Inspire…Educate…Transform.

# 5. Natural Language Processing

**Dr. Manish Gupta**

Sr. Mentor – Academics, INSOFE

Adapted from Dan Jurafsky's course slides on
https://web.stanford.edu/~jurafsky/NLPCourseraSlides.html
https://www.cse.iitb.ac.in/~nlp-ai/WSD.ppt

# Course Content

- Collection of three main topics of high recent interest.
  - Search engines (Crawling, Indexing, Ranking)
    - Language Modeling
    - Text Indexing and Crawling
    - Relevance Ranking
    - Link Analysis Algorithms
  - Text Processing (NLP, NER, Sentiments)
    - **Natural Language Processing**
    - Named Entity Recognition
    - Sentiment Analysis
    - Summarization
  - Social networks (Properties, Influence Propagation)
    - Social Network Analysis
    - Influence Propagation in Social Networks

# Agenda

- **What is NLP?**

# Question Answering: 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

# Information Extraction

Subject: **curriculum meeting**

Date: January 15, 2012

To: Dan Juraf

Hi Dan, we've now scheduled the curriculum

It will be in Gates 159 tomorrow from 10:00-

-Chris

Event: Curriculum meeting
Date: Jan-16-2012
Start: 10:00am
End: 11:30am
Where: Gates 159

Create new Calendar entry

# Information Extraction & Sentiment Analysis

Attributes:

- zoom
- affordability
- size and weight
- flash
- ease of use

Size and weight

- ✓ nice and compact to carry!
- ✓ since the camera is small and light, I w
  around those heavy, bulky professiona
- ✗ the camera feels flimsy, is plastic and very light in weight you have to be very delicate in the handling of this camera

# Machine Translation

- Fully automatic

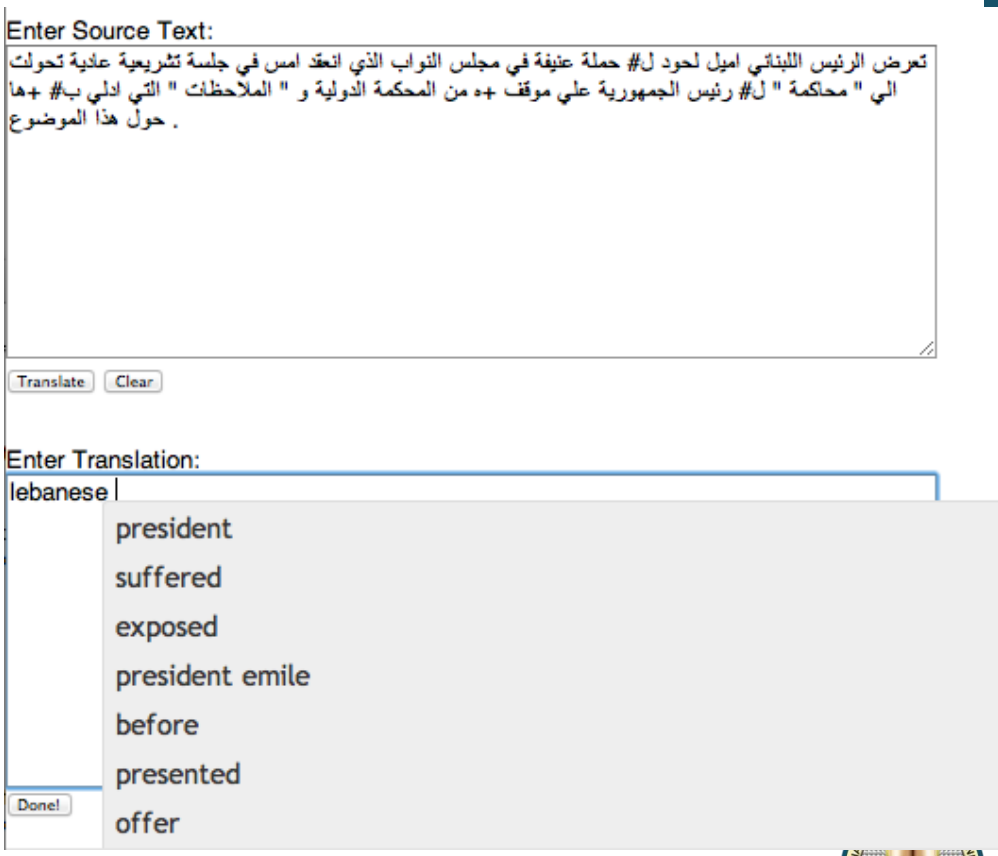
Enter Source Text:

这 不过 是 一 个 时间 的 问题 .

Translation from Stanford's *Phrasal*:

This is only a matter of time.

- Helping human translators



# 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

Party
May 27
add

## 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 Sonu Nigam playing in Hyderabad?

Ella Hotels at 7:30. Do you want a ticket?

# Other NLP Tasks

- Discourse analysis
- Morphological segmentation
- Natural language generation
- Natural language understanding
- Optical character recognition (OCR)
- Relationship extraction
- Sentence breaking
- Speech recognition
- Speech segmentation
- Topic segmentation and recognition
- Word segmentation
- Speech processing
- Native Language Identification
- Stemming
- Text simplification
- Text-to-speech
- Text-proofing
- Natural language search
- Query expansion
- Automated essay scoring
- Truecasing

# NLP Challenges

- Teacher Strikes Idle Kids
- Red Tape Holds Up New Bridges
- Hospitals Are Sued by 7 Foot Doctors
- Juvenile Court to Try Shooting Defendant
- Local High School Dropouts Cut in Half

Ambiguity

## non-standard English

Great job @justinbieber! Were SOO PROUD of what youve 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

## world knowledge

Mary and Sue are sisters.

Mary and Sue are mothers.

## tricky entity names

Where is *A Bug's Life* playing …

*Let It Be* was recorded …

… a mutation on the *for* gene …

# Agenda

- What is NLP?
- **Tokenization, Stemming**

# Tokenization

- Issues in English tokenization
  - `Finland's capital` → `Finland Finlands Finland's ?`
  - `what're, I'm, isn't` → `What are, I am, is not`
  - `Hewlett-Packard` → `Hewlett Packard ?`
  - `state-of-the-art` → `state of the art ?`
  - `Lowercase` → `lower-case lowercase lower case ?`
  - `San Francisco` → one token or two?
  - Acronyms: m.p.h., PhD. → ??
- French
  - ***L'ensemble*** → one token or two? ***L*** ? ***L'*** ? ***Le*** ?
    - La - feminine, Le - masculine, L' - starting a vowel, Les - plural, Un - masculine a, Une - feminine a, Des - plural a
- German noun compounds are not segmented
  - ***Lebensversicherungsgesellschaftsangestellter***
  - 'life insurance company employee'
  - German information retrieval needs **compound splitter**

# Tokenization: Language issues

- Chinese and Japanese no spaces between words:
  - 莎拉波娃现在居住在美国东南部的佛罗里达。
  - 莎拉波娃 现在 居住 在 美国 东南部 的 佛罗里达
  - Sharapova now lives in US southeastern Florida
- Further complicated in Japanese, with multiple alphabets intermingled
  - Dates/amounts in multiple formats

***フォーチュン500社は情報不足のため時間あた$500K(約6,000万円)***

Katakana Hiragana Kanji Romaji

End-user can express query entirely in hiragana!

# Word Tokenization in Chinese

- Also called **Word Segmentation**
- Chinese words are composed of characters
  - Characters are generally 1 syllable and 1 morpheme.
  - Average word is 2.4 characters long.
- Standard baseline segmentation algorithm:
  - Maximum Matching (also called Greedy)

# Max-match segmentation illustration

- Thecatinthehat — the cat in the hat
- Thetabledownthere — the table down there

  theta bled own there
- Doesn’t generally work in English!
- But works astonishingly well in Chinese
  - 莎拉波娃现在居住在美国东南部的佛罗里达。
  - 莎拉波娃 现在 居住 在 美国 东南部 的 佛罗里达
- Modern probabilistic segmentation algorithms even better

# Maximum Matching Word Segmentation Algorithm

- Given a wordlist of Chinese, and a string.

1) Start a pointer at the beginning of the string
2) Find the longest word in dictionary that matches the string starting at pointer
3) Move the pointer over the word in string
4) Go to 2

# Lemmatization

- Reduce inflections or variant forms to base form
  - *am, are, is* $\rightarrow$ *be*
  - *car, cars, car's, cars'* $\rightarrow$ *car*
- *the boy's cars are different colors* $\rightarrow$ *the boy car be different color*
- Lemmatization: have to find correct dictionary headword form
- Machine translation
  - Spanish quiero (‘I want’), quieres (‘you want’) same lemma as querer ‘want’

# Stemming

- Reduce terms to their stems in information retrieval
- *Stemming* is crude chopping of affixes
  - language dependent
  - e.g., ***automate(s), automatic, automation*** all reduced to ***automat***.

*for example compressed and compression are both accepted as equivalent to compress.*

for exampl compress and compress ar both accept as equival to compress

# Porter's algorithm
# The most common English stemmer

Step 1a

```
sses → ss caresses → caress
ies  → i ponies    → poni
ss   → ss caress   → caress
s    → ø cats      → cat
```

Step 1b

```
(*v*)ing → ø
walking  → walk
sing     → sing
(*v*)ed  → ø
plastered → plaster
…
```

Step 2 (for long stems)

```
ational→ ate relational→ relate
izer→ ize digitizer → digitize
ator→ ate operator  → operate
…
```

Step 3 (for longer stems)

```
al    → ø   revival    → reviv
able  → ø   adjustable → adjust
ate   → ø   activate   → activ
…
```

# Viewing morphology in a corpus
# Why only strip –ing if there is a vowel?

```
(*v*)ing → ø  walking   → walk
              sing      → sing
```

| | | | |
|---|---|---|---|
| 1312 | King | 548 | being |
| 548 | being | 541 | nothing |
| 541 | nothing | 152 | something |
| 388 | king | 145 | coming |
| 375 | bring | 130 | morning |
| 358 | thing | 122 | having |
| 307 | ring | 120 | living |
| 152 | something | 117 | loving |
| 145 | coming | 116 | Being |
| 130 | morning | 102 | going |

# Problems with stemming

- Lack of domain-specificity and context can lead to occasional serious retrieval failures (which “stocking” is meant)
- Stemmers are often difficult to understand and modify
- Sometimes too aggressive in conflation
  - e.g., “policy”/“police”, “execute”/“executive”, “university”/“universe”, “organization”/“organ” are conflated by Porter
- Miss good conflations
  - e.g., “European”/“Europe”, “matrices”/“matrix”, “machine”/“machinery” are not conflated by Porter
- Produce stems that are not words and are often difficult for a user to interpret
  - e.g., with Porter, “iteration” produces “iter” and “general” produces “gener”

# Stemming vs lemmatization

- Stemming:
  - crude heuristic process
  - chops off the ends of words in the hope of achieving this goal correctly most of the time.
- Lemmatization:
  - does things properly with the use of a vocabulary and morphological analysis of words
  - normally aims to remove inflectional endings only
  - returns the base or dictionary form of a word, which is known as the lemma.
- For the token “saw”
  - stemming might return just “s”
  - lemmatization would attempt to return either “see” or “saw” depending on whether the use of the token was as a verb or a noun.

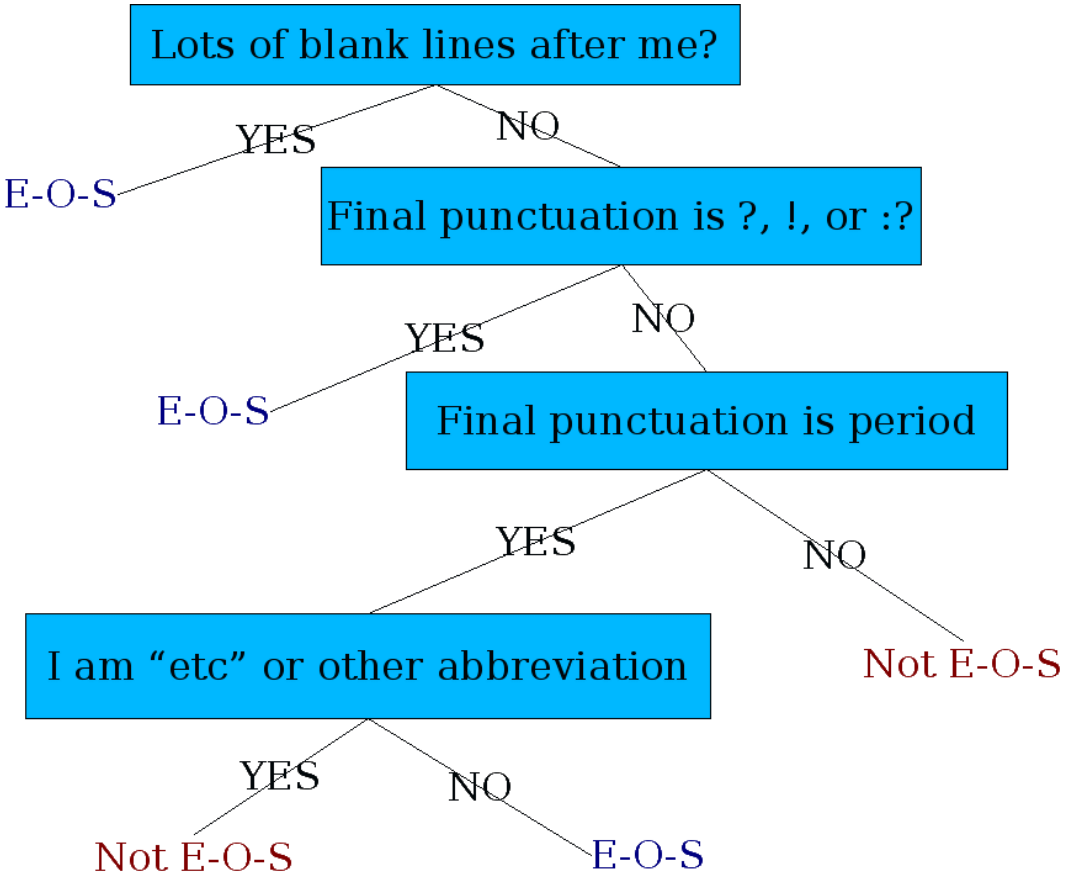
# Agenda

- What is NLP?
- Tokenization, Stemming
- **Sentence Segmentation, Phrase Identification**

# Sentence Segmentation

- !, ? are relatively unambiguous
- Period "." is quite ambiguous
  - Sentence boundary
  - Abbreviations like Inc. or Dr.
  - Numbers like .02% or 4.3
- Build a binary classifier
  - Looks at a "."
  - Decides EndOfSentence/NotEndOfSentence
  - Classifiers: hand-written rules, regular expressions, or machine-learning

# Determining if a Word is End-of-sentence: A Decision Tree




- Other features
  - Case of word with/after ".": Upper, Lower, Cap, Number
  - Length of word with "."

# Phrase Identification

- Goal is to use phrases as indexing units
  - Makes general words more specific
  - blood → blood hound, blood test, blood brother, …
- Statistical approach
  - Index all pairs of adjacent words ("bigrams")
  - Explosion in index elements makes this non-feasible
  - Also, it adds lots of "nonsense" phrases
    - "also it", "it adds", "adds lots", "lots of", "of nonsense", "nonsense phrases"
- NLP approaches
  - Runs of words
  - Sentence parsing
  - Statistical models

# Phrases as Runs of Words

- Consider all runs of words between stop words
  - Can easily be extended to allow some stopwords
    - e.g., Library of Congress, cats and dogs
- Scan a large body of text for occurrences of phrases
- Any that occur more than n times are valid
  - Small n (e.g., 4) works impressively well

# “Phrase identification”

- “Goal” is to “use phrases” as “indexing units”
  - Makes “general words” more “specific”
  - “blood” → “blood hound”, “blood test”, “blood brother”, …
- “Statistical approach”
  - “Index” all “pairs” of “adjacent words” (“bigrams”)
  - “Explosion” in “index elements” makes this “non-feasible”
- “NLP approaches”
  - “Runs” of “words”
  - “Sentence parsing”
  - “Statistical models”

# Phrases and Their Counts

## TREC

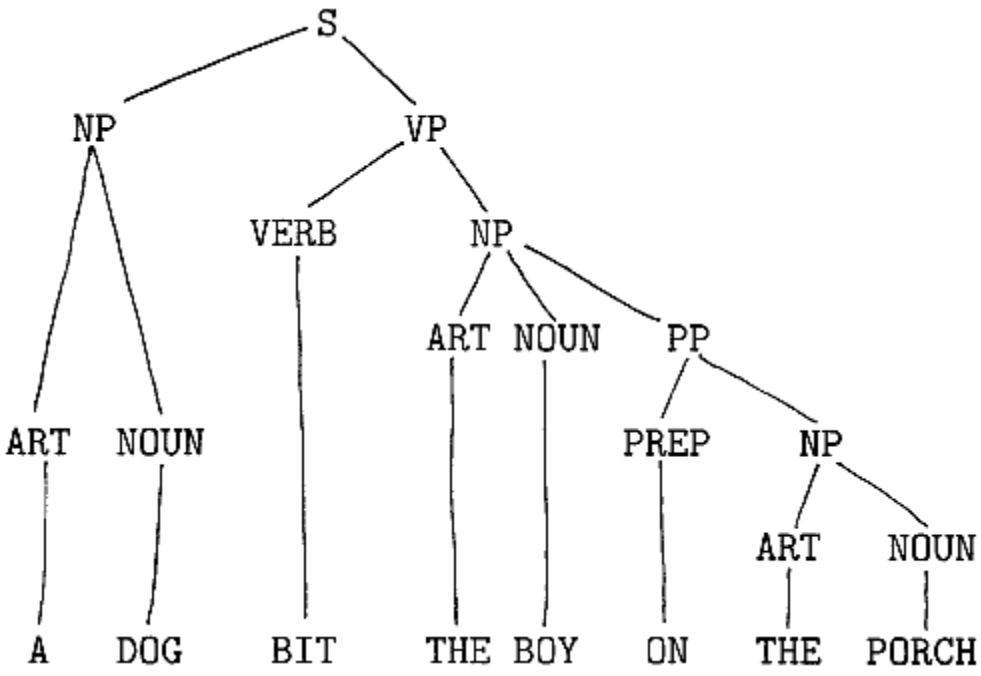
65824 United States
61327 Article Type
33864 Los Angeles
18062 Hong Kong
17788 North Korea
17308 New York
15513 San Diego
15009 Orange County
12869 prime minister
12799 first time
12067 Soviet Union
10811 Russian Federation
9912 United Nations
8127 Southern California
7640 South Korea
7620 end recording
7524 European Union
7436 South Africa
7362 San Francisco
7086 news conference
6792 City Council
6348 Middle East
6157 peace process
5955 human rights
5837 White House
5778 long time
5776 Armed Forces
5636 Santa Ana
5619 Foreign Ministry
5527 Bosnia-Herzegovina
5458 words indistinct
5452 international community
5443 vice president
5247 Security Council
5098 North Korean
5023 Long Beach
4981 Central Committee
4872 economic development
4808 President Bush
4652 press conference
4602 first half
4565 second half
4495 nuclear weapons
4448 UN Security Council
4426 South Korean
4219 first quarter
4166 Los Angeles County
4107 State Duma
4085 State Council
3969 market economy
3941 World War II

## U.S. patents

975362 present invention
191625 U.S. Pat
147352 preferred embodiment
95097 carbon atoms
87903 group consisting
81809 room temperature
78458 SEQ ID
75850 BRIEF DESCRIPTION
66407 prior art
59828 perspective view
58724 first embodiment
56715 reaction mixture
54619 DETAILED DESCRIPTION
54117 ethyl acetate
52195 Example 1
52003 block diagram
46299 second embodiment
41694 accompanying drawings
40554 output signal
37911 first end
35827 second end
34881 appended claims
33947 distal end
32338 cross-sectional view
30193 outer surface
29635 upper surface
29535 preferred embodiments
29252 present invention provides
29025 sectional view
28961 longitudinal axis
27703 title compound
27434 PREFERRED EMBODIMENTS
27184 side view
25903 inner surface
25802 Table 1
25047 lower end
25047 plan view
24513 third embodiment
24432 control signal
24296 upper end
24275 methylene chloride
24117 reduced pressure
23831 aqueous solution
23618 SEQUENCE DESCRIPTION
23616 SEQUENCE CHARACTERISTICS
22382 weight percent
22070 closed position
21356 light source
21329 image data
21026 flow chart
21003 PREFERRED EMBODIMENT

# Phrases from Sentence Parsing

- Run a shallow or deep parsing system
  - Simplest and common approach uses noun phrases
  - Can use other types, too, of course
    - Verb phrases, noun phrases with adjectives, prepositional phrases, noun+verb phrases, …



# Phrases from statistical models

- Build a dictionary of phrases using heuristic methods
  - High-frequency phrases (1-6 words) that occur frequently
  - Some POS tagging for some lower-frequency phrases
    - e.g., throw away verbs or phrases ending with adjectives
- Estimate probabilities for Markov model
  - …that first word is the start of a phrase
  - …that next word is part of the same phrase
  - …that a phrase follows this phrase
- Using pointwise mutual information
  - $\mathrm{MI}(w_i, w_j) = \frac{P(w_i, w_j)}{P(w_i)P(w_j)}$

# Agenda

- What is NLP?
- Tokenization, Stemming
- Sentence Segmentation, Phrase Identification
- **Word Sense Disambiguation**

# Word Sense Disambiguation

- Computationally determining which ***sense*** of a word is activated by its use in a particular ***context***.
  - E.g. I am going to withdraw money from the ***bank.***
- A classification problem:
  - Senses → Classes
  - Context → Used to compute features

# WSD Using Selectional Preferences & Arguments

| Sense 1 | Sense 2 |
|---|---|
| This airlines ***serves*** dinner in the evening flight. | This airlines ***serves*** the sector between Agra & Delhi. |
| serve (Verb) | serve (Verb) |
| • agent | • agent |
| • object – dinner | • object – sector |

Requires exhaustive enumeration of:

- Argument-structure of verbs.
- Selectional preferences of arguments.
- Description of properties of words such that meeting the selectional preference criteria can be decided.

  E.g. This flight serves the "**region**" between Mumbai and Delhi

  How do you decide if "region" is compatible with "sector"

# Overlap Based Approaches

- Require a ***Machine Readable Dictionary*** *(MRD).*
- Find the overlap between the features of different senses of an ambiguous word (sense bag) and the features of the words in its context (context bag).
- These features could be sense definitions, example sentences, hypernyms etc.
- The features could also be given weights.
- The sense which has the maximum overlap is selected as the contextually appropriate sense.

- Lesk's and Walker's algorithms.

# LESK'S ALGORITHM

**Sense Bag**: *contains the words in the definition of a candidate sense of the ambiguous word.*

**Context Bag**: *contains the words in the definition of each sense of each context word.*

E.g. "On burning ***coal*** we get ***ash***."

| Ash | Coal |
|---|---|
| **Sense 1**<br>Trees of the olive family with pinnate leaves, thin furrowed bark and gray branches. | **Sense 1**<br>A piece of glowing carbon or ***burn***t wood. |
| **Sense 2**<br>The ***solid*** residue left when ***combustible*** material is thoroughly ***burn***ed or oxidized. | **Sense 2**<br>charcoal. |
| **Sense 3**<br>To convert into ash | **Sense 3**<br>A black ***solid combustible*** substance formed by the partial decomposition of vegetable matter without free access to air and under the influence of moisture and often increased pressure and temperature that is widely used as a fuel for ***burn***ing |

In this case Sense 2 of ash would be the winner sense.

# Walker's Algorithm

- A Thesaurus Based approach.
- ***Step 1:*** *For each sense of the target word find the thesaurus category to which that sense belongs.*
- ***Step 2:*** *Calculate the score for each sense by using the context words. A context word will add 1 to the score of the sense if the thesaurus category of the word matches that of the sense.*
  - *E.g. The money in this **bank** fetches an interest of 8% per annum*
  - Target word: ***bank***
  - Clue words from the context: ***money, interest, annum, fetch***

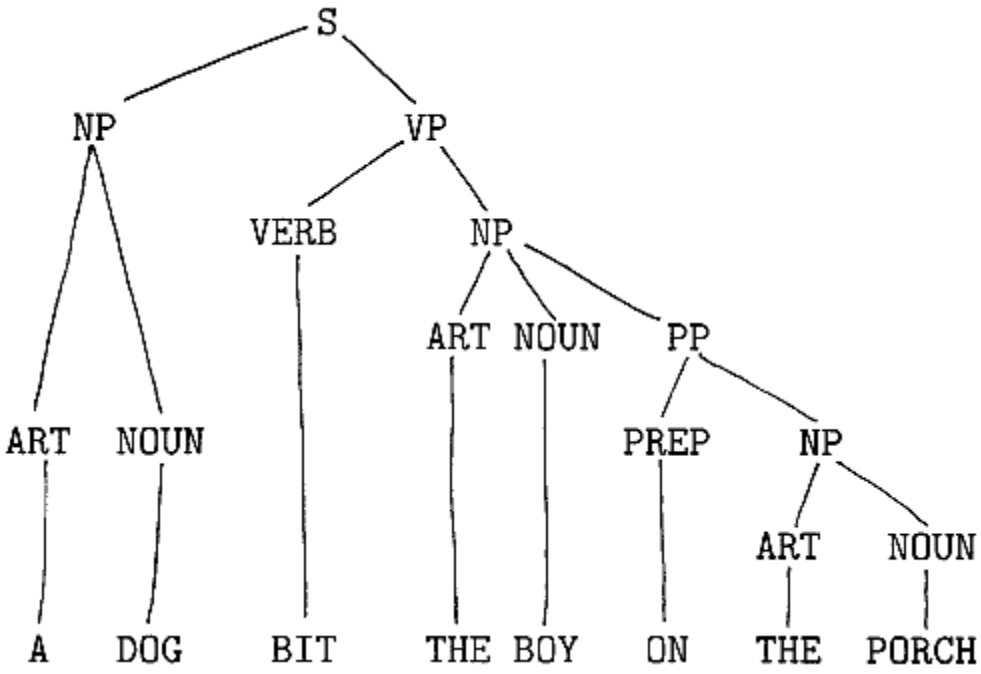
| | Sense1: Finance | Sense2: Location |
|---|---|---|
| Money | +1 | 0 |
| Interest | +1 | 0 |
| Fetch | 0 | 0 |
| Annum | +1 | 0 |
| Total | 3 | 0 |

Context words add 1 to the sense when the topic of the word matches that of the sense

# Agenda

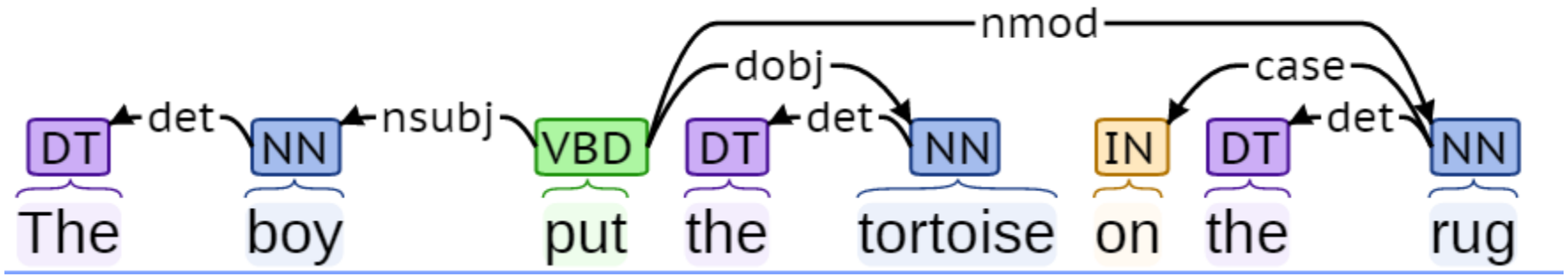
- What is NLP?
- Tokenization, Stemming
- Sentence Segmentation, Phrase Identification
- Word Sense Disambiguation
- **Parsing**

# Constituency (phrase structure) parsing



# Dependency Parsing

- Dependency structure shows which words depend on (modify or are arguments of) which other words.




*The boy put the tortoise on the rug*

- http://nlp.stanford.edu:8080/corenlp/process

# Take-aways

- NLP is set of tasks for natural language processing.
- We looked at a few of such tasks in detail
  - Tokenization
  - Stemming
  - Sentence Segmentation
  - Phrase Identification
  - Word Sense Disambiguation
  - Parsing

# Further Reading

- Books
  - Foundations of Statistical Natural Language Processing: Christopher D. Manning, Hinrich Schütze
  - Speech and Language Processing, 2nd Edition: Daniel Jurafsky, James H. Martin
- https://www.coursera.org/course/nlp

# International School of Engineering

Plot 63/A, 1st Floor, Road # 13, Film Nagar, Jubilee Hills, Hyderabad - 500 033

For Individuals: +91-9502334561/63 or 040-65743991

For Corporates: +91-9618483483

Web: http://www.insofe.edu.in

Facebook: https://www.facebook.com/insofe

Twitter: https://twitter.com/Insofeedu

YouTube: http://www.youtube.com/InsofeVideos

SlideShare: http://www.slideshare.net/INSOFE

LinkedIn: http://www.linkedin.com/company/international-school-of-engineering