part in how people interact

However, some NLP problems can be surprisingly easy

Different levels of NLP

pragmatics/discourse: how does the context affect the interpretation?

semantics: what does it mean?

syntax: phrases, how do words interact

words: morphology, classes of words

NLP problems and applications

What are some places where you have seen NLP used?

What are NLP problems?
Lots of problems of varying difficulty

Easier

Word segmentation: where are the words?

I would’ve liked Prof. Kauchak to finish early. But he didn’t.

Easier continued

- Speech segmentation
- Sentence splitting (aka sentence breaking, sentence boundary disambiguation)
  - I would’ve liked Prof. Kauchak to finish early. But he didn’t.
- Language identification
  - Soy un maestro con queso.
NLP problems and applications

Moderately difficult

- morphological analysis/stemming
  - smarter
  - smarterly
  - smartest
  - smart
- speech recognition
- text classification

NLP problems and applications

moderately difficult continued

- text segmentation: break up the text by topics
- part of speech tagging (and inducing word classes)
- parsing

As he walked along the side of the stream, he spotted some money by the bank. The money had gotten muddy from being so close to the water.

We are good at grammar.

NLP problems and applications

Hard (many of these contain many smaller problems)

- 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 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.
NLP problems and applications

Information extraction

IBM hired Fred Smith as president.

<table>
<thead>
<tr>
<th>person</th>
<th>company</th>
<th>position</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fred Smith</td>
<td>IBM</td>
<td>president</td>
</tr>
</tbody>
</table>

Summarization

A company that acts as a middle-man between content providers and Internet service providers is accusing Comcast Corp., the nation’s largest broadband provider, of anti-competitive behavior. (article 1) Comcast Corp. and NBC Universal made a new promise to the Federal Communications Commission that the company hopes will help get the regulatory agency to approve the proposed deal between the media giants. (article 2) At issue is the cable operator’s decision to offer the Tennis Channel on a specialty sports networks as opposed to its widely distributed basic tier. (article 3) The quality of television news could deteriorate further under a Comcast-controlled NBC Universal, the Writers Guild of America has warned. Wednesday in letters to key Washington officials overseeing the government’s review of the proposed merger. (article 4) With regulatory approval still weeks if not months away, Comcast and NBC Universal have extended the terms of their merger agreement to March of next year. (article 5) Disney’s Michael Cogswell, the point person, would not say too much about content and the terms of a company that also controls how consumers access the Internet and television. (article 6) Isaac Fox added on Wednesday with two senior staff members of FCC Commissioners Meredith Artwell Baker (article 7)

Natural language understanding

- Text => semantic representation (e.g. logic, probabilistic relationships)

Information retrieval and question answering

- “How many programmers in the child care department make over $50,000?”
- “Who was the fourteenth president?”
- “How did he die?”

Text simplification

Alfonso Perez Munoz, usually referred to as Alfonso, is a former Spanish footballer, in the striker position.

Alfonso Perez is a former Spanish football player.
Where are we now?

Many of the “easy” and “medium” problems have reasonable solutions:
- spell checkers
- sentence splitters
- word segmenters/tokenizers

Parsing
- Stanford Parser (http://nlp.stanford.edu:8080/parser/)

Machine translation
- Getting better every year
- enough to get the jist of most content, but still no where near a human translation
- better for some types of text

http://translate.google.com

Many commercial versions…
- systran
- language weaver

Information extraction
- Structured documents (very good!)
  - www.dealtime.com
  - www.froogle.com
- AIT technologies
  - Lots of these
    - FlipDog
    - WhizBang! Labs
    - ... 
  - work fairly well
Where are we now?
CMU’s NELL (Never Ending Language Learner)
http://rtw.ml.cmu.edu/rtw/

Where are we now?
Information retrieval/query answering
- search engines:
  - pretty good for some things
  - does mostly pattern matching and ranking
  - no deep understanding
  - still requires user to “find” the answer

Where are we now?
Question answering
- Wolfram Alpha

Where are we now?
Question answering
- Wolfram Alpha
Question answering

Where are we now?

Question answering

- Many others...
  - TREC question answering competition
  - language computer corp
  - answerbus
  - ...

Where are we now?

Summarization

- NewsBlaster (Columbia)
- http://newsblaster.cs.columbia.edu/

Where are we now?

Voice recognition

- pretty good, particularly with speaker training
- Apple OS has one built in:
  - “What time is it?”
  - “Switch to finder”
  - “Hide this application”
- IBM ViaVoice
- Dragon Naturally Speaking

Speech generation

- The systems can generate the words, but getting the subtle nuances right is still tricky
- Apple OS
- translate.google.com
A combination of problems…

- Many problems untackled/undiscovered
- “That’s What She Said: Double Entendre Identification”
- ACL 2011

Other problems

Language translation

Where have you seen machine translation systems?

Language translation

Yo quiero Taco Bell
The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an email from someone claiming to be the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport.

Machine translation is becoming very prevalent

Even PowerPoint has translation built into it!

The Beijing Youth Daily pointed out that under the Ministry of Agriculture, the beef would be disposed of after being examined according to advice from the Ministry of Agriculture.
Which is the human?

Pakistan President Pervez Musharraf Wins Senate Confidence Vote

There was not a single vote against him.

No members vote against him.

Data-Driven Machine Translation

Man, this is so boring.

Hello, every time he sees “banco”, he either types “bank” or “bench”... but if he sees “banco de...” he always types “bank” never “bench”...

Welcome to the Chinese Room

You can teach yourself to translate Chinese using only bilingual data (without grammar books, dictionaries, any people to answer your questions...)

Translated documents
Your assignment, translate this to Arcturan:

1a. ok-voon ororok sprok .
1b. at-voon bichat dat .
2a. ok-drubel ok-voon anok plok sprok .
2b. at-drubel at-voon pippat rrat dat .
3a. avok sprok izok hihok ghirok .
3b. totat dat ardat vat bulat .
4a. avok sprok izok kokk jok .
4b. at-voon krat pippat sat lat .
5a. wiwok farok izok stok .
5b. totat jjat quat cat .
6a. lalok sprok izok jok stok .
6b. wat dat kratquat cat .
7a. lalok farok ororok sprok izok enemok .
7b. wat jat bichat wat dat vat anmat .
8a. lalok brok anok plok nok .
8b. iat lat pippat rrat mat .
9a. wiwok nek izok kantok ok-yarp .
9b. wat mat mat mat bat hilat .
10a. lalok mok nok yorok ghirok clok .
10b. wat nnat gat mat bat hilat .
11a. lalok nok crrrok hihok yorok zanzanok .
11b. wat nnat arrat mat zanzanat .
12a. lalok nok kokk kokk mek .
12b. wat mat furat ardat vat gat .

Your assignment, translate this to Arcturan:

farok crrrok hihok yorok clok kantok ok-yarp

Centauri/Arcturan [Knight, 1997]
Your assignment, translate this to Arcturan:

1a. ok-voon ororok sprok .
1b. at-voon bichat dat .
2a. ok-drubel ok-voon anok plok sprok .
2b. at-drubel at-voon pippat rrat dat .
3a. avok sprok izok hihok ghirok .
3b. totat dat arat vat hilat .
4a. ok-voon anok brok jok .
4b. at-voon krat pippat sat lat .
5a. wiwok farok izok stok .
5b. totat jat quit cat .
6a. lalok sprok izok jok stok .
6b. wat dat krat quit cat .
7a. lalok farok ororok lalok sprok izok enemok .
7b. wat jat bichat vat dat vat enemat .
8a. lalok brok anok plok nok .
8b. iat lat pippat rrat matat .
9a. wiwok nok izok hihok ghirok .
9b. totat dat arat vat hilat .
10a. lalok mok nok yorok ghirok clok .
10b. wat nnat gat mat bat hilat .
11a. lalok nok crrrok hihok yorok zanzanok .
11b. wat nnat arrat mat zanzanat .
12a. lalok mok nok izok hihok ghirok .
12b. wat nnat forat arrat vat gat .

Your assignment, translate this to Arcturan:

1a. ok-voon ororok sprok .
1b. at-voon bichat dat .
2a. ok-drubel ok-voon anok plok sprok .
2b. at-drubel at-voon pippat rrat dat .
3a. avok sprok izok hihok ghirok .
3b. totat dat arat vat hilat .
4a. ok-voon anok brok jok .
4b. at-voon krat pippat sat lat .
5a. wiwok farok izok stok .
5b. totat jat quit cat .
6a. lalok sprok izok jok stok .
6b. wat dat krat quit cat .
7a. lalok farok ororok lalok sprok izok enemok .
7b. wat jat bichat vat dat vat enemat .
8a. lalok brok anok plok nok .
8b. iat lat pippat rrat matat .
9a. wiwok nok izok hihok ghirok .
9b. totat dat arat vat hilat .
10a. lalok mok nok yorok ghirok clok .
10b. wat nnat gat mat bat hilat .
11a. lalok nok crrrok hihok yorok zanzanok .
11b. wat nnat arrat mat zanzanat .
12a. lalok mok nok izok hihok ghirok .
12b. wat nnat forat arrat vat gat .
+ Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan:

Your assignment, translate this to Arcturan:  

Your assignment, put these words in order:  

{jjat, arrat, mat, bat, oloat, at-yurp}

Your assignment, translate this to Arcturan:  

Your assignment, translate this to Arcturan:  

Your assignment, put these words in order:  

{jjat, arrat, mat, bat, oloat, at-yurp}
### It’s Really Spanish/English

<table>
<thead>
<tr>
<th>English</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clients do not sell pharmaceuticals in Europe</td>
<td>Clientes no venden medicinas en Europa</td>
</tr>
</tbody>
</table>

1a. Garcia and associates.  
1b. Garcia y asociados.  
7a. the clients and the associates are enemies.  
7b. los clientes y los asociados son enemigos.  

2a. Carlos Garcia has three associates.  
2b. Carlos Garcia tiene tres asociados.  
8a. the company has three groups.  
8b. la empresa tiene tres grupos.  

3a. his associates are not strong.  
3b. sus asociados no son fuertes.  
9a. its groups are in Europe.  
9b. sus grupos estan en Europa.  

4a. Garcia has a company also.  
4b. Garcia tambien tiene una empresa.  
10a. the modern groups sell strong pharmaceuticals.  
10b. los grupos modernos venden medicinas fuertes.  

5a. its clients are angry.  
5b. sus clientes estan enfadados.  
11a. the groups do not sell zenzanine.  
11b. los grupos no venden zenzanine.  

6a. the associates are also angry.  
6b. los asociados tambien estan enfadados.  
12a. the small groups are not modern.  
12b. los grupos pequenos no son modernos.  

---

### Warren Weaver (1947)

- ingcmpnqsnwf cv fpn owoktvcv
- hu ihgznwfv rqcffnw cw owgcnwf
- kowazoany ...
Warren Weaver (1947)

Decipherment is the analysis of documents written in ancient languages.

Can this be computerized?

Maybe this can be done for translation of languages. All I need is a pair-wise word frequencies between two languages...
Noisy channel model

Some message is sent along the way the message gets changed/mutated. What was originally sent?

We have the mutated message, but would like to recover the original.

Noisy channel model: \( p(\text{sent} | \text{received}) \)

Probabilistic model: \( p(s | r) \)

- \( p(\text{English} | \text{Foreign}) \)
- \( p(\text{English} | \text{speech}) \)
- \( p(\text{simplified} | \text{un simplified}) \)

Given sentence pairs, gives us the probability.

Noisy channel model: training

Input: aligned sentence pairs

\[
\begin{array}{c|c}
\text{Chinese sentence fragment} & \text{English sentence fragment} \\
\hline
\hline
\hline
\hline
\hline
\end{array}
\]

Learning \( p(s | r) \)
Noisy channel model: applying

Conditioned on the input, what is the most-likely output

\[ \arg \max_s p(s | r) \]

Bayes' rule

\[ p(s | r) = \frac{p(r | s) p(s)}{p(r)} \]

- \( p(s) \): language model: what are likely word sequences?
- \( p(r) \): probability of the received message
- \( p(r | s) \): translation model: how does the mutation/translation process happen? what operations are valid?
- \( p(s | r) \): probability of the received message

Bayes' rule

\[ p(s | r) \propto p(r | s) p(s) \]

why?

- \( p(r) \): probability of the received message
- \( p(s) \): language model: what are likely word sequences?
- \( p(r | s) \): translation model: how does the translation process happen? what operations are valid?
**Machine translation**

$p(s \mid t) \propto p(t \mid s)p(s)$

- **Channel model**
- **Translation model**
- **Language model**

How do English words/phrases translate to Chinese?

What are likely English words/word sequences?

**One way to think about it...**

Que hambre tengo yo → What hunger have I, Hungry I am so, I am so hungry, Have I that hunger …

**Data available**

- Many languages
  - Europarl corpus has all European languages
  - [http://www.statmt.org/europarl/](http://www.statmt.org/europarl/)
  - French/English from French parliamentary proceedings
  - Lots of Chinese/English and Arabic/English from government projects/interests
  - Chinese-English: 440 million words (15-20 million sentence pairs)
  - Arabic-English: 790 million words (30-40 million sentence pairs)
  - Smaller corpora in many, many other languages
- Lots of monolingual data available in many languages
- Even less data with multiple translations available
- Available in limited domains
  - Most data is either news or government proceedings
  - Some other domains recently, like blogs

**Statistical MT Overview**

- Find the best translation given the foreign sentence and the model
- English sentence
- Foreign sentence
- Bilingual data
- Monolingual data
- Trained parameters
- Model
Language modeling

Answers the question of how likely a sentence is to be an English sentence

I think today is a good day to be me

Must also worry about the ordering of the words

How likely is “I think” to be preceded by “today is a good day”?

The main challenge with language modeling is dealing with data sparsity
**Language modeling**

Most common: n-gram language models

\[ p(\text{word} \mid \text{previous words}) \]

More data the better (Google n-grams)

Domain is important!

**Translation models: MT Pyramid**

Morgen fliege ich nach Kanada zur Konferenz
Phrase-Based Statistical MT

Foreign input segmented into phrases
- “phrase” is any sequence of words

Example: Morgen fliege ich nach Kanada zur Konferenz
- Tomorrow I will fly to the conference in Canada

Phrase-Based Statistical MT

Phrases are probabilistically re-ordered

Advantages of Phrase-Based

Many-to-many mappings can handle non-compositional phrases
- Easy to understand
- Local context is very useful for disambiguating
  - “Interest rate” → ...
  - “Interest in” → ...
- The more data, the longer the learned phrases
  - Sometimes whole sentences
+Available Resources

- Bilingual corpora
  - Large words of Chinese/English and Arabic/English, LDC (www.ldc.upenn.edu)
  - Lots of French/English, Spanish/French, LDC
  - European Parliament (sentence-aligned), 11 languages, Philip Koehn, ISI
  - Small words (sentence-aligned) of English/Spanish, Eric Germann, ISI
    (www.isi.edu/natural-language/downlift.html)
- Sentence alignment
  - Dan Melamed, NYU (www.cs.nyu.edu/~melamed/GMA/docs/README.html)
  - Xiaoyi Ma, LDC (Champion)
- Word alignment
  - GIZA++, JHU Workshop '99
  - GIZA++, RWTH Aachen
  - Manually word-aligned test corpus (500 French/English sentence pairs), RWTH Aachen
  - Shared task, NAACL-HLT'03 workshop
- Decoding
  - IS ReWrite Model 4 decoder (www.isi.edu/licensed-sw/rewrite-decoder/)
  - IS Pharaoh phrase-based decoder
- Statistical MT Tutorial Workbook, ISI (www.isi.edu/~knight/)
- Annual common-data evaluation, NIST (www.nist.gov/speech/tests/mt/index.htm)

+Some Papers Referenced on Slides

- ACL
  - [Och, Tillmann, & Ney, 1999]
  - [Och & Ney, 2000]
  - [Germann et al, 2001]
  - [Yamada & Knight, 2001, 2002]
  - [Papineni et al, 2002]
  - [Alshawi et al, 1998]
  - [Collins, 2002]
  - [Koehn & Knight, 2003]
  - [Och & Ney, 2003]
  - [Och, 2002]
  - [Koehn et al, 2003]
- EMNLP
  - [Marcu & Wong, 2002]
  - [Tan, 2002]
  - [Murez & Marcu, 2002]
- AI Magazine
  - [Knight, 1997]
  - [www.aij.na.ac.jp/~knight/]
  - [MT Tutorial Workbook]