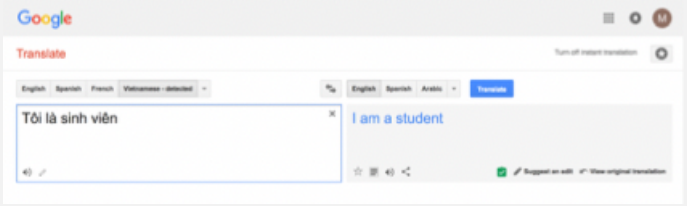
NATURAL LANGUAGE PROCESSING

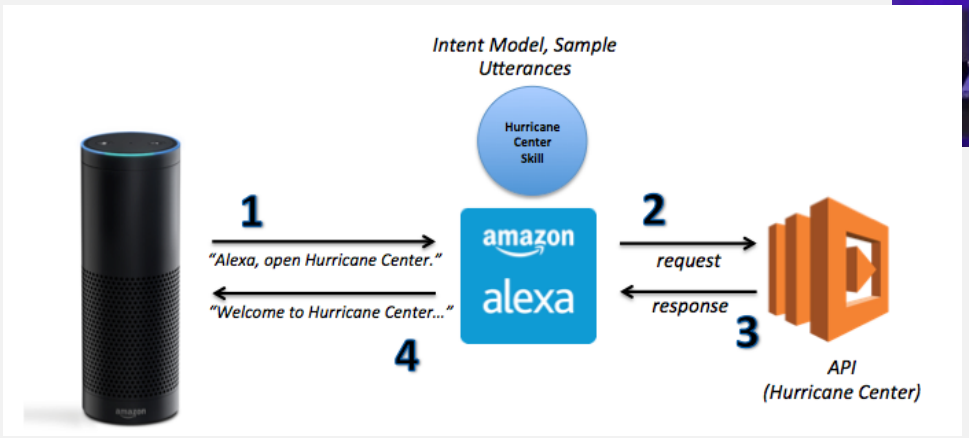
(based heavily on Dr. Pham Quang Nhat Minh's 2016 lecture,
"Introduction to Natural Language Processing")

Lecture 3

CSCI 8360

GOOGLE "NATURAL LANGUAGE PROCESSING"





WHAT IS NLP?

- A field of computer science, artificial intelligence, and computational linguistics
- To get computers to perform useful tasks involving human languages
  - Human-machine communication
  - Machine translation
  - Extracting information from text

WHY NLP?

- Languages pervades almost all human activities
  - Reading, writing, speaking, listening…
- Voice-actuated interfaces
  - Remote controls, virtual assistants, accessibility…
- We have tons of text data
  - Social networks, blogs, electronic health care records, publications…
- NLP bridges all these areas to create interesting applications
- NLP is challenging!

WHY IS NLP CHALLENGING?

- Language is **ambiguous**
- From Jurafsky book: *"I made her duck"* could mean
  - I cooked waterfowl for her.
  - I cooked the waterfowl that belongs to her.
  - I created the (plaster?) duck she owns.
  - I caused her to quickly lower her head or body.
  - I waved a magic wand and turned her into waterfowl.
- Nevermind the infamous *"Buffalo buffalo Buffalo buffalo buffalo buffalo Buffalo buffalo"*…

WHY IS NLP CHALLENGING?

- *"I shot an elephant in my pajamas."*

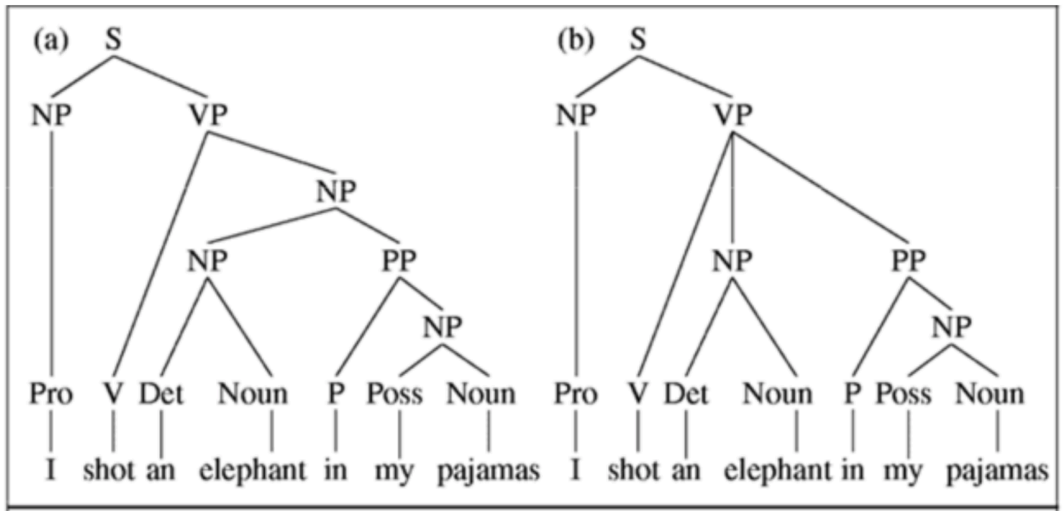


Figure 10.11 Two parse trees for an ambiguous sentence. Parse (a) corresponds to the humorous reading in which the elephant is in the pajamas, parse (b) to the reading in which Captain Spaulding did the shooting in his pajamas.

WHY IS NLP CHALLENGING?

- Ambiguity of language exists at every level
  - Lexical (word meaning)
  - Syntactic
  - Semantic
  - Discourse (conversations)
- Natural languages are fuzzy
- Natural languages rely on *a priori* knowledge of the surrounding world
  - E.g., it is unlikely that an elephant will wear pajamas

BRIEF HISTORY OF NLP

- 1940s and 1950s
  - Foundational insights
  - Automaton
  - Probabilistic and information-theoretic models
- 1957-1970
  - Two camps: symbolic (Chomsky *et al*, formal language theory and generative syntax) and stochastic (pure statistics)
- 1970-1983
  - Four paradigms, explosion in research into NLP
  - Stochastic, logic-based, natural language understanding (knowledge models), discourse modeling
- 1983-1993
  - Empiricism and finite state models, redux
- 1994-1999
  - The fields come together: probabilistic and data-driven models become the standard
- 2000-present
  - The Rise of the Planet of the Crystal Skull of Machine Learning
  - Large amount of digital data available
  - Widespread availability of high-performance computing hardware

COMMON NLP TASKS

Easy

- Chunking
- Part-of-Speech Tagging
- Named Entity Recognition
- Spam Detection
- Thesaurus

Medium

- Syntactic Parsing
- Word Sense Disambiguation
- Sentiment Analysis
- Topic Modeling
- Information Retrieval

Hard

- Machine Translation
- Text Generation
- Automatic Summarization
- Question Answering
- Conversational Interfaces

WORD SEGMENTATION

- In some languages, there's no space between words, or a word may contain smaller symbols

毎年うちの研究室の学生が1-2名国語研でアルバイトさせてもらっているので、今日は新しくアルバイトする B4 学生の紹介である。

Nhật Bản luôn là thị trường thương mại quan trọng của Việt Nam (Nhật_Bản luôn là thị_ trường thương_mại quan_trọng của Việt_Nam)

- In such cases, word segmentation is the first step in any NLP pipeline

WORD SEGMENTATION

- A possible solution is **maximum matching**
  - Start by pointing at the beginning of a string, then choose the longest word in the the dictionary that matches the input at the current position

Nhật_Bản luôn là thị trường thương mại quan trọng của Việt Nam

- Nhật_Bản is a word in dictionary, but “Nhật Bản luôn” is not

- Problems:
  - Maxmatching can’t deal with unknown words
  - Dependency between words in the same sentences is not exploited

WORD SEGMENTATION

- Most successful word segmentation tools are based on ML techniques
- Word segmentation tools obtain a high accuracy
  - vn.vitk (https://github.com/phuonglh/vn.vitk) obtained **97% accuracy** on test data

- Not necessarily a problem with whitespace-delimited languages (like English) but still have corner cases

POS TAGGING

- Each word in a sentence can be classified in to classes, such as verbs, adjectives, nouns, etc
- *POS Tagging* is a process of tagging words in a sentences to particular part-of-speech, based on:
  - Definition
  - Context
- *The/DT grand/JJ jury/NN commented/VBD on/IN a/DT number/NN of/IN other/JJ topics/NNS ./.*

SEQUENCE LABELING

- Many NLP problems can be viewed as sequence labeling
- Each token in a sequence is assigned a label
- Labels of tokens are dependent on the labels of other tokens in the sequence, particularly their neighbors

John	saw	the	saw	and	decided	to	take	it	to	the	table.
NNP	VBD	DT	NN	CC	VBD	TO	VB	PRP	IN	DT	NN

PROBABILISTIC SEQUENCE MODELS

- Model probabilities of pairs (token sequences, tag sequences) from annotated data
- Exploit dependency between tokens
- Typical sequence models
  - Hidden Markov Models (HMMs)
  - Conditional Random Fields (CRF)

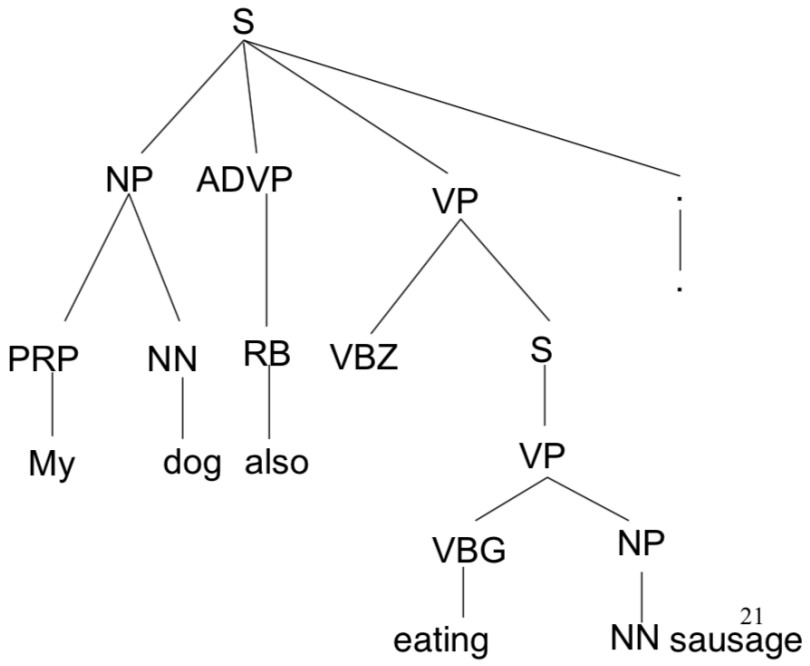
SYNTAX ANALYSIS

- The task of recognizing a sentence and assigning a syntactic structure to it
- An important task in NLP with many applications
  - Intermediate stage of representation for semantic analysis
  - Play an important role in applications like question answering and information extraction
  - E.g., *What books were written by British women authors before 1800?*

SYNTAX ANALYSIS

My dog also likes eating sausage.

```
(ROOT
  (S
    (NP (PRP$ My) (NN dog))
    (ADVP (RB also))
    (VP (VBZ likes)
      (S
        (VP (VBG eating)
          (NP (NN sausage)))))
    (. .)))
```



APPROACHES TO SYNTAX ANALYSIS

- Top-down parsing
- Bottom-up parsing
- Dynamic programming methods
  - CYK algorithm
  - Earley algorithm
  - Chart parsing
- Probabilistic Context-Free Grammars (PCFG)
  - Assign *probabilities* for derivations

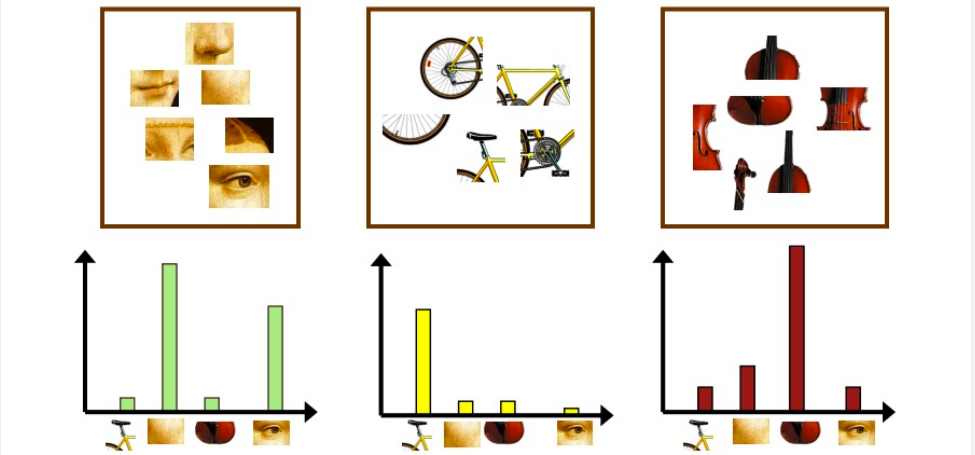
SEMANTIC ANALYSIS

- Two levels

1. Lexical semantics
   - Representing meaning of words
   - Word sense disambiguation (e.g., word *bank*)
2. Compositional semantics
   - How words combined to form a larger meaning.

SEMANTIC ANALYSIS TECHNIQUES

- Bag-of-words
  - Word order doesn't matter, only word frequency
  - Works surprisingly well in practice (e.g., Naïve Bayes)
  - Fails hilariously at times (word order *does* matter, stop words, etc)



SEMANTIC ANALYSIS TECHNIQUES

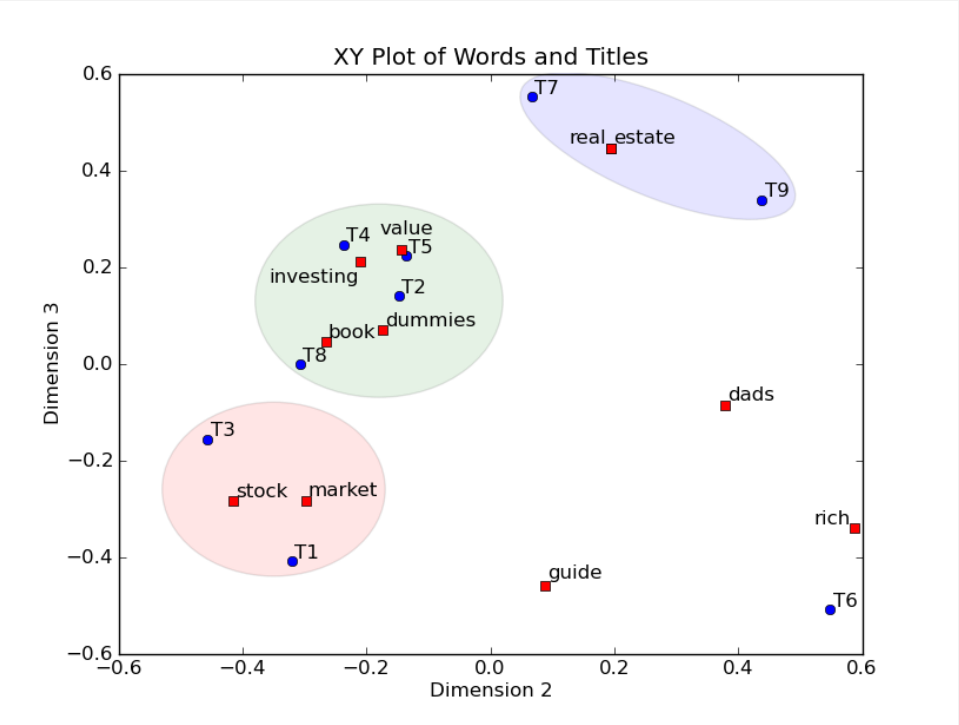
- TF-IDF
  - Slight modification on standard bag-of-words
  - Includes an *inverse document frequency* term to offset effects of stopwords
  - Works even better in practice
  - Term counts are now document-specific

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

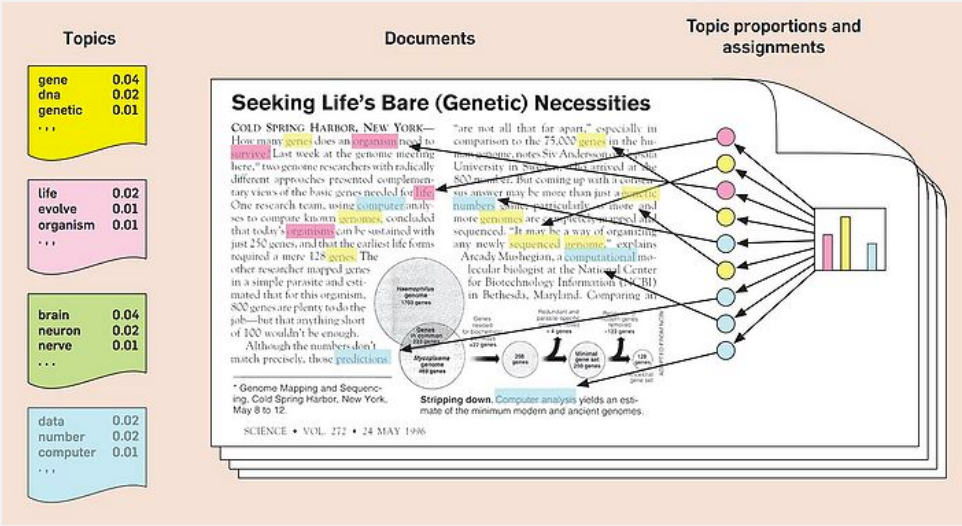
SEMANTIC ANALYSIS TECHNIQUES

- Latent Semantic Analysis (LSA)
  - Basically matrix factorization of term frequencies
  - Pulls out semantic "concepts" present in the documents
  - Sometimes "concepts" defy intuitive interpretation



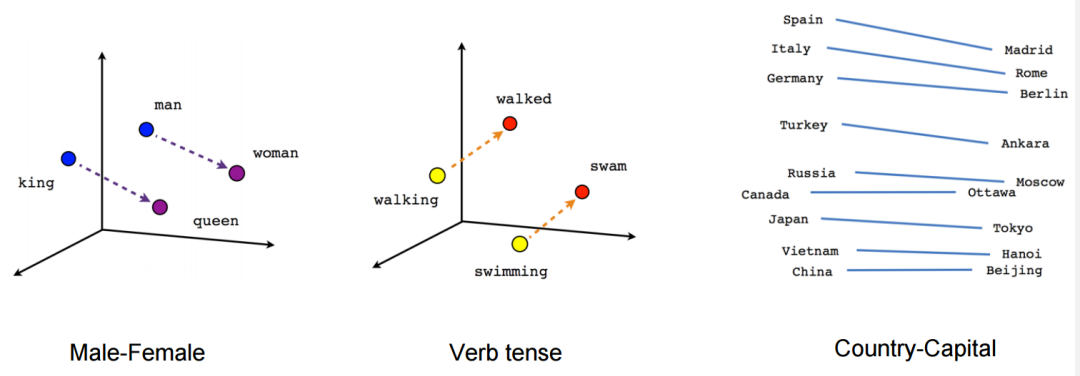
SEMANTIC ANALYSIS TECHNIQUES

- Latent Dirichlet Allocation (LDA)
  - Explicitly models topic distributions even within the same document
  - Generative model that can "simulate" documents belonging to a single topic
  - Really hard to train
  - Topics again defy intuitive interpretation



SEMANTIC ANALYSIS TECHNIQUES

- Word embeddings
  - word2vec, doc2vec, GloVe
  - Build a vector representation of a word
  - Define it by its *context* (neighboring words)
  - Can perform "word algebra"
  - Embeddings dependent on corpus used to train them



PROJECT 0

- Out now! Check it out (links on AutoLab and the course website)
- Due **Tuesday, January 16 at 11:59pm**
- Can't use nltk, breeze, or other NLP-specific packages
  - Really, you won't need them
- Spark & "NLP"
  - Count words in documents (term frequencies)
  - Incorporate stopword filtering (will **need** broadcast variables for this)
  - Truncate out punctuation
  - Implement TF-IDF for improved word counting

PROJECT 0

- **Pay attention to the requirements of the deliverables**
  - Incorrectly-named or formatted JSON files will cause autograder to fail
  - Name GitHub repo correctly
  - Include README and CONTRIBUTORS files
  - Practice using git (commit, push, branch, merge) and GitHub functionality (issues, milestones, pull requests)

REFERENCES

- "Introduction to natural language processing", https://www.slideshare.net/minhpqn/introduction-to-natural-language-processing-67212472
- NLP slides from Stanford Coursera course https://web.stanford.edu/~jurafsky/NLPCourseraSlides.html