Natural Language Processing
Introduction
Part 2

Sudeshna Sarkar

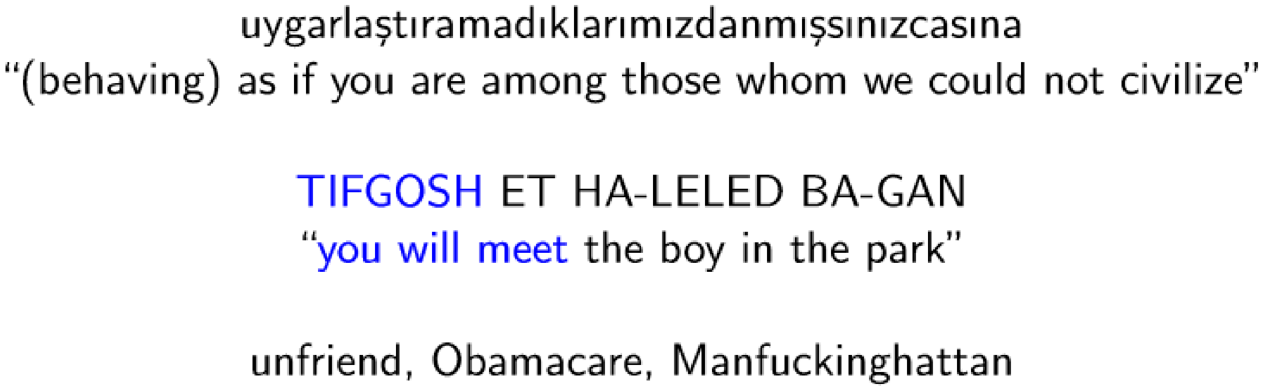
18 July 2019

Orthography

- Linking the symbols of an alphabet to the sounds of a language.
- Enable written communication
- How the symbols (graphemes) represent the sounds (phonemes) used in spoken language.

Morphology

The identification, analysis and description of the structure of words



uygarlaştıramadıklarımızdanmışsınızcasına

“(behaving) as if you are among those whom we could not civilize”

TIFGOSH ET HA-LELED BA-GAN

“you will meet the boy in the park”

unfriend, Obamacare, Manfuckinghattan

Ambiguity

Turkish word izin

1. Yerdeki ***izin*** temizlenmesi gerek.
   iz + Noun+A3sg+Pnon+Gen

The trace on the floor should be cleaned.

2. *Ü*zerinde parmak ***izin*** kalmis¸
   iz + Noun+A3sg+P2sg+Nom

Your finger **print** is left on (it).

3. Ic¸eri girmek ic¸in *izin* alman gerekiyor.
   izin + Noun+A3sg+Pnon+Nom

You need a **permission** to enter.

- Bengali word মাতাল

ওরে **মাতাল**, দুয়ার ভেঙে দিয়ে

Noun (drunkard)

মাতাল হয়ে পাতাল-পানে ধাওয়া।

Adjctive

মোর ভাবনারে কী হাওয়ায় **মাতালো**

Verb, simple past

Morphology

- The identification, analysis and description of the structure of words

- Morpheme: the smallest linguistic unit with semantic meaning
- Lexeme: corresponds to a set of forms taken by a single word.

The Challenge of Words

- Segmenting text into words (Thai)
- Sandhi splitting (Sanskrit)
- Morphological variation
- Words with multiple meanings (based on context, domain)
- Multiword expression

Lexicon

- The lexicon contains information about particular idiosyncratic properties of words; eg. what sound or orthography goes with what meaning

Syntax

- Syntax concerns the way in which words can be combined together to form (grammatical) sentences

1. revolutionary new ideas appear infrequently
2. colourless green ideas sleep furiously
3. *ideas green furiously colourless sleep

Syntax

- Words combine syntactically in certain orders in a way which mirrors the meaning conveyed
  - John loves Mary
  - Mary loves John
- John gave her dog biscuits
  - (john (gave (her) (dog biscuits)))
  - (john (gave (her dog) (biscuits)))

Semantics

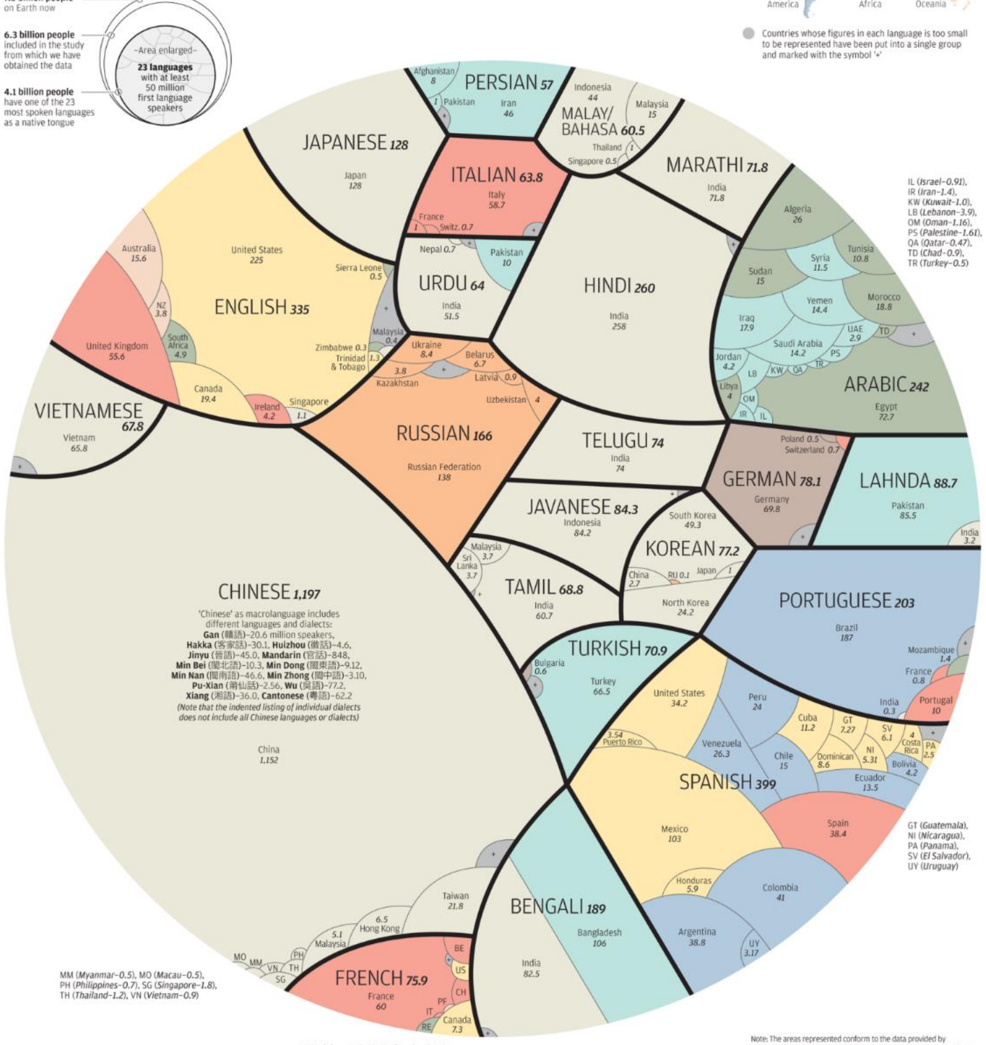
- The manner in which lexical meaning is combined morphologically and syntactically to form the meaning of a sentence
  - Concerns the meaning of words, phrases and sentences
  - The meaning of a sentence is usually a productive combination of the meaning of its words

Discourse analysis

- The meaning of a sentence depends upon the sentences that preceded it and also invokes the meaning of the sentences that follow it.
- The discourse structure of connected text, i.e. the nature of the discourse relationships between sentences (e.g. elaboration, explanation, contrast)

Pragmatics

- Pragmatics is about the use of language in context
  - the linguistic and situational context of an utterance
  - “Draw the curtains”
  - “Could you pass the salt?”
- **General Knowledge** also plays an important role in language interpretation



Countries whose figures in each language is too small to be represented have been put into a single group and marked with the symbol '+'

'Chinese' as macrolanguage includes different languages and dialects: **Gan** (贛語)–20.6 million speakers, **Hakka** (客家話)–30.1, **Huizhou** (徽話)–4.6, **Jinyu** (晉語)–45.0, **Mandarin** (官話)–848, **Min Bei** (閩北語)–10.3, **Min Dong** (閩東語)–9.12, **Min Nan** (閩南語)–46.6, **Min Zhong** (閩中語)–3.10, **Pu-Xian** (莆仙話)–2.56, **Wu** (吳語)–77.2, **Xiang** (湘語)–36.0, **Cantonese** (粵語)–62.2 *(Note that the indented listing of individual dialects does not include all Chinese languages or dialects)*

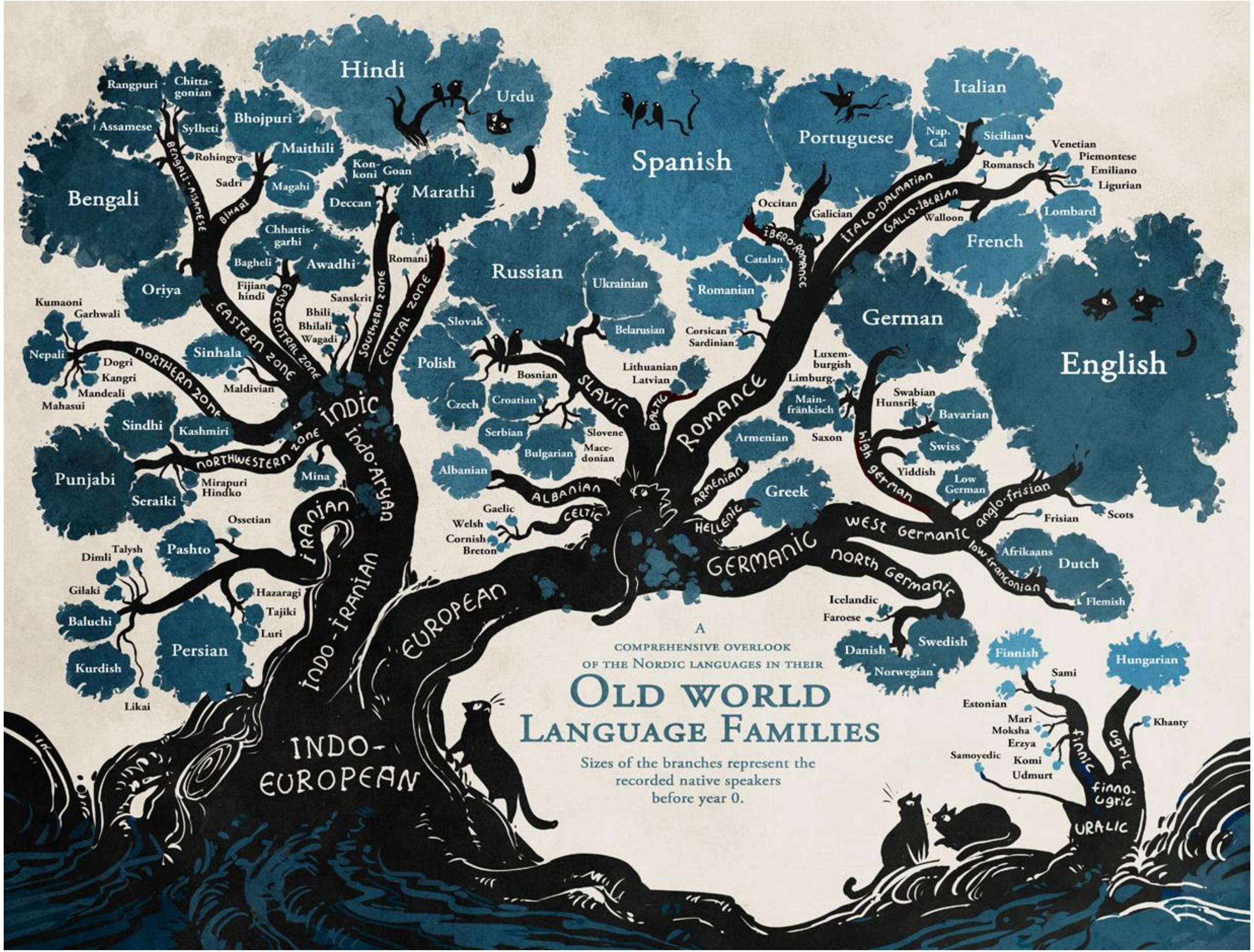
IL (*Israel–0.91*),
IR (*Iran–1.4*),
KW (*Kuwait–1.0*),
LB (*Lebanon–3.9*),
OM (*Oman–1.16*),
PS (*Palestine–1.61*),
QA (*Qatar–0.47*),
TD (*Chad–0.9*),
TR (*Turkey–0.5*)

GT (*Guatemala*),
NI (*Nicaragua*),
PA (*Panama*),
SV (*El Salvador*),
UY (*Uruguay*)

MM (*Myanmar–0.5*), MO (*Macau–0.5*),
PH (*Philippines–0.7*), SG (*Singapore–1.8*),
TH (*Thailand–1.2*), VN (*Vietnam–0.9*)

BE (*Belgium–3.9*), CH (*Switzerland–1.9*),
IT (*Italy–0.1*), PF (*French Polynesia–0.2*),
RE (Réunion–*0.7*), US (*United States–1.3*)

Note: The areas represented conform to the data provided by "Ethnologue-Languages of the World". These estimates are not absolute because the demography is constantly evolving. Some studies are based on data from old census and may date back more than 8 years



A COMPREHENSIVE OVERLOOK OF THE NORDIC LANGUAGES IN THEIR

OLD WORLD LANGUAGE FAMILIES

Sizes of the branches represent the
recorded native speakers
before year 0.

All India | Press Trust of India | Updated: July 01, 2018 15:30 IST

- More Than 19,500 Languages Spoken In India: Census
- There are 121 languages which are spoken by 10,000 or more people in India, which has a population of 121 crore, according to a census analysis.
- However, 96.71 per cent population in the country have one of the 22 scheduled languages as their mother tongue.

Hardness of NLP

- Mappings across levels are complex.
  - A string may have many possible interpretations in different contexts, and resolving ambiguity correctly may rely on knowing a lot about the world.
  - Richness: any meaning may be expressed many ways, and there are immeasurably many meanings.
  - Linguistic diversity across languages, dialects, genres, styles

Challenges: Ambiguity

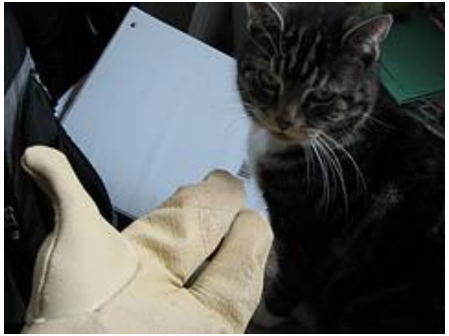
- Word-level ambiguity
  - “design” can be a noun or a verb (Ambiguous POS)
  - “root” has multiple meanings (Ambiguous sense)
  - Different morphological derivations: Bengali word “maataala”, “taaraa”

Challenge: Ambiguity

- Syntactic Analysis
- Attachment ambiguity
  - Word: "get the cat with the gloves"

Challenge: Ambiguity

- Syntactic Analysis
- Attachment ambiguity
  - Word: “get the cat with the gloves”



Challenge: Ambiguity

- Syntactic Analysis
- Attachment ambiguity
  - Word: “get the cat with the gloves”

    “I ate spaghetti with chopsticks” vs. “I ate spaghetti with meatballs.”
  - Phrase: I saw a tiger running across the field
  - Clause: I told the child that I liked that he came to the playground early.

Challenge: Ambiguity

- Semantic Analysis
  - “The dog is in the pen.” vs. “The ink is in the pen.”
  - “I put the plant in the window” vs. “Ford put the plant in Mexico”
  - Visiting aunts can be trying
  - आपको मुझे मीठाई खीलानी पड़ेगी

Challenge: Ambiguity

- Pragmatic Analysis From “The Pink Panther Strikes Again”:

Clouseau: Does your dog bite?
Hotel Clerk: No.
Clouseau: [bowing down to pet the dog] Nice doggie.
[Dog barks and bites Clouseau in the hand]
Clouseau: I thought you said your dog did not bite!
Hotel Clerk: That is not my dog.

Humor and Ambiguity

- Many jokes rely on the ambiguity of language:
  - Groucho Marx:

    One morning I shot an elephant in my pajamas.

    How he got into my pajamas, I'll never know.

Humor and Ambiguity

- She criticized my apartment, so I knocked her flat.
- Noah took all of the animals on the ark in pairs. Except the worms, they came in apples
- Policeman to little boy: “We are looking for a thief with a bicycle.” Little boy: “Wouldn’t you be better using your eyes.”
- Why is the teacher wearing sun-glasses. Because the class is so bright.
- A car owner after coming back from a party finds the sticker “parking fine” on his car. He goes and thanks the policeman for appreciating his parking skill.

Why is Language Ambiguous?

- Why not have a unique linguistic expression for every possible conceptualization that could be conveyed
- This would make language overly complex and linguistic expressions unnecessarily long.
- Allowing resolvable ambiguity permits shorter linguistic expressions, i.e. data compression.
- Language relies on people's ability to use their knowledge and inference abilities to properly resolve ambiguities.
- Infrequently, disambiguation fails, i.e. the compression is lossy.

Linguistic Methodology

- Descriptive / Empirical as opposed to Prescriptive
  - Linguists are interested in what people do say (or write)

- Generative Methodology: Noam Chomsky published Syntactic Structures in 1957.
  - generative linguists started out with a metatheory of what grammars of human languages look like and attempted to express specific grammars within this metatheory.

Generative grammar

- Generative grammars consist of finite sets of rules which should predict all and only the infinite grammatical sentences of a given human language (and what is conveyed about their meaning by their grammatical structure).
- Much of the theory of parsing and compiling programming languages has its antecedents in early generative linguistics (The Chomsky Hierarchy etc.)

Natural Language Tasks

- Processing natural language text involves many various syntactic, semantic and pragmatic tasks in addition to other problems.

SYNTACTIC TASKS

Word Segmentation

- Breaking a string of characters (graphemes) into a sequence of words.

Morphological Analysis

- Morphological analysis is the task of segmenting a word into its morphemes:
  - carried $\Longrightarrow$ carry + ed (past tense)
  - independently $\Longrightarrow$ in + (depend + ent) + ly
  - Googlers $\Longrightarrow$ (Google + er) + s (plural)
  - unlockable $\Longrightarrow$ un + (lock + able) ?

    $\Longrightarrow$ (un + lock) + able ?

Part Of Speech (POS) Tagging

- Annotate each word in a sentence with a part-of-speech.

I	ate	the	spaghetti	with	meatballs.
Pro	V	Det	N	Prep	N

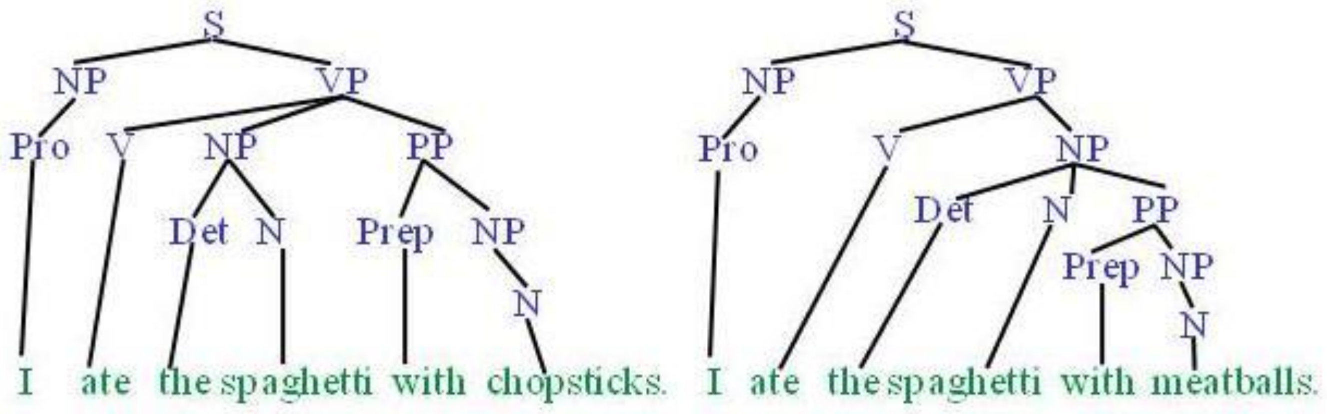
John	saw	the	saw	and	decided	to	take	it	to	the	table.
PN	V	Det	N	Con	V	Part	V	Pro	Prep	Det	N

Phrase Chunking

- Find all non-recursive noun phrases (NPs) and verb phrases (VPs) in a sentence.
  - [NP I] [VP ate] [NP the spaghetti] [PP with] [NP meatballs].
  - [NP He] [VP reckons] [NP the current account deficit] [VP will narrow] [PP to] [NP only # 1.8 billion] [PP in] [NP September]

Syntactic Parsing

- Produce the correct syntactic parse tree for a sentence.


S
NP
VP
Pro
V
NP
PP
Det
N
Prep
NP
N
I ate the spaghetti with chopsticks.
S
NP
VP
Pro
V
NP
Det
N
PP
Prep
NP
N
I ate the spaghetti with meatballs.

SEMANTIC TASKS

Word Sense Disambiguation (WSD)

- The proper sense of each ambiguous word in a sentence must be determined.
  - Ellen has a strong interest in computational linguistics.
  - Ellen pays a large amount of interest on her credit card.

Semantic Role Labeling (SRL)

- For each clause, determine the semantic role played by each noun phrase that is an argument to the verb.

  agent patient source destination instrument

  - John drove Mary from Austin to Dallas in his Toyota Prius.
  - The hammer broke the window.

Semantic Parsing

- A *semantic parser* maps a natural-language sentence to a complete, detailed semantic representation (*logical form*).

- Example: Mapping an English database query to Prolog:

  How many cities are there in the US?

  ```
  answer(A, count(B, (city(B), loc(B, C),
                      const(C, countryid(USA))),
                 A))
  ```

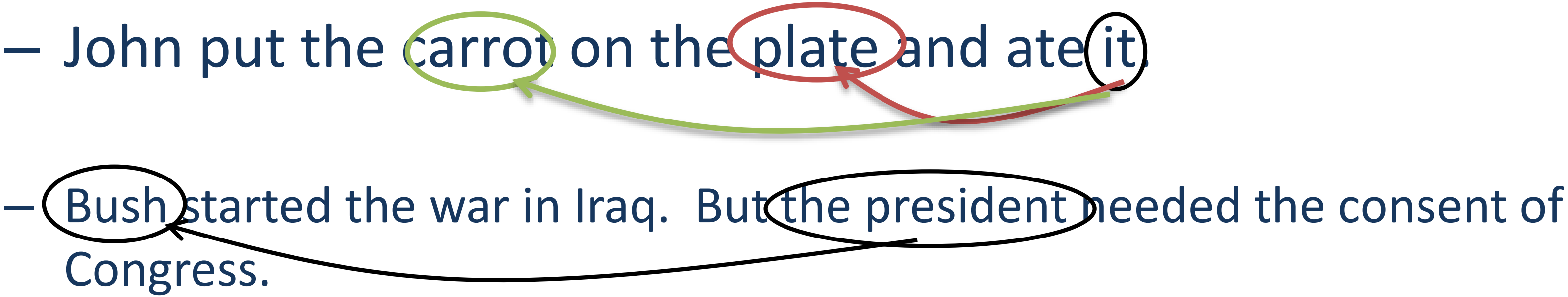
Textual Entailment

- Determine whether one natural language sentence entails (implies) another under an ordinary interpretation.
  - Eyeing the huge market potential, currently led by Google, Yahoo took over search company Overture Services Inc last year.
  - Yahoo bought Overture

PRAGMATICS/DISCOURSE TASKS

Anaphora Resolution/ Co-Reference

- Determine which phrases in a document refer to the same underlying entity.



  - John put the carrot on the plate and ate it.
  - Bush started the war in Iraq. But the president needed the consent of Congress.
  - Today was Jack's birthday. Penny and Janet went to the store. They were going to get presents. Janet decided to get a kite. "Don't do that," said Penny. "Jack has a kite. He will make you take it back."

Ellipsis Resolution

- Frequently words and phrases are omitted from sentences when they can be inferred from context.

"Wise men talk because they have something to say; fools, because they have to say something.“ (Plato)

"Wise men talk because they have something to say; fools talk because they have to say something.“ (Plato)

Other Tasks

- Information Extraction (IE)
- Question Answering
- Reading Comprehension
- Text Summarization
- Machine Translation
- Ambiguity Resolution

Manual Knowledge Acquisition

- Traditional, “rationalist,” approaches to language processing require human specialists to specify and formalize the required knowledge.
  - difficult, time-consuming, and error prone.
  - “Rules” in language have numerous exceptions and irregularities.
    - “All grammars leak.”: Edward Sapir (1921)
- Manually developed systems were expensive to develop and their abilities were limited and “brittle”

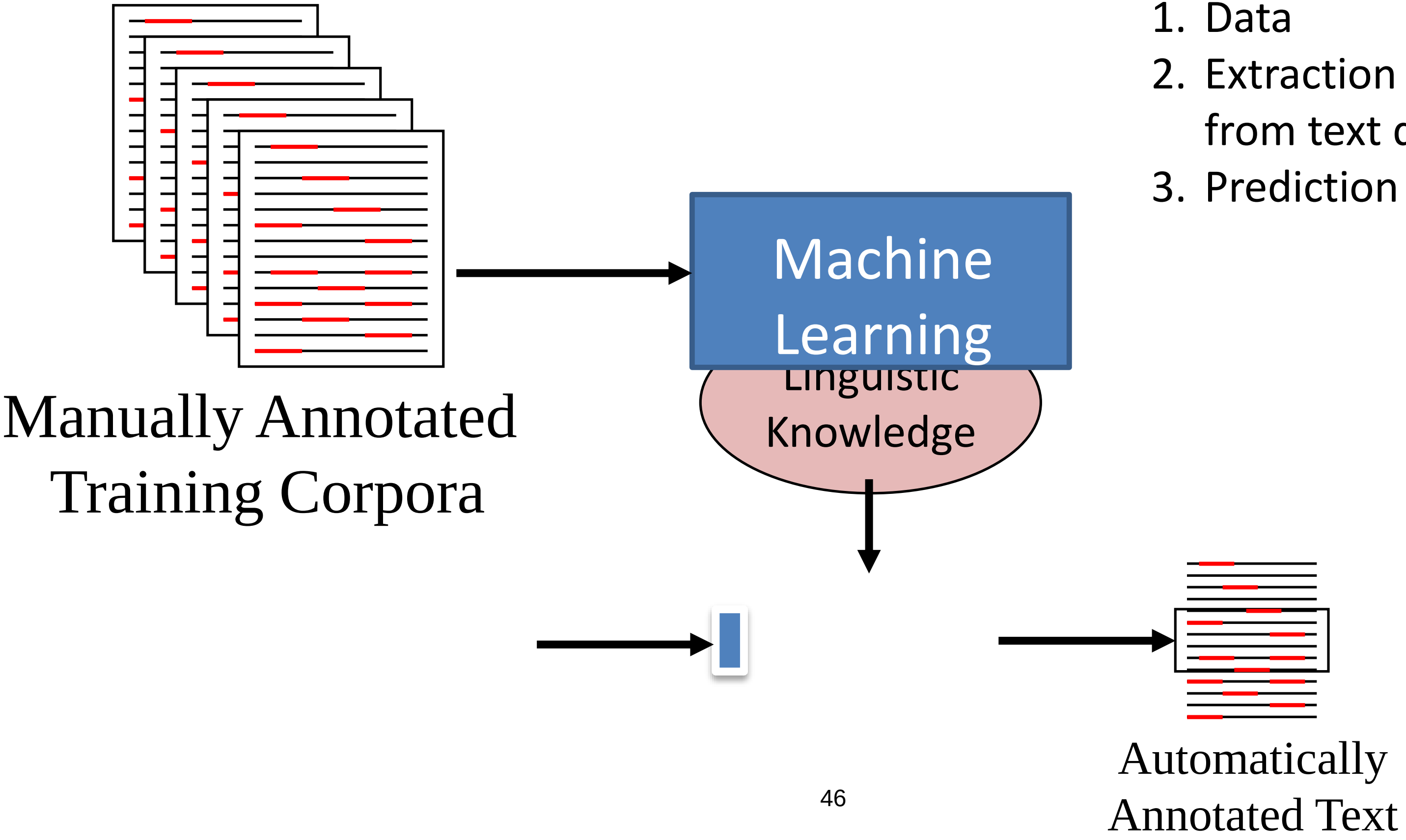
Automatic Learning Approach

- Use machine learning methods to automatically acquire the required knowledge from appropriately annotated text corpora.
  - the “corpus based,” “statistical,” or “empirical” approach

- During the 1990’s, the statistical training approach expanded and came to dominate almost all areas of NLP.

Machine Learning Approach

1. Data
2. Extraction of “features” from text data
3. Prediction of output



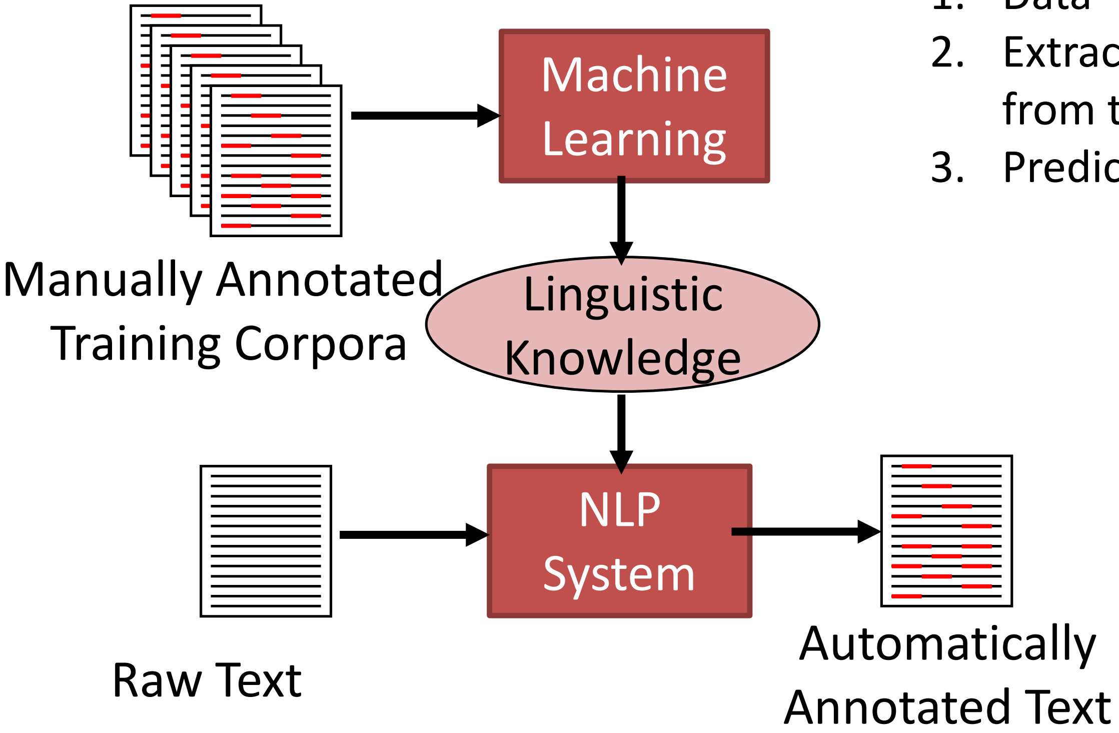
Manually Annotated Training Corpora

Machine Learning

Linguistic Knowledge

Automatically Annotated Text

Machine Learning Approach


Machine Learning
Manually Annotated Training Corpora
Linguistic Knowledge
Raw Text
NLP System
Automatically Annotated Text

1. Data
2. Extraction of “features” from text data
3. Prediction of output

Advantages of the Learning Approach

- Large amounts of electronic text available.
- Annotating corpora is easier and requires less expertise.
- Learning algorithms are now able to handle large amounts of data and produce accurate probabilistic knowledge.
- The probabilistic knowledge acquired allows robust processing that handles linguistic regularities as well as exceptions.

The Importance of Probability

- Unlikely interpretations of words can combine to generate spurious ambiguity:
  - “Time flies like an arrow” has 4 parses, including those meaning:
    - Insects of a variety called “time flies” are fond of a particular arrow.
    - A command to record insects’ speed in the manner that an arrow would.
- Some combinations of words are more likely than others:
  - “vice president Gore” vs. “dice precedent core”
- Statistical methods allow computing the most likely interpretation by combining probabilistic evidence from a variety of uncertain knowledge sources.

Human Language Acquisition

- Human children learn languages from experience.
- However, it is controversial to what extent prior knowledge of “universal grammar” (Chomsky, 1957) facilitates this acquisition process.
- Computational studies of language learning may help us to understand human language learning
- Existing empirical results indicate that a great deal of linguistic knowledge can be effectively acquired from reasonable amounts of real linguistic data without specific knowledge of a “universal grammar.”

Pipelining Problem

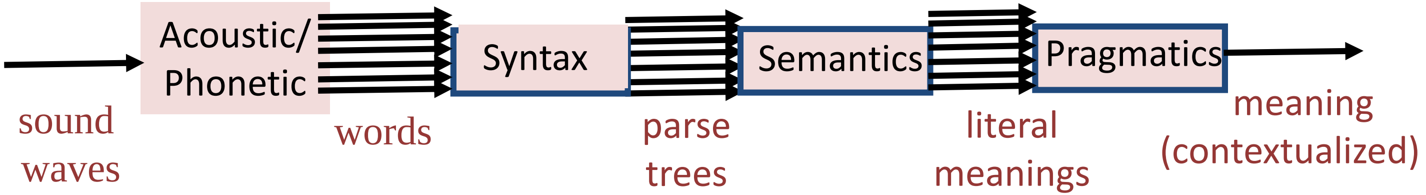
- Assuming separate independent components for speech recognition, syntax, semantics, pragmatics, etc. allows for more convenient modular software development.
- However, frequently constraints from "higher level" processes are needed to disambiguate "lower level" processes.
  - Example of syntactic disambiguation relying on semantic disambiguation:
    - At the zoo, several men were showing a group of students various types of flying animals. Suddenly, one of the students hit the man **with** a **bat**.

Pipelining Problem (cont.)

- If a hard decision is made at each stage, cannot backtrack when a later stage indicates it is incorrect.
  - If attach “with a bat” to the verb “hit” during syntactic analysis, then cannot reattach it to “man” after “bat” is disambiguated during later semantic or pragmatic processing.

Increasing Module Bandwidth

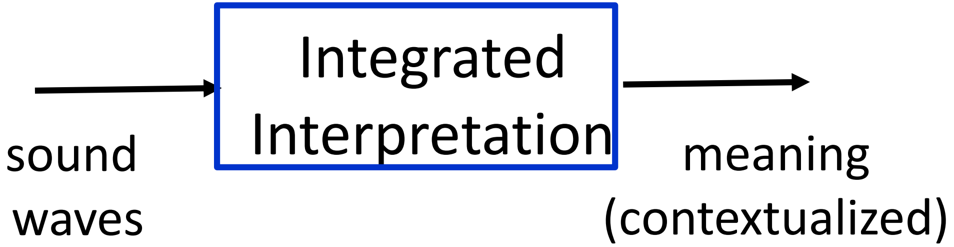
- If each component produces multiple scored interpretations, then later components can rerank these interpretations.



- Problem: Number of interpretations grows combinatorially.
- Solution: Efficiently encode combinations of interpretations.
  - Word lattices
  - Compact parse forests

Global Integration/ Joint Inference

- Integrated interpretation that combines phonetic/ syntactic/ semantic/ pragmatic constraints.



- Difficult to design and implement.
- Potentially computationally complex.