Lecture 8

Natural Language Processing

Introduction
• Introduction to Natural Language Processing (NLP)
  • What is NLP
  • Applications of NLP
  • Why NLP is hard
  • Brief history of NLP
• Linguistic Essentials
Natural Language Processing

• Computers would be more useful if they could handle our email, do our library research, talk to us, etc ...

• But computers are fazed by natural human language
  • or at least their programmers are, most avoid the language problem by using mice, menus, drop boxes

• How can we tell computers about language?
  • or help them learn it as kids do?

• Can machines understand human language?
  • define ‘understand’
  • understanding is the ultimate goal
  • however, one doesn’t need to fully understand to be useful

• NLP is also known as Computational Linguistics (CL), Human Language Technology (HLT), Natural Language Engineering (NLE)
IBM’s Watson Won Jeopardy on February 16, 2011!

WILLIAM WILKINSON’S “AN ACCOUNT OF THE PRINCIPALITIES OF WALLACHIA AND MOLDOVIA” INSPIRED THIS AUTHOR’S MOST FAMOUS NOVEL

Bram Stoker (Dracula)
Hi Dan, we’ve now scheduled the curriculum meeting. It will be in Gates 159 tomorrow from 10:00-11:30.

-Chris
• nice and compact to carry!

• since the camera is small and light, I won't need to carry around those heavy, bulky professional cameras either!

• the camera feels flimsy, is plastic and very light in weight you have to be very delicate in the handling of this camera
Application: Machine Translation

• Fully automatic

Enter Source Text:

这不过是一个时间的问题。

Translation from Stanford’s Phrasal:

This is only a matter of time.
Where is Language Technology

• Goals can be very far reaching
  • True text understanding and interpretation
  • Real-time participation in spoken dialogs
  • High quality machine translation

• Or very application oriented
  • Finding the price of products on the web
  • Analyzing reading level or authorship statistically
  • Sentiment detection about products or stocks
  • Extracting names, facts or relations from documents

• These days, the latter predominate
  • As NLP becomes increasingly possible, it becomes increasingly engineering-oriented
Where is Language Technology

mostly solved

Spam detection
- Let's go to Agra!
- Buy V1AGRA ...

Part-of-speech (POS) tagging
- ADJ  ADJ  NOUN  VERB  ADV
- Colorless  green  ideas  sleep  furiously.

Named entity recognition (NER)
- PERSON  ORG  LOC
- Einstein met with UN officials in Princeton

making good progress

Sentiment analysis
- Best roast chicken in San Francisco!
- The waiter ignored us for 20 minutes.

Coreference resolution
- Carter told Mubarak he shouldn't run again.

Word sense disambiguation (WSD)
- I need new batteries for my mouse.

Parsing
- I can see Alcatraz from the window!

Machine translation (MT)
- 第13届上海国际电影节开幕...
- The 13th Shanghai International Film Festival...

Information extraction (IE)
- You're invited to our dinner party, Friday May 27 at 8:30

still really hard

Question answering (QA)
- Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase
- XYZ acquired ABC yesterday
- ABC has been taken over by XYZ

Summarization
- The Dow Jones is up
- The S&P500 jumped
- Housing prices rose
- Economy is good

Dialog
- Where is Citizen Kane playing in SF?
- Castro Theatre at 7:30. Do you want a ticket?
• 1950’s, empirical approach:
  • data-driven, co-occurrences in language are important sources of information: “You shall know a word by the company it keeps”, J. Firth, 1957
  • First speech systems (Davis et al. Bell labs)
  • Text authorship (Hamilton vs. Madison), solved based on patterns of word occurrences in 1941 by F. Mosteller and F. Williams
  • Machine translation: toy system, basically word-substitution, on machines less powerful than pocket calculators
  • Little understanding of natural language syntax and semantics
  • Problem soon appeared intractable: can’t store enough data on computers
Brief NLP History

- 1960’s and 1970’s
  - Data-driven approach falls out of favor
  - Language is to be analyzed at deeper level than surface statistics
  - N. Chomsky:
    1. “Colorless green ideas sleep furiously”
    2. “Furiously sleep ideas green colorless”
    - Neither (1) nor (2) will never occur. Yet (1) is grammatical, while (2) is not. Therefore (1) should have higher probability of occurrence than (2)
    - However, since neither (1) nor (2) will ever occur, they will both be assigned the same probability of 0
    - The criticism is that the data driven approach will always lack suffer from the lack of data, and therefore doomed to failure
  - Knowledge-based (rule based) approach becomes dominant, human expert encodes relevant information
    - Development of linguistic
    - Complex language models, parsing, CF grammars
    - Applications in toy domains
• **Drawbacks of knowledge-based (rule-based) approach:**
  • Rules are often too strict to characterize people’s use of language (people tend to stretch and bend rules in order to meet their communicative needs.)
  • Need expert people to develop rules (knowledge acquisition bottleneck)

• **1980’s: the empirical revolution**
  • In part motivated by success in speech recognition
    • Based on learning from lots of data
  • Corpus-based (data-driven) methods become central
  • Sophisticated machine learning algorithms are developed to learn from the data
  • Linguistics (the rules) is still used
  • Deep analysis often traded for robust and simple approximations
Why is NLP difficult?

• Key problem: language is **ambiguous** at all levels
  • Semantic (word meaning)
  • Syntactic (sentence structure)
  • Acoustic (parsing of speech signal)

• To resolve these ambiguities we often need to use complex knowledge about the world

• Other difficulties
  • Language only reflects the surface of meaning
    • humor, sarcasm, “between the lines” meaning
  • Language presupposes communication between people
    • Persuading, insulting, amusing them
  • Lots of subtleties
“At last, a computer that understands you like your mother”
- 1985 advertisement from a company claimed to program computer to understand human language

- At least three different interpretations:
  1. The computer understands you as well as your mother understands you
  2. The computer understands that you like your mother
  3. The computer understands you as well as it understands your mother

- Humans would rule out the last two interpretation from their knowledge of the world: we know advertisement is trying to convince us of something

different sentence structure leads to different interpretations
At last, a computer that understands you like your mother

- Word “mother” has several meanings:
  - “a female parent”
  - “a cask or vat used in vinegar-making”
“At last, a computer that understands you like your mother”

- For speech recognition:
  - “a computer that understands you like your mother”
  - a computer that understands your lie cured mother
More Ambiguity

“At last, a computer that understands you like your mother”

- Even if we interpret this as “The computer understands you as well as your mother understands you” does that mean it understands you “well” or “not so well”
- sarcasm
Another Example Syntactic Ambiguity

• How about simpler sentences?
• Even simple sentences are highly ambiguous
• “Get the cat with the gloves”
Headline Ambiguity

- Iraqi Head Seeks Arms
- Ban on Nude Dancing on Governor’s Desk
- Juvenile Court to Try Shooting Defendant
- Teacher Strikes Idle Kids
- Kids Make Nutritious Snacks
- British Left Waffles on Falkland Islands
- Red Tape Holds Up New Bridges
- Bush Wins on Budget, but More Lies Ahead
- Hospitals are Sued by 7 Foot Doctors
- Stolen Painting Found by Tree
- Local HS Dropouts Cut in Half
Why else NLP Difficult?

• Non-standard English (language in the “wild”)
  • @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either♥

• Segmentation issues
  • break-up
  • 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

• Tricky entity names
  • where A Bug’s Life playing
  • when Let It Be was recorded
Tools and Resources Needed

• Probability/Statistical Theory:
  • Statistical Distributions, Bayesian Decision Theory.

• Linguistics Knowledge:
  • Morphology, Syntax, Semantics, Pragmatics...

• Corpora:
  • Bodies of marked or unmarked text
    • The more, the better
  • to train classifiers
  • to apply statistical methods