Computational Linguistics

Today, we explore this interstice
Computational Linguistics
(aka Natural Language Processing)

Christopher Manning
Ling 1
November 4, 2011
What is Computational Linguistics?

- Getting computers to perform useful tasks involving human languages whether for:
  - Enabling human-machine communication
  - Improving human-human communication
  - Doing stuff with language data … email, blogs, etc.

- Examples:
  - Machine Translation
  - Automatic Question Answering
  - Speech Recognition
  - Text-to-Speech Synthesis
  - Text Understanding
• Why say “Computational Linguistics (CL)” versus “Natural Language Processing” (NLP)?

• Either choose either freely or …

• Computational Linguistics
  • The science of computers dealing with language
  • Some interest in modeling what people do

• Natural Language Processing
  • Developing computer systems for processing and understanding human language text
I’m sorry, Dave. I can’t do that.

Oh, dear!

That is correct, captain.
Language: 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 (Not to mention linguistic knowledge!)
NLP: Goals of the field

• From the lofty ...
  • full-on natural language understanding
  • participation in spoken dialogues
  • open-domain question answering
  • real-time bi-directional translation

• ... to the mundane
  • identifying spam
  • categorizing news stories
  • finding & comparing product information on the web
  • assessing sentiment toward products, brands, stocks, ...

Predominant in recent years
NLP in the commercial world
Current motivations for NLP

What’s driving NLP? Three trends:

• The explosion of machine-readable natural language text
  • Exabytes (10^{18} bytes) of text, doubling every year or two
  • Web pages, emails, IMs, SMSs, tweets, docs, PDFs, ...
  • Opportunity — and increasing necessity — to extract meaning

• Mediation of human interactions by computers
  • Opportunity for the computer in the loop to do much more

• Growing role of language in human-computer interaction
Further motivation for CL

One reason for studying language — and for me personally the most compelling reason — is that it is tempting to regard language, in the traditional phrase, as a “mirror of mind”.

Chomsky, 1975

For the same reason, computational linguistics is a compelling way to study human language acquisition and processing.

Sometimes, the best way to understand something is to build a model of it.

What I cannot create, I do not understand.  Feynman, 1988
### Subfields and tasks

<table>
<thead>
<tr>
<th>Mostly solved</th>
<th>Making good progress</th>
<th>Still really hard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spam detection</td>
<td>Sentiment analysis</td>
<td>Semantic search</td>
</tr>
<tr>
<td>OK, let’s meet by the big ...</td>
<td>The pho was authentic and yummy.</td>
<td>people protesting globalization</td>
</tr>
<tr>
<td>Dick too small? Buy V1AGRA ...</td>
<td>Waiter ignored us for 20 minutes.</td>
<td>Search</td>
</tr>
<tr>
<td>Text categorization</td>
<td>Coreference resolution</td>
<td>Question answering (QA)</td>
</tr>
<tr>
<td>Phillies shut down Rangers 2-0</td>
<td>Obama told Mubarak he shouldn’t run again.</td>
<td>Q. What currency is used in China?</td>
</tr>
<tr>
<td>Jobless rate hits two-year low</td>
<td></td>
<td>A. The yuan</td>
</tr>
<tr>
<td>Part-of-speech (POS) tagging</td>
<td>Word sense disambiguation (WSD)</td>
<td>Textual inference &amp; paraphrase</td>
</tr>
<tr>
<td>ADJ ADJ NOUN VERB ADV</td>
<td>I need new batteries for my mouse.</td>
<td>T. Thirteen soldiers lost their lives ...</td>
</tr>
<tr>
<td>Colorless green ideas sleep furiously.</td>
<td></td>
<td>H. Several troops were killed in the ...</td>
</tr>
<tr>
<td>Named entity recognition (NER)</td>
<td>Syntactic parsing</td>
<td>Summarization</td>
</tr>
<tr>
<td>PERSON ORG LOC</td>
<td>I can see Russia from my house!</td>
<td>Sheen continues rant against ...</td>
</tr>
<tr>
<td>Obama met with UAW leaders in Detroit ...</td>
<td></td>
<td>Sheen is nuts</td>
</tr>
<tr>
<td>Information extraction (IE)</td>
<td>Machine translation (MT)</td>
<td>Discourse &amp; dialog</td>
</tr>
<tr>
<td>You’re invited to our bunga bunga party, Friday May 27 at 8:30pm in Cordura Hall</td>
<td>Our specialty is panda fried rice.</td>
<td>Where is Thor playing in SF?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Metreon at 4:30 and 7:30</td>
</tr>
</tbody>
</table>
Why is computational linguistics hard?

Human languages:

• are highly ambiguous at all levels
• are complex, with recursive structures and coreference
• subtly exploit context to convey meaning
• are fuzzy and vague
• require reasoning about the world for understanding
• are part of a social system: persuading, insulting, amusing, ...

(Nevertheless, simple features often do half the job!)
OK, why else is NLP hard?

Oh so many reasons!

**non-standard English**
Great job @justinbieber! Were SOO PROUD of what you’ve accomplished! U taught us 2 #neversaynever & you yourself should never give up either❤

**segmentation issues**
- the New York-New Haven Railroad
- the New York-New Haven Railroad

**idioms**
- dark horse
- get cold feet
- lose face
- throw in the towel

**neologisms**
- unfriend
- retweet
- bromance
- teabagger

**garden path sentences**
The man who hunts ducks out on weekends.
The cotton shirts are made from grows here.

**tricky entity names**
... a mutation on the for gene ...
Where is A Bug’s Life playing ...
Most of Let It Be was recorded ...

**world knowledge**
- Mary and Sue are sisters.
- Mary and Sue are mothers.

**prosody**
- I never said she stole my money.
- I never said she stole my money.

**lexical specificity**

But that’s what makes it fun!
Meanings and expressions

- soda
- soft drink
- pop
- beverage
- Coke
One meaning, many expressions

Consider a semantic search application:

<table>
<thead>
<tr>
<th>Russia increasing price of gas for Georgia</th>
<th>Search</th>
</tr>
</thead>
<tbody>
<tr>
<td>Russia hits Georgia with huge rise in its gas bill</td>
<td></td>
</tr>
<tr>
<td>Russia plans to double Georgian gas price</td>
<td></td>
</tr>
<tr>
<td>Russia gas monopoly to double price of gas</td>
<td></td>
</tr>
<tr>
<td>Gazprom confirms two-fold increase in gas price for Georgia</td>
<td></td>
</tr>
<tr>
<td>Russia doubles gas bill to “punish” neighbour Georgia</td>
<td></td>
</tr>
<tr>
<td>Gazprom doubles Georgia's gas bill</td>
<td></td>
</tr>
</tbody>
</table>
One expression, many meanings
Syntactic & semantic ambiguity

Fruit flies like a banana

photos from worth1000.com
Ambiguous headlines

Teacher Strikes Idle Kids
China to Orbit Human on Oct. 15
Juvenile Court to Try Shooting Defendant
Clinton Wins on Budget, but More Lies Ahead
Local High School Dropouts Cut in Half
Police: Crack Found in Man’s Buttocks
• Parsing is the process of taking a string and a grammar and returning a parse tree or trees for that string.

The flight left

S
  / | 
 NP VP
  / | 
 Det N Verb
   / | 
   the flight left
Parsing involves search

- As with most everything of interest, parsing involves a search which involves the making of choices
- We’ll look at some basic methods to give you an idea of the problem
Top-Down Parsing

• Since we’re trying to find trees rooted with an S (Sentence) start with the rules that give us an S.

• Then work your way down from there to the words.
Top-Down Space
Bottom-Up Parsing

- Of course, we also want trees that cover the input words. So start with trees that link up with the words in the right way.
- Then work your way up from there.
Bottom-Up Space

Book that flight

Book that flight

Book that flight

Book that flight

Book that flight
Control

• We need to keep track of the search space and have a strategy to make choices
  • We need to systematically explore everything to make sure we find the right parse for a sentence
    • Which node to try to expand next?
    • Which grammar rule to use to expand a node?
Top-Down, Depth-First, Left-to-Right Search
Example

[Does]  Does [this]  Does [this]

AUX  NP  VP  AUX  NP  VP  AUX  NP  VP

Det  Nom
Example

```
S
  /\   /
AUX NP VP
  /\   /
Det Nom  
  Does  this [flight]
  /\   /
Det Nom  
  Does  this [flight]
  /\   /
Det  Noun
  Does  this [flight]
```
Example

```
S
├── AUX
│   ├── Det
│   │   └── Does
│   └── NP
│       └── flight [include]
└── VP
      └── Noun

S
├── AUX
│   ├── Det
│   │   └── Does
│   └── NP
│       └── flight [include]
└── VP
      └── Verb
      └── Noun
```
Efficient parsing

• That should give you the general idea of how a parser can work by exploring hypotheses systematically

• But really we need to do much more to make parsing efficient ... this leads into dynamic programming, memoization, and other tricky stuff that I won’t mention further here.
How to choose between parses?

• Probabilistic methods!

• Augment the grammar rules with probabilities

• Modify the parser to keep only most probable parses

• At the end, return the most probable parse
A statistical scientific revolution

• Computational Linguistics before 1990:
  • Hand-built parsers, hand-built dialogue systems
  • High precision, low coverage methods

• Computational Linguistics after 1995:
  • Automatically trained parsers, unsupervised clustering, statistical machine translation
  • High coverage, low precision methods
  • Build models exploiting data
Demos!
If you might like NLP / CompLing ...

• learn Java or Python (and play with [JavaNLP](https://javannlp.org) or [NLTK](https://nltk.org))
• get some exposure to linguistics (LING1, ...)
• and to logic, probability, statistics, linear algebra
• study AI and machine learning (CS121, CS221, CS229)
• read [Jurafsky & Martin](https://nlp.stanford.edu/tales) or [Manning & Schütze](https://nlp.stanford.edu/tales)
• Take
  • Ling 180/CS124: From Languages to Information
  • Ling 284/CS224N: Natural Language Processing
  • Ling 281/CS224S: Speech Recognition & Synthesis
  • Ling /CS224U: Natural Language Processing
One more for the road

Natural language processing researchers are all, “oh boo hoo, computationally generating and understanding natural language is hard, we totally need more time to build Data from Star Trek, boo hoo hoo!”

Looks like you dropped the ball, NLP researchers!!

Because MY email autoresponder takes in text with all sorts of imperfect and irregular language, disambiguates word senses, builds a semantic understanding AND generates an appropriate response in the time it would take a real person to type it out. Oh snap! Did a rank amateur, wanting only to ensure that his body wouldn't be discovered for a few years after he died, just brutally advance the state of the art? SEEMS LIKE IT, BABY!

So how’s it work?

Just fine, thank you!

No, I mean, how'd you do it? How do you figure out the semantics and word senses? Statistical methods? But those have problems with recall and precision, even with - are you using the web as a corpus? How are you handling the noise?

Utahraptor, please! So many questions!

I just wrote "Email Responder TWO THOUSAND" on a giant novelty chef's hat, taped a page from the dictionary to it, and then stuffed the whole deal into my computer's CD tray!

w- what?

GRANT MONEY PLEASE

cartoon from qwantz.com