Chapter 4

Natural Language Processing

What is NLP?

Natural Language Processing (NLP) is a both a modern computational technology and a method of investigating and evaluating claims about human language itself.

Also called **Computational Linguistics** which links to Artificial Intelligence (AI), the general study of cognitive function by computational processes, normally with an emphasis on the role of knowledge representations, that is to say the need for representations of our knowledge of the world in order to understand human language with computers.

What is Text Processing?

Text processing is a process of **manipulating a written text in a way that will be useful** for further processing or higher level of **NLP application**. Text processing might have different scope based on the *application domain* or NLP type.

- *Understanding* text, *identifying* relevant elements, *manipulating* elements of a text and *analyzing* the structure and semantics of text elements is vital.

Language Technology

mostly solved

Spam detection

Let's go to Agra! ✓

Buy DraG… ✗

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! 👍

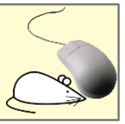
The waiter ignored us for 20 minutes. 👎

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 Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?

Background

Solving the language-related problems, is the main concern of the fields known as Natural Language Processing, Computational Linguistics, and Speech Recognition and Synthesis

Few applications of language processing

- Spelling correction,
- Grammar checking,
- Information retrieval, and
- Machine translation,
- Speech processing, etc.

Knowledge in NLP

- Tasks of being capable of analyzing an incoming audio signal and recovering the exact sequence of words and generating its response require knowledge about **phonetics and phonology**, which can help model how words are pronounced in colloquial speech.
- Producing and recognizing the variations of individual words (e.g., recognizing that *doors* is plural) requires knowledge about **morphology**, which captures information about the shape and behavior of words in context.

Cont..

Syntax: the knowledge needed to order and group words together

I'm I do, sorry that afraid Dave I'm can't.

(Dave, I'm sorry I'm afraid I can't do that.)

Lexical semantics: knowledge of the meanings of the component words

Compositional semantics: knowledge of how these components combine to form larger meanings

data + base = database

Cont..

Pragmatics: the appropriate use of the kind of polite and indirect language.

You’re smart! You got 10 out of 100.

Discourse conventions: knowledge of correctly structuring these such conversations (intonation, gesturer, style, speech act, etc)

Dave, I’m sorry I’m afraid I can’t do ***that***.

✓The word “that” is referring to something which is not part of the sentences

Knowledge in Language Processing

- ✓**Phonetics and Phonology**: The study of linguistic sounds
- ✓**Morphology**: The study of the meaningful components of words
- ✓**Syntax**: The study of the structural relationships between words
- ✓**Semantics**: The study of meaning
- ✓**Pragmatics**: The study of how language is used to accomplish goals.
- ✓**Discourse**: The study of linguistic units larger than a single utterance.

Methods and Resources

❖Linguistic Knowledge

Linguistic knowledge resources for many languages are utilized: dictionaries, morphological and syntactic grammars, rules for semantic interpretation, pronunciation and intonation.

❖Corpora and Corpus Tools

Large collections of application-specific or generic collections of spoken and written language are exploited for the acquisition and testing of statistical or rule-based language models.

Approach to NLP

❖ **Rule Based (Hand Crafted Rules)**

Develop the rules to process the natural languages based on known facts and exceptions

❖ **Machine Learning**

Capture rules from examples and apply on new instances

- Supervised: learn by comparing with expected output
- Unsupervised: blind learning.

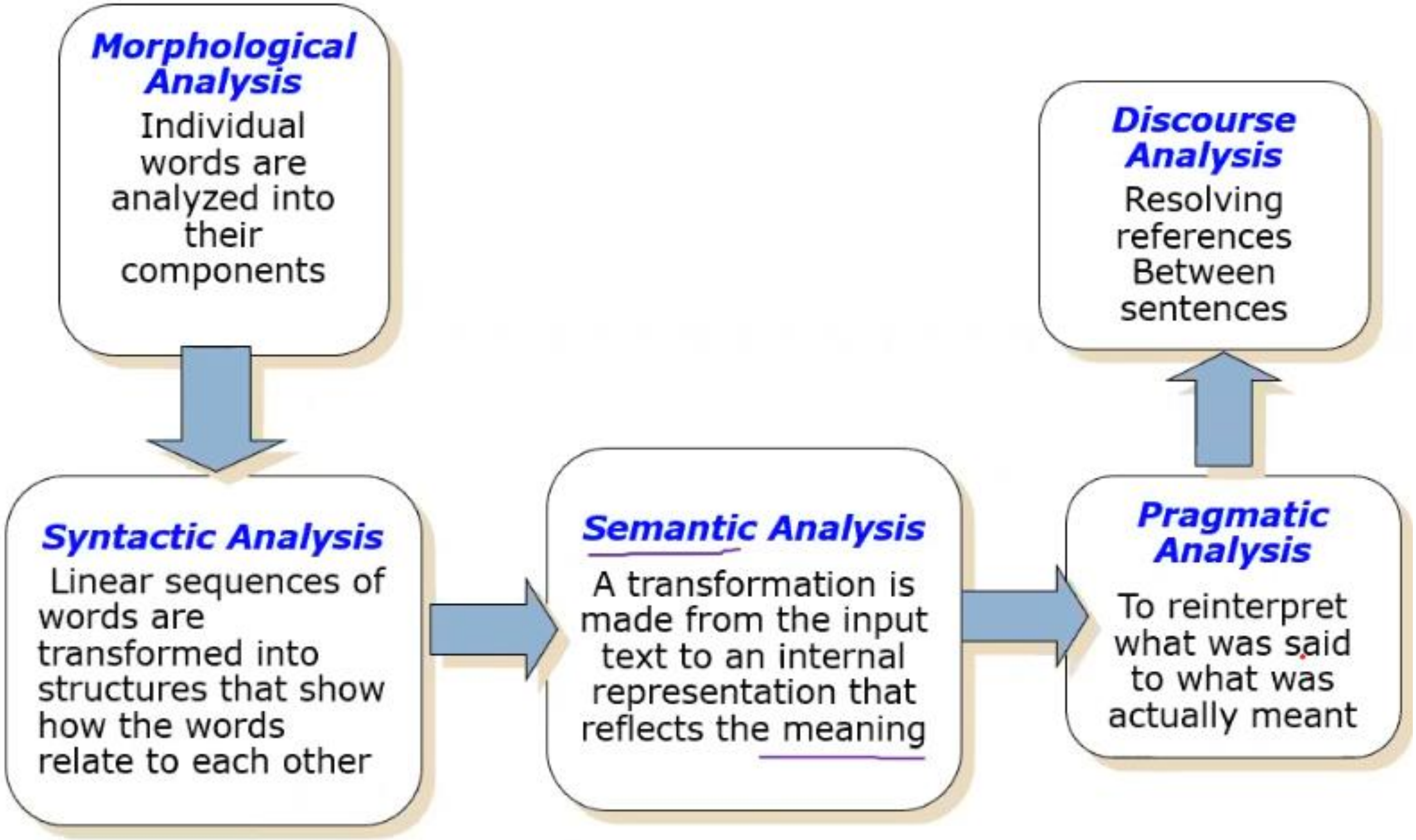
Cont..

- **Machine learning** research has focused on ways to automatically learn the various representations described above;

  - Automata, Rule Systems, Search Heuristics, Classifiers.

These systems can be trained on large corpora and can be used as a powerful modeling technique, especially in places where we don’t yet have good causal models.

Stages of NLP (Textual form)



Morphological Analysis
Individual words are analyzed into their components

Syntactic Analysis
Linear sequences of words are transformed into structures that show how the words relate to each other

Semantic Analysis
A transformation is made from the input text to an internal representation that reflects the meaning

Pragmatic Analysis
To reinterpret what was said to what was actually meant

Discourse Analysis
Resolving references Between sentences

Lecture 2.2

Morphology

What is Morphology?

- **Morphology** deals with the syntax of complex words and parts of words, also called ***morphemes***, as well as with the semantics of their lexical meanings.
- Understanding how ***words are formed*** and what ***semantic properties*** they convey through their forms *enables human beings to easily recognize individual words and their meanings* in discourse.

Cont..

Morphology is the branch of linguistics that studies the structure of words.

In English and many other languages, many words can be broken down into parts. For example:

- Un-happi-ness
- Madaxweyn-aha

Cont..

The smallest unit which has a meaning or grammatical function that words can be broken down into are known as **morphemes**.

Morphemes are classified into two types:

- **Free Morphemes**: in Af-Soomaali, words like Madax, Weyne, Inan, Hooyo IWM, and In English words, like; girl, boy, mother, etc. These are words with a complete meaning, so they can stand alone as an *independent word* in a sentence.
- **Bound Morphemes**: These are lexical items incorporated into a word as a *dependent part*. They cannot stand alone, but must be connected to another morpheme to give meaning. For Instance, In Af-Soomali, words like; aha, ooyin, ka, tii IWM. Un & ness.

Word Formation Methods

(1) Affixation is concerned with the way morphemes are connected to existing lexical forms as attachments to show different grammatical feature.

We distinguish affixes of various types:

Prefixes - attached at the beginning of a lexical item or base-morpheme –

e.g. ma, waan, ka, soo

Suffixes – attached at the end of a lexical item

e.g. yaa, sha, ha

(2) Compounding, words can be created by Compounding, which is forming new words from two or more independent words: the words can be free morphemes, words derived by affixation, or even words formed by compounds themselves.

e.g. textbook, database, air-condition

Cont..

(3) Reduplication, which is forming new words either by doubling an entire free morpheme (total reduplication) or part of a morpheme (partial reduplication).

e.g. in *Af-Soomaali*, the word "**jajabay**"

(4) Derivational morphemes create or *derive* new words by changing the meaning or the *word class* of the word (change verb into noun), while

(5) Inflectional morphemes creates a word with similar meaning but more grammatical feature without affecting the word class. For example:

happy	→	unhappy	(Inflectional)

Both words are adjectives, but the meaning changes.

quick	→	quickness	(Derivational)

The affix changes both meaning and word class - adjective to a noun.

Lemmatization and Stemming

- **Lemmatization** usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the **lemma**.
- Lemmatization is the process of identifying lexical/dictionary term after removing all affixes.

Morphological Analysis in its general form involves recovering the LEMMA of a word and all its affixes, together with their grammatical properties.

- **Stemming** a simplified form of morphological analysis – simply find the stem.

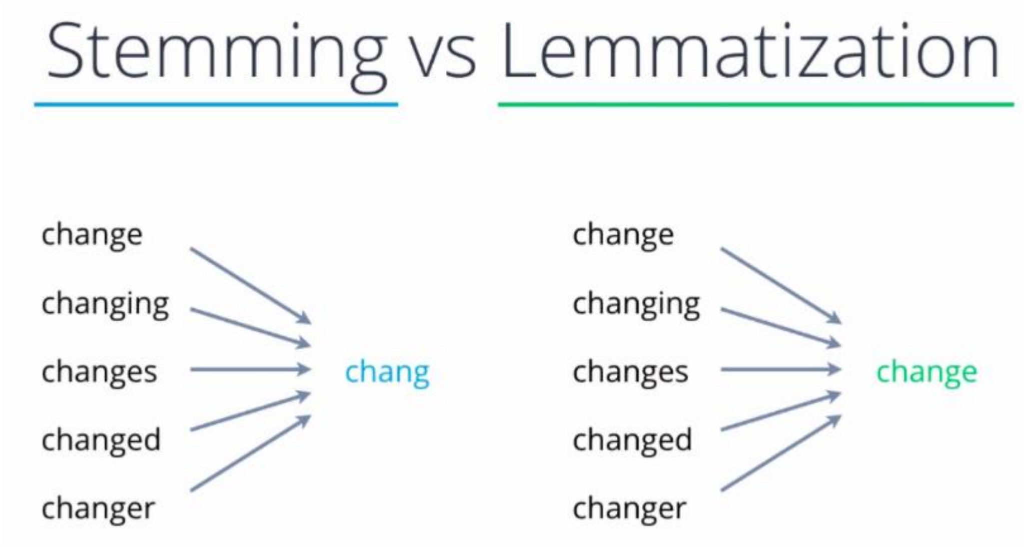
Cont..

- The goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form.
- 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.



Roots & Stem

Roots:

The root is generally the principle carrier of the lexical meaning of a word, while affixes generally carry grammatical meanings.

For example, in cats, the root cat carries the basic meaning, while -s carries the grammatical information 'plural.'

Stems:

In addition to roots, we also distinguish stems. A stem may be also a root, as cat in cats.

Exercise: Somali Root vs Stem?

Tokenization

Tokenization is the process of tokenizing or splitting a string, text into a list of tokens. One can think of token as parts like a word is a token in a sentence, and a sentence is a token in a paragraph.

The are two types of tokenization:

- Word Tokenization
- Sentence Tokenization

Cont..

they lay back on the San Francisco grass and looked at the stars and their

Type: an element of the vocabulary.

Token: an instance of that type in running text.

How many?

- 15 tokens
- 13 types

Practical Sessions of this Lecture!

First thing first, Install NLTK for your Computer. Then try to do:

Tokenization in NLP

Stemming in NLP

Lemmatization in NLP

Syntax, POS Tagging & Parsing

Syntax

Syntax, is the study of grammatical relations between words and other units within the sentence.

- Study of structure of sentence in a language
- Word order or subconscious grammatical knowledge
- Refers to the way words are arranged together, and the relationship between them.
- Roughly, goal is to relate surface form (what we perceive when someone says something) to semantics (what that utterance means)
- Representational device is tree structure

Syntax useful for: -

- Grammar checkers
- Question answering
- Information extraction
- Machine translation

Constituency

- How would the blocks relate to one another? e.g.: I hit the man with a stick
  - Two possibilities:
    - I hit [the man with a stick]
    - I hit [the man] with a stick
- Af Somali Exercise:

Write down three examples of Somali constituency

Two kinds of ambiguity:

- She called her friend from Australia.
  - STRUCTURAL AMBIGUITY
    - [She called] [her friend] [from Australia].
    - [She called] [her friend from Australia].

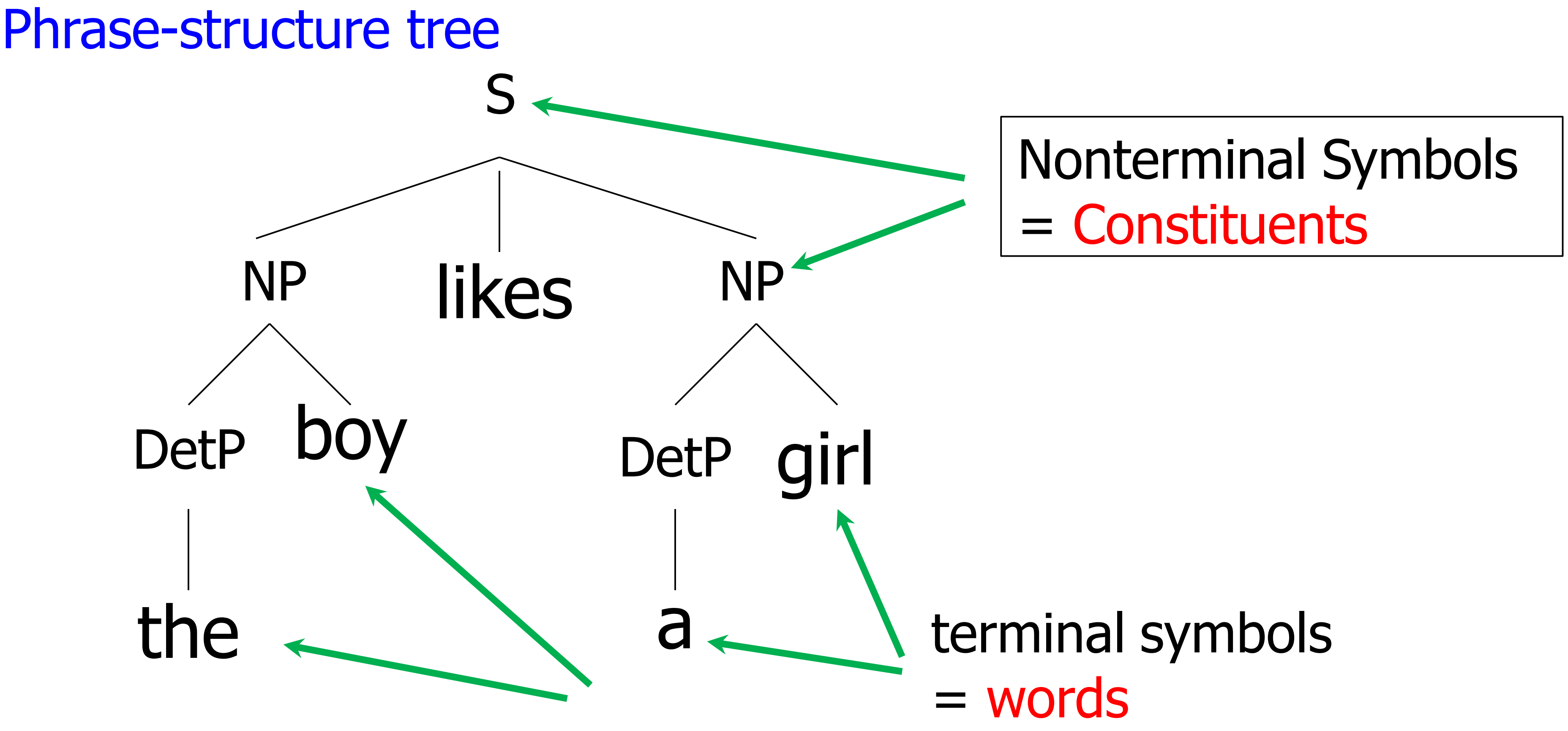
- We went down to the bank yesterday
  - LEXICAL AMBIGUITY
    - [bank] river bank
    - [bank] financial institute bank

Basic Word Order

- SVO (English, Chinese)
  - *The boy saw the man.*
- SOV (Amharic, Russian, Turkish, Japanese)
  - *Pensive poets painful vigils keep.* (*Pope*)
  - ልጁ ቤቱ ሄደ ። አበበ አልማዝን መታት
- VSO (Irish, Arabic, Welsh)
  - *Govern thou my song.* (*Milton*)

Types of Nodes

- (((the/Det) boy/N) likes/V ((a/Det) girl/N))



Phrase-structure tree

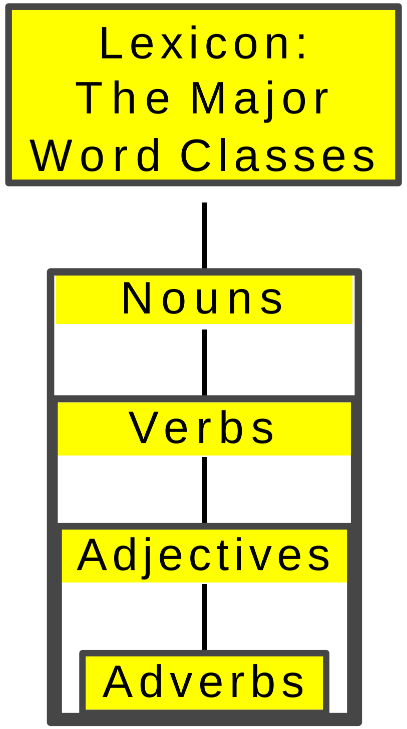
Nonterminal Symbols
= Constituents

terminal symbols
= words

Determining Part-of-Speech

- Determining part of speech is crucial for building the hierarchical structure of sentences.

The Lexicon



Context-Free Grammars

- Defined in formal language theory
- Composed of
  - Terminals,
  - nonterminals,
  - start symbol, and
  - rules
- CFG is a String-rewriting system/method
- Start with start symbol, rewrite using rules, done until only terminals are left
- ***NOT A LINGUISTIC THEORY***, just a formal device

CFG: Example

- Many possible CFGs for English, here is an example (fragment):
  - S → NP VP
  - VP → V NP
  - NP → DetP N | AdjP NP
  - AdjP → Adj | Adv AdjP
  - N → boy | girl
  - V → sees | likes
  - Adj → big | small
  - Adv → very
  - DetP → a | the

the very small boy likes a girl

Derivations in a CFG

the boy likes a girl

S → NP VP

VP → V NP

NP → DetP N | AdjP

NP AdjP → Adj | Adv

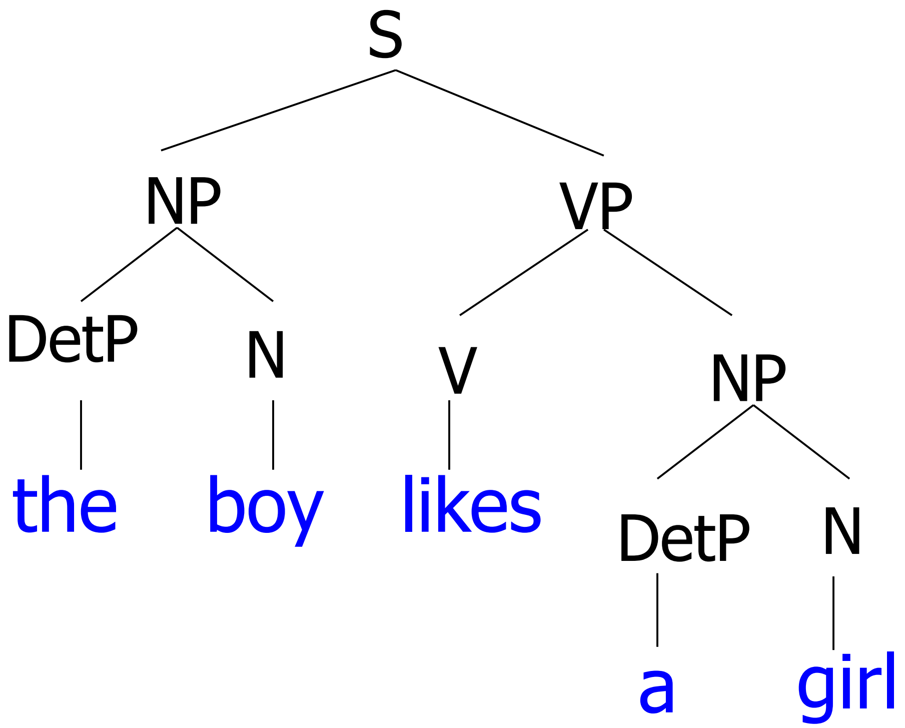
AdjP N → boy | girl

V → sees | likes

Adj → big | small

Adv → very

DetP → a | the

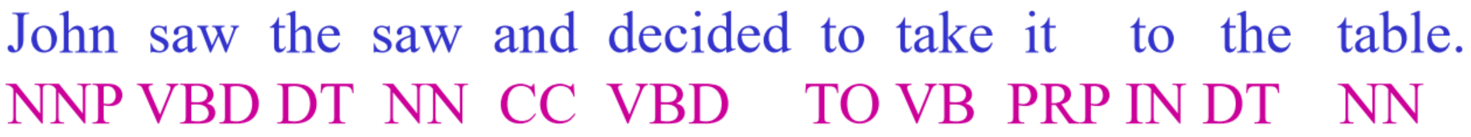


Part Of Speech Tagging

- ❑ Syntax requires word classes to be identified
- ❑ Words can be divided into classes that behave similarly.
  - ❖ Traditionally eight parts of speech:
    - ✓ noun, verb, pronoun, preposition, adverb, conjunction, adjective and article
- ❑ They tell us a lot about a word (and the words near it).
- ❑ Tell us what words are likely to occur in the neighborhood
  - ▪ adjectives often followed by nouns
  - ▪ personal pronouns often followed by verbs (you, he, she, it..)
  - ▪ possessive pronouns by nouns (yours, his, hers, its,….

Part of Speech Tagging

- □ **PoS Tagging** is the process of annotating each word in a sentence with a part-of-speech marker.
- □ Lowest level of syntactic analysis is PoS Tagging.



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

- □ Useful for subsequent syntactic parsing and word sense disambiguation.

Tagging Terminology

- **Tagging**
  - The process of associating labels with each token (word) in a text
- **Tags**
  - The labels (Noun, Verb, Adjective, etc)
- **Tag Set**
  - The collection of tags used for a particular task

Common Tagsets

- Brown corpus: 87 tags
- Penn Treebank: 45 tags
- Lancaster UCREL C5 (used to tag the British National Corpus - BNC): 61 tags
- Lancaster C7: 145 tags

Word Class/Categories

- Word categories: also called parts of speech
  - *Noun*: Names of things boy, cat, truth
  - *Verb*: Action or state become, hit
  - *Pronoun*: Used for noun like I, you, we
  - *Adjective*: modifies noun happy, clever
  - *Adverb*: modifies V, Adj, Adv sadly, very
  - *Conjunction*: Joins things and, but, while
  - *Preposition*: Relation of N to, from, into
  - *Interjection*: An outcry ouch, oh, alas, psst

POS Tagging Approaches

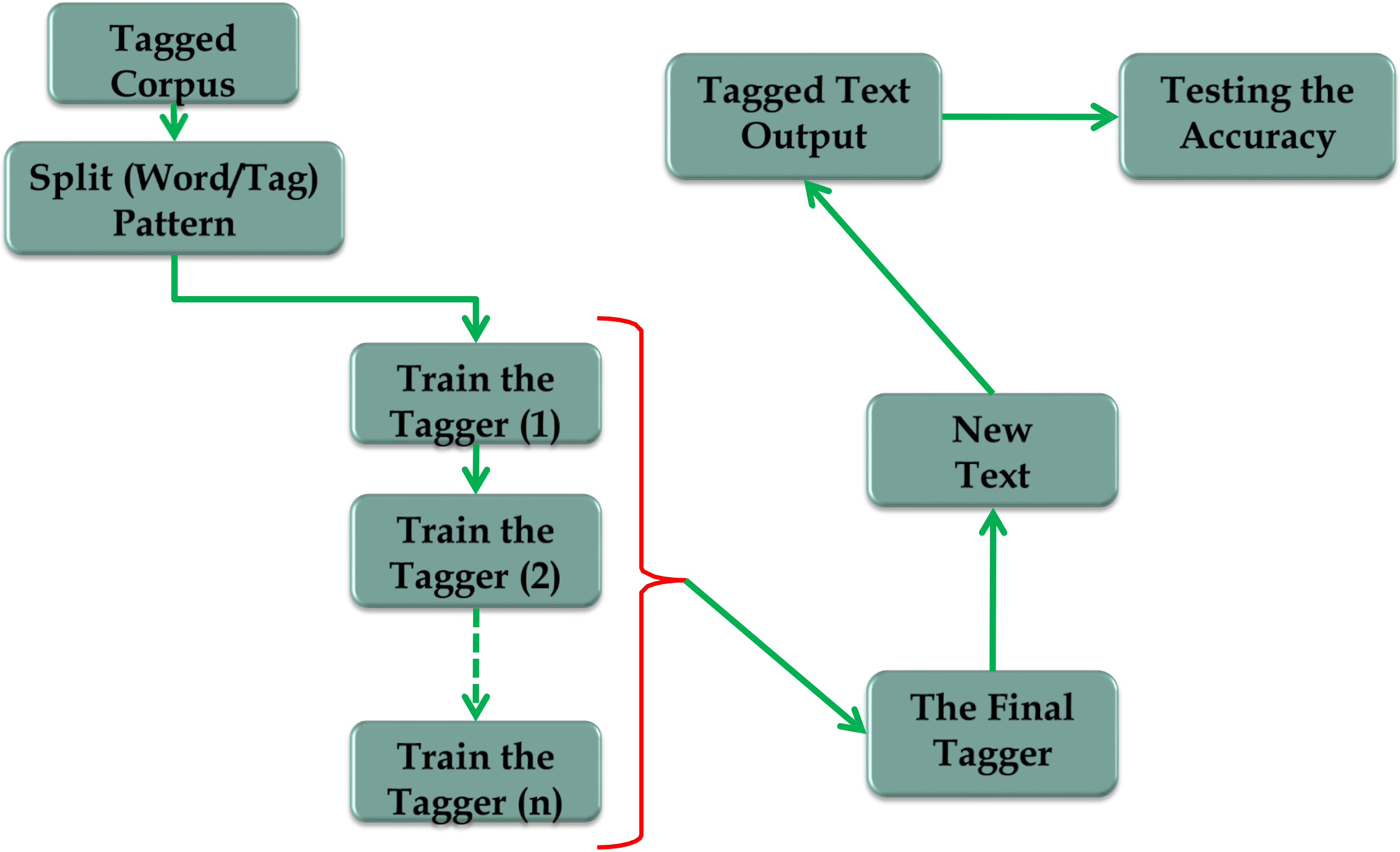
- **Rule-Based**: Human crafted rules based on lexical and other linguistic knowledge.
- **Learning-Based**: Trained on human annotated corpora like the Penn Treebank.
  - **Statistical models**: Hidden Markov Model (HMM), Maximum Entropy Markov Model (MEMM), Conditional Random Field (CRF)
  - **Rule learning**: Transformation Based Learning (TBL)
- Generally, learning-based approaches have been found to be more effective overall, taking into account the total amount of human expertise and effort involved.

Stochastic Tagging

- Based on probability of certain tag occurring given various possibilities
  - *Requires a training corpus*
  - *No probabilities for words not in corpus.*
  - *Training corpus may be different from test corpus.*

- Simple Method: Choose most frequent tag in training text for each word!
  - **Result: 90% accuracy**
  - **Unknown for words never encountered before**
  - **HMM is an example**

Setting the Scene



Parsing

- Parsing is the process of recognizing and assigning STRUCTURE
- Parsing a string with a CFG:
  - Finding a derivation of the string consistent with the grammar
  - The derivation gives us a **Parse Tree**

Parsing

- ✓ *A parser processes input sentences according to the productions of a grammar, and builds one or more constituent structures that conform to the grammar.*
- ✓ *A parser is a procedural interpretation of the grammar. It searches through the space of trees allowed by a grammar to find one that has the required sentence along its edge.*

Parsing

- ✓ *Parsing is the process of taking a string and a grammar and **returning parse tree(s)** for that string*
- ✓ A parser permits a grammar to be evaluated against a collection of test sentences
- ✓ A parser can also be used to check the permissibility of a sentences
- ✓ *A parser can serve as a model of psycholinguistic processing, helping to **explain the difficulties that humans have with processing certain syntactic constructions**.*

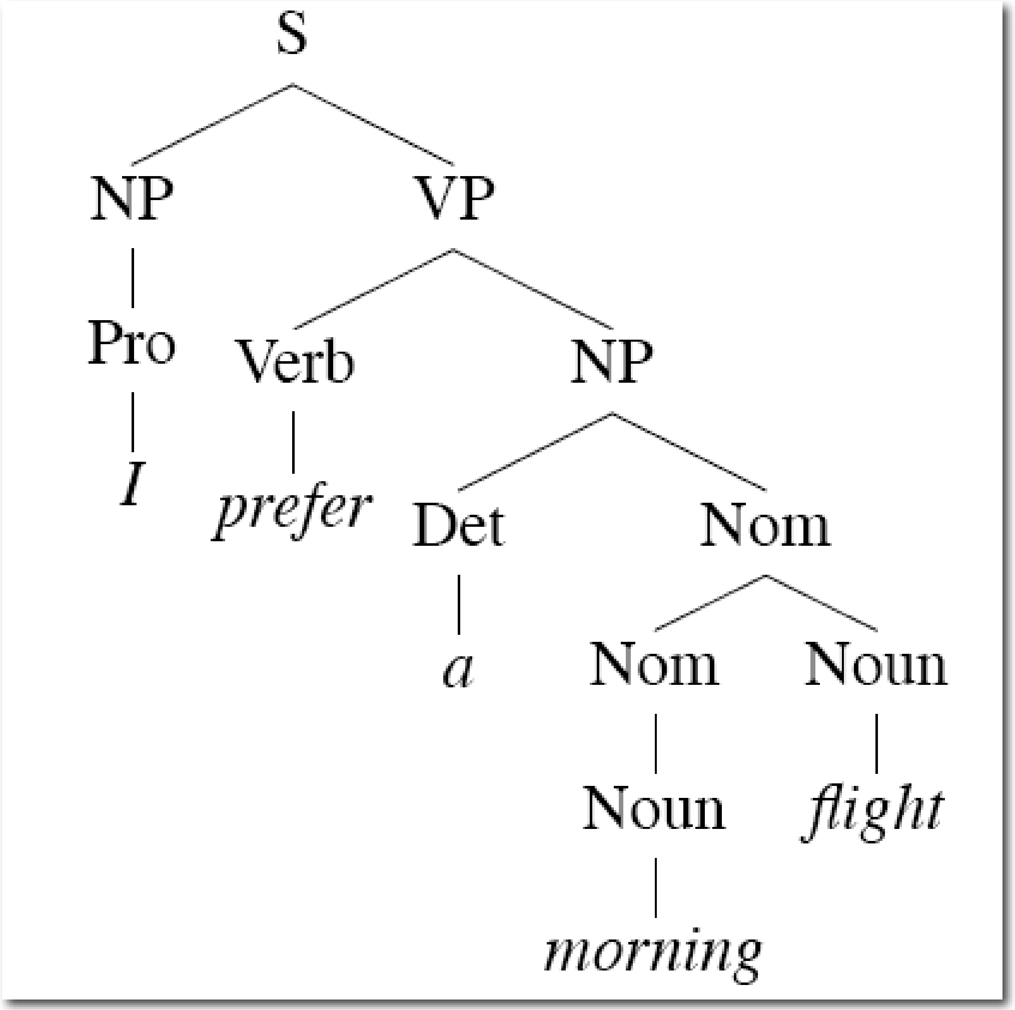
Parsing as Search

- Search within a space defined by
  - Start State
  - Goal State
  - State to state transformations
- Two distinct parsing strategies:
  - Top down
  - Bottom up
- Different parsing strategy, different state space, different problem.

Derivations

□ A *derivation* is a sequence of rules applied to a string that *accounts* for that string (sequence of words)

1. **Covers all the elements in the string**
2. **Covers only the elements in the string**



Top-Down Parsing Method

- Recursive Descent Parsing
  - ✓ break a high-level goal into several lower-level sub-goals
  - ✓ First question will be how to break the top level goal?
  - ✓ The top-level goal is to find an **S** **… … .** **Sentences**.
  - ✓ For the grammar, the $S \rightarrow NP\ VP$ production permits the parser to replace this goal with two sub-goals:
    - ✓ *find an NP, then*
    - ✓ *find a VP.*
    - ✓ *Then replace VP and NP with others until we reach a terminal*

Top-Down Parsing Method

- Recursive Descent Parsing
  - ✓ Keep doing this until a terminal is found and compare the terminal with the input string.
    - ✓ If no match then backup and look other alternatives
  - ✓ Once a parse has been found, we can get the parser to look for additional parses.
    - ✓ … in case the sentences has more than one possible structure
  - ✓ Top-down parsers use a grammar to predict what the input will be, before inspecting the input.
    - ✓ Check the part of speech before the word itself
  - Demo: **nltk.app.rdparser()**
- Recursive Descent Parsing in NLTK:
  - **nltk.RecursiveDescentParser(yourGrammar)**

Bottom-Up Parsing Method

□ Shift-Reduce Parsing

- ✓ shift-reduce parser tries to find sequences of words and phrases that correspond to the right hand side of a grammar production, and replace them with the left-hand side, until the whole sentence is reduced to an **S**.
- ✓ Since the input is available to the parser all along, it would be more sensible to consider the input sentence from the very beginning.
- ✓ This approach is called bottom-up parsing

Bottom-Up Parsing Method

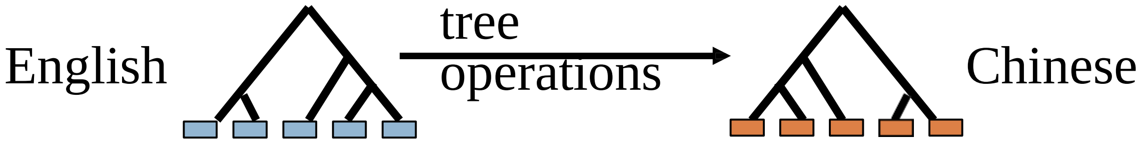
- Shift–reduce parsing is a bottom up derivation strategy, that is, it starts from the words in the string, and tries to work upwards towards the root symbol in the grammar.
- parse(sent):
  - if sent is [S] then finish
  - otherwise, for every rule, check if the RHS of the rule matches any substring of the sentence
  - if it does, replace the substring in the LHS of the rule
  - continue with this sentence
- Demo: **nltk.app.srparser()**

Top Down vs Bottom Up Searching

- The search has to be guided by the INPUT and the Grammar
- TOP-DOWN search: the parse tree has to be rooted in the start symbol S
  - EXPECTATION-DRIVEN parsing
- BOTTOM-UP search: the parse tree must be an analysis of the input
  - DATA-DRIVEN parsing

Applications of parsing

- Machine translation (Alshawi 1996, Wu 1997, ...)



- Speech recognition using parsing (Chelba et al 1998)

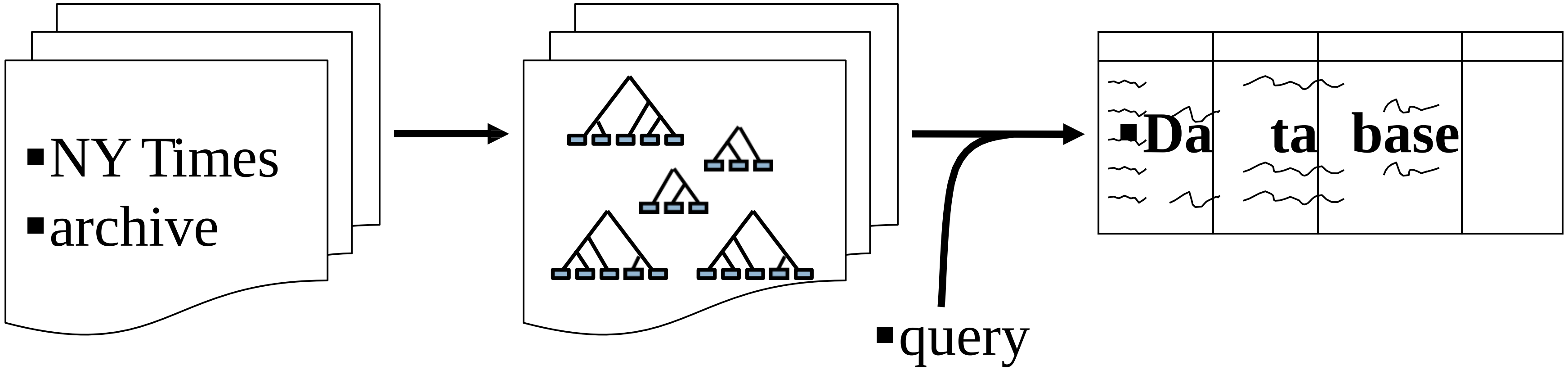
  Put the file in the folder.

  Put the file and the folder.

Applications of parsing

- Grammar checking (Microsoft)

- Information extraction (Hobbs 1996)



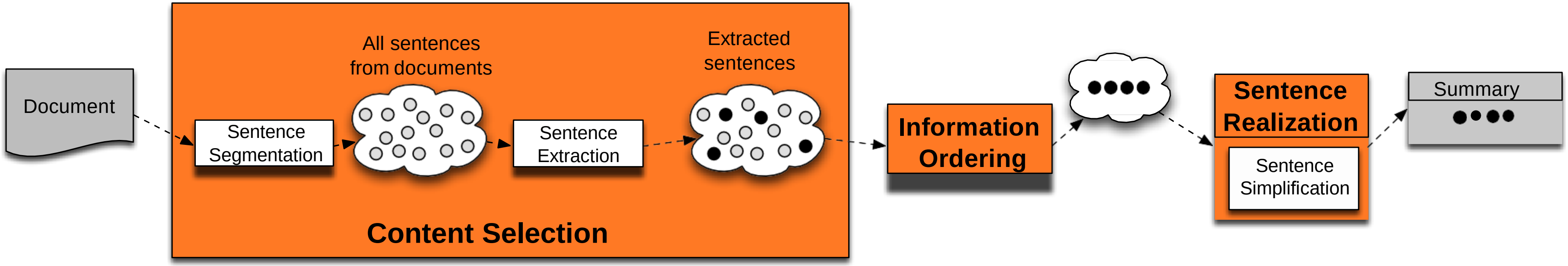
Text Summarization

- **Goal**: produce a reduced version of a text that contains information that is important or relevant to understand the content.

- **Summarization Applications**
  - **outlines or abstracts** of any document, article, etc
  - **summaries** of email threads
  - **action items** from a meeting
  - **simplifying** text by compressing sentences
  - News **summarization**

Summarization: Three Stages

1. content selection: choose sentences to extract from the document
2. information ordering: choose an order to place them in the summary
3. sentence realization: clean up the sentences



Practical Sessions of this Lecture!

Let’s try to do:

CFG in NLP

POS Tagging in NLP

N-grams in NLP

Plagiarism Checker in NLP

The End