Text Analytics

MODULE 4: ESSENTIAL LINGUISTICS & NATURAL LANGUAGE PROCESSING

Dr. Fan Zhenzhen

Institute of Systems Science

National University of Singapore

Email: zhenzhen@nus.edu.sg

Objectives

At the end of this module, you will:

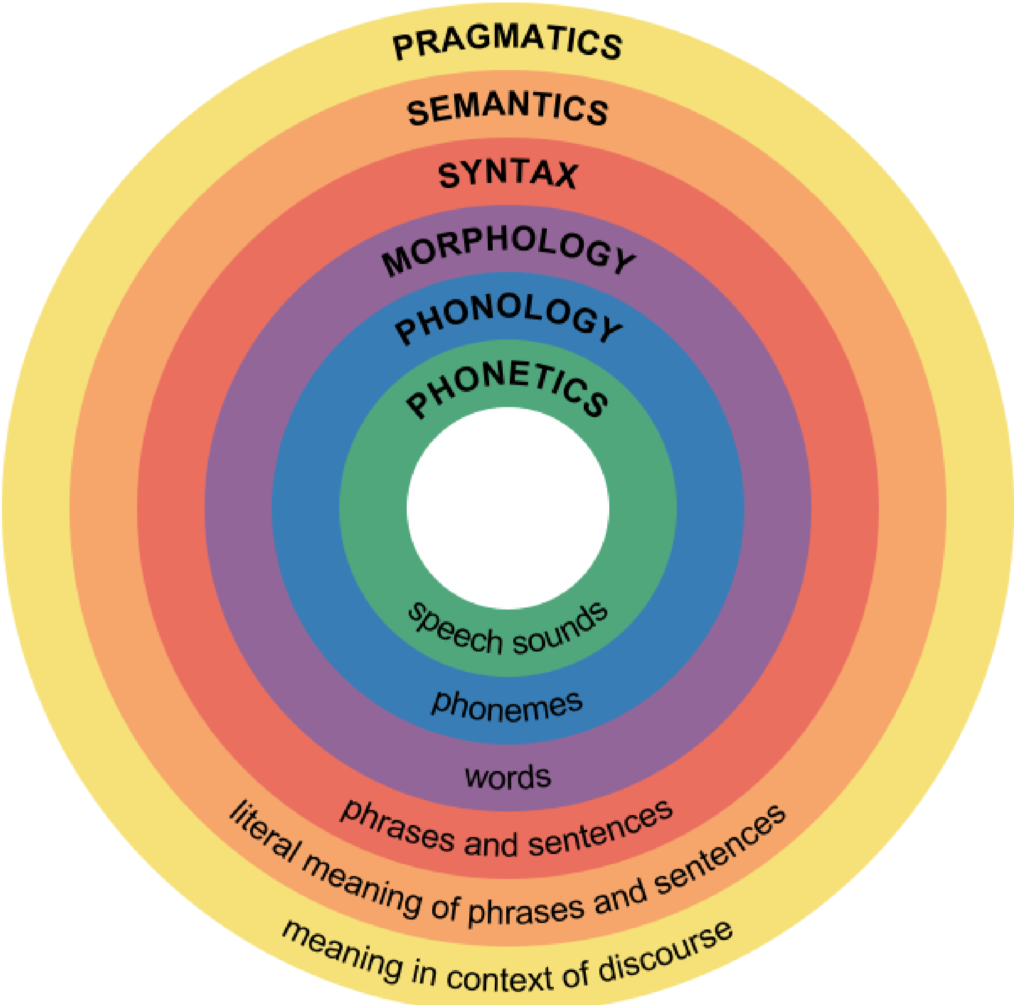
- Have essential linguistic knowledge for text processing
- Appreciate the NLP tasks that some TA systems perform.

Outline

- Essential Linguistics
- Natural language processing (NLP) tasks for TA

Essential Linguistics

Linguistics – the scientific study of language



https://courses.lumenlearning.com/boundless-psychology/chapter/introduction-to-language/

Morphology

- The structure of words and their part-of-speech (POS, major lexical syntactic categories)
  - Open-class, or content words: nouns, verbs, adjectives, adverbs,
  - Closed-class, or functional words: pronouns, determiners, prepositions, conjunctions, pronouns, numerals, auxiliary verbs, etc.
- Parts of speech group words that have similar neighbouring words (their distributional properties) or take similar affixes (their morphological properties).
- Many words are ambiguous between multiple lexical categories (with >1 POS) E.g. "*book*" can be a noun ("*my* ***book***") or a verb ("*to* ***book*** *a room*")

POS Tags

- LDC Penn Tree Bank has 36 POS tags + 12 other tags with detailed information, e.g.

```
1. CC   Coordinating conjunction  25.TO  to
2. CD   Cardinal number           26.UH  Interjection
3. DT   Determiner                27.VB  Verb, base form
4. EX   Existential there         28.VBD Verb, past tense
5. FW   Foreign word              29.VBG Verb, gerund/present participle
6. IN   Preposition/subord.       30.VBN Verb, past participle
218z      conjunction
7. JJ   Adjective                 31.VBP Verb, non-3rd ps. sing. present
8. JJR Adjective, comparative     32.VBZ Verb, 3rd ps. sing. present
9. JJS Adjective, superlative     33.WDT wh-determiner
10.LS   List item marker          34.WP  wh-pronoun
11.MD   Modal                     35.WP  Possessive wh-pronoun
12.NN   Noun, singular or mass    36.WRB wh-adverb
13.NNS Noun, plural               37. #  Pound sign
14.NNP Proper noun, singular      38. $  Dollar sign
15.NNPS Proper noun, plural       39. .  Sentence-final punctuation
16.PDT Predeterminer              40. ,  Comma
17.POS Possessive ending          41. :  Colon, semi-colon
18.PRP Personal pronoun           42. (  Left bracket character
19.PP   Possessive pronoun        43. )  Right bracket character
20.RB   Adverb                    44. "  Straight double quote
21.RBR Adverb, comparative        45. `  Left open single quote
22.RBS Adverb, superlative        46. "  Left open double quote
23.RP   Particle                  47. '  Right close single quote
24.SYM Symbol                     48. "  Right close double quote
        (mathematical or scientific)
```

Affixation

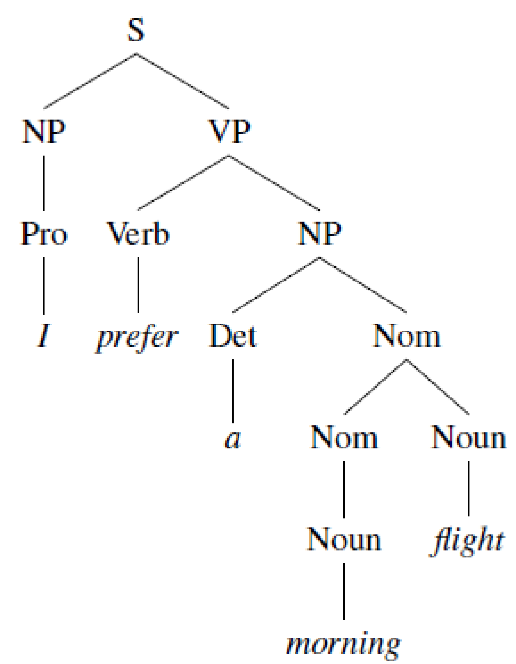
- Word stems (lemmas) + affixes (prefixes, suffixes)
  - May involve spelling changes, e.g. *able* -> *ability*
  - Can be productive, e.g. *unreprogramability*
- Inflectional suffixes – to create variants of the same POS as the stem:
  - *+s, +es* for plural nouns – e.g. *noun -> nouns*, *class -> classes*, *story -> stories*
  - *+s, +ed, +ing* for verbs in different tenses and aspects – e.g. *like -> lik**es***, *lik**ed***, *lik**ing***
- Derivational affixes - often change the inherent meaning of the word and/or its POS
  - Suffixes: e.g. *teach* (V) -> *teach**er*** (N), *produce* (V) -> *produc**tion*** (N)
  - Prefixes: e.g. *apply* -> ***re**apply*, *happy* -> ***un**happy*
- There are irregular forms and ambiguities
  - "*corpus*" vs. "*corpora*", "*seek*" vs. "*sought*"
  - Is "*bore*" the present tense of "*bore*" or past tense of "*bear*"?

Lexical Features

- Words are described with lexical features based on their syntactic categories and variant forms.
  - Number (num): sg/pl e.g. *word/words*
  - N-type: mass, count, name
  - Person (per): 1, 2, 3 e.g. *I*(1sg), *you*(2sg), *he*(3sg), *they*(3pl)
  - Case: nom, acc e.g. *he, him*
  - Valence: intransitive, transitive, ditransitive, scomp, etc. e.g. *smile, eat, give, believe*, …
  - A-type: base/comparative/superlative e.g*. old, older, oldest*

Syntax

- Words go together to form syntactic units of various kinds called constituents – words, phrases, clauses
- Parse trees represent the syntactic structure of sentences, showing the constituents.



Phrases

- Phrasal Categories (with the corresponding **head** word)
  - **NP** (noun phrases) – e.g. "*all the non-stop morning flights from Denver to Tampa leaving before 10*"
    - Head noun
    - Before head noun: determiners, cardinal/ordinal numbers, quantifiers, adjectives
    - Postmodifiers: prepositional phrases, non-finite clauses, relative clauses
  - **VP** (verb phrases) – e.g. *"book a flight that goes from Denver to Tampa"*
    - Head verb
    - Other constituents: NPs, PPs, Sentential Complements, VP
  - **AP** (adjectival phrases) – head adjective, may be preceded by adverbs. E.g. "*very early*"
  - **PP** (prepositional phrases) – a preposition followed by an NP, e.g. "from Denver"

Clauses

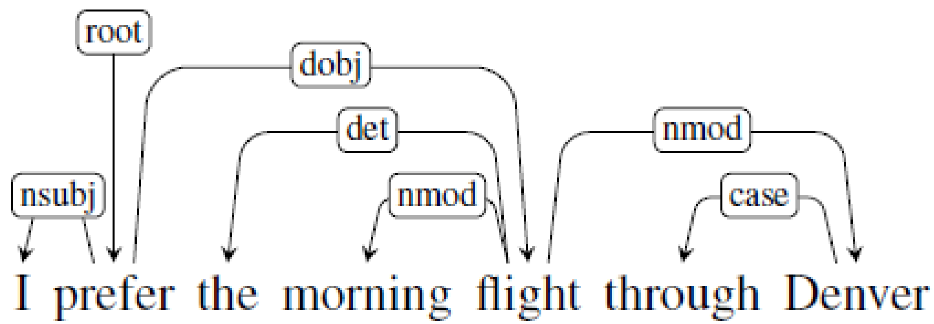
- Clausal Categories
  - **Declarative** clauses (e.g. *The taxi arrived early*.)
  - **Interrogative** clauses
    - yes-no questions (e.g. *Is he coming?*)
    - wh-questions (e.g. *When will the taxi arrive?*)
  - **Imperative** clauses (e.g. *Close the door.*)
  - **Relative** clauses (e.g. *Here's the taxi* ***that you called.***)
  - **Complement** clauses (e.g. *I know* ***that the taxi is here***.)
  - **Passive** clause (e.g. *The building was bought by a tycoon*.)

Grammatical Relations of the Constituents

- Subject – ***Alice*** *smiled*.
- Direct object – *Alice ate* ***a burger***.
- Indirect object – *Alice gave* ***me*** *a book*.
- Infinitive complement – *Alice wanted* ***to dance***.
- Specifier/modifier – *Alice is a* ***very clever*** *student. She studies* ***diligently***.

Dependency Relations

- Typed dependency structures, encoding important information in the sentences
- Illustrated as labelled arcs from heads to dependents



- Approximating the semantic relations between predicates and their arguments

Dependency Relations

- Selected dependency relations from the Universal Dependency set

Clausal Argument Relations	Description
NSUBJ	Nominal subject
DOBJ	Direct object
IOBJ	Indirect object
CCOMP	Clausal complement
XCOMP	Open clausal complement
Nominal Modifier Relations	**Description**
NMOD	Nominal modifier
AMOD	Adjectival modifier
NUMMOD	Numeric modifier
APPOS	Appositional modifier
DET	Determiner
CASE	Prepositions, postpositions and other case markers
Other Notable Relations	**Description**
CONJ	Conjunct
CC	Coordinating conjunction

Semantics

- Linguistic expressions -> meaning representation (knowledge representation, e.g. FOL, frames)
  - Propositions (predicates, referring expressions)
  - Correctness (true/false), contradiction
  - E.g. *I ate a turkey sandwich for lunch at my desk.*

$$\exists e\, Eating(e) \wedge Eater(e, Speaker) \wedge Eaten(e, TurkeySandwich) \wedge Meal(e, Lunch) \wedge Location(e, Desk)$$

- Ambiguity – some sentences can convey more than one proposition.
- Entailment – The assertion of some propositions implies the truth of other propositions. =>Inference!

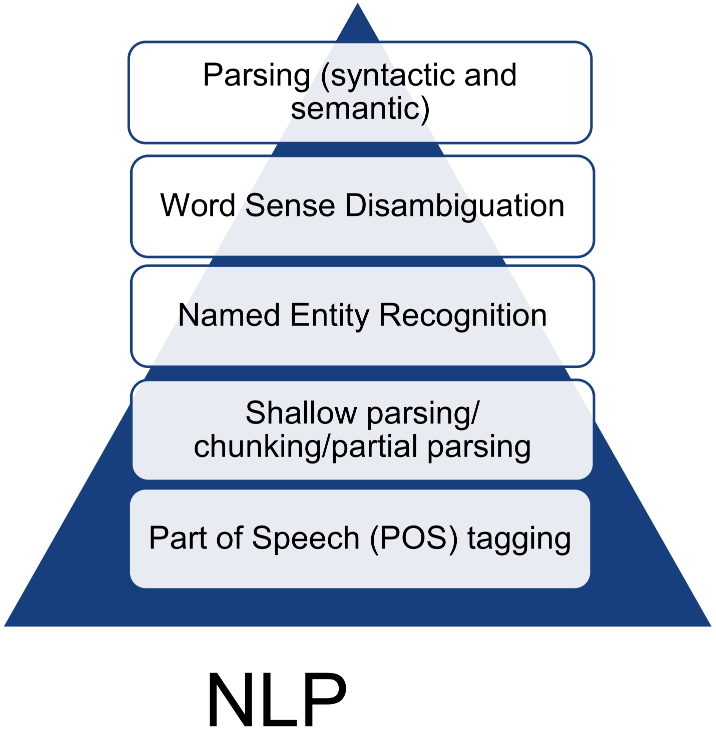
Pragmatics

- The use of language in context (both linguistic and situational)
- Utterances and speech acts (to achieve some effect on hearer)
  - Locution
    - Physical utterance with context and reference, i.e., who is the speaker and the hearer, which is the object, etc.
  - Illocution
    - The act of conveying intentions, i.e., the speaker wants the hearer to do something or to think something as a consequence of its utterance
  - Perlocutions
    - Actions that occur as a result of the illocution
  - Example: **Open the window!**
    - Locution: Monique is the speaker, Steve is the hearer, the window is the last left one
    - Illocution: Monique wants Steve to open the window
    - Perlocution: Steve opens the window
- Discourse, coherence

Natural Language Processing Tasks

Natural Language Processing Tasks

- To extract more sophisticated features, additional linguistic analyses of the text is needed.
- Input: the original text string



POS Tagging

- To determine POS or grammatical category of a term
  - Dictionary with word-POS correspondence is needed

IN/ About CD/ six CC/ and DT/ a JJ/ half NNS/ hours RB/ later ,/ , NNP/ Mr. NNP/ Armstrong VBD/ opened DT/ the NN/ landing NN/ craft POS/ 's NN/ hatch ,/ , VBD/ stepped RB/ slowly IN/ down DT/ the NN/ ladder CC/ and VBD/ declared IN/ as PRP/ he VBD/ planted DT/ the JJ/ first NN/ human NN/ footprint IN/ on DT/ the NN/ lunar NN/ crust :/ : ``/ " DT/ That VBZ/ 's CD/ one JJ/ small NN/ step IN/ for NN/ man ,/ , CD/ one JJ/ giant NN/ leap IN/ for NN/ mankind ./ . "/ "

Generated by UIUC POS Tagger

- POS disambiguation
  - 14-15% of words in the vocabulary, mostly common words, are ambiguous, hence 55-67% of word tokens in running text are ambiguous.
  - Baseline: choose the tag which is the most frequent in the training corpus
  - Using rule-based or stochastic approach

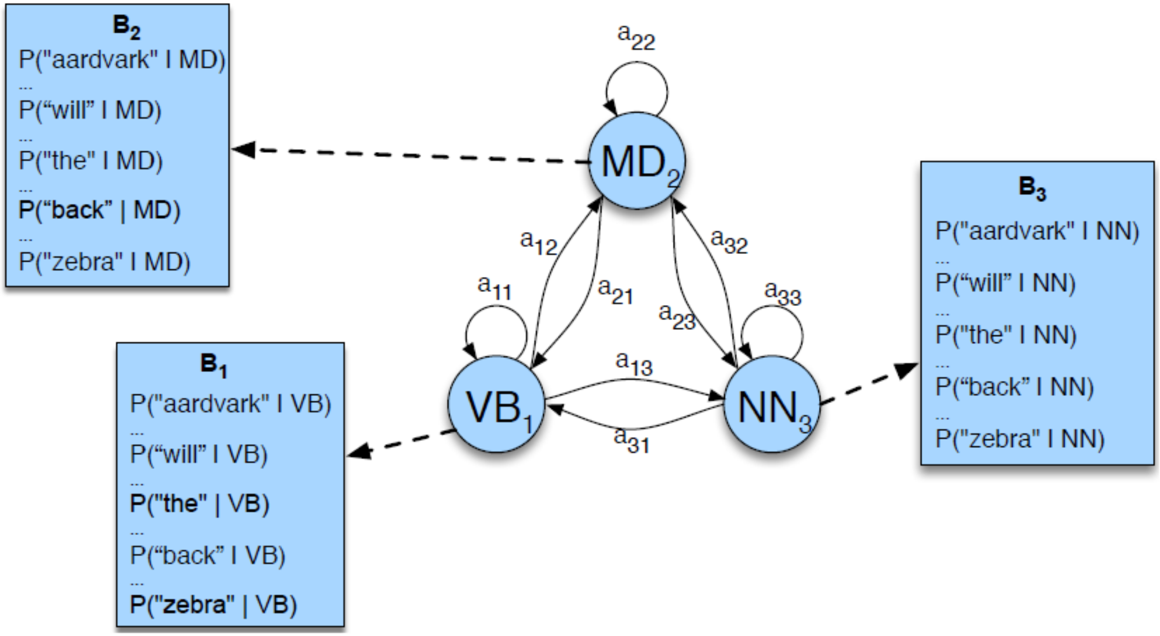
earnings growth took a **back/JJ** seat
a small building in the **back/NN**
a clear majority of senators **back/VBP** the bill
Dave began to **back/VB** toward the door
enable the country to buy **back/RP** about debt
I was twenty-one **back/RB** then

POS Taggers

- Rule-based - e.g. Brill’s tagger by Eric Brill
  - Error-driven transformation-based tagger
  - Initially assign the most frequent tag to each word, based on dictionary and morphological rules
  - Contextual rules are then applied repeatedly to correct any errors
- Stochastic taggers – e.g. CLAWS, Viterbi, Baum-Welch, etc.
  - based on Hidden Markov Models (HMMs) and n-gram probabilities
  - Manually tagged corpus is needed to estimate probabilities
- Many machine learning methods have also been applied. (Stanford’s Statistical NLP website lists many free taggers.)

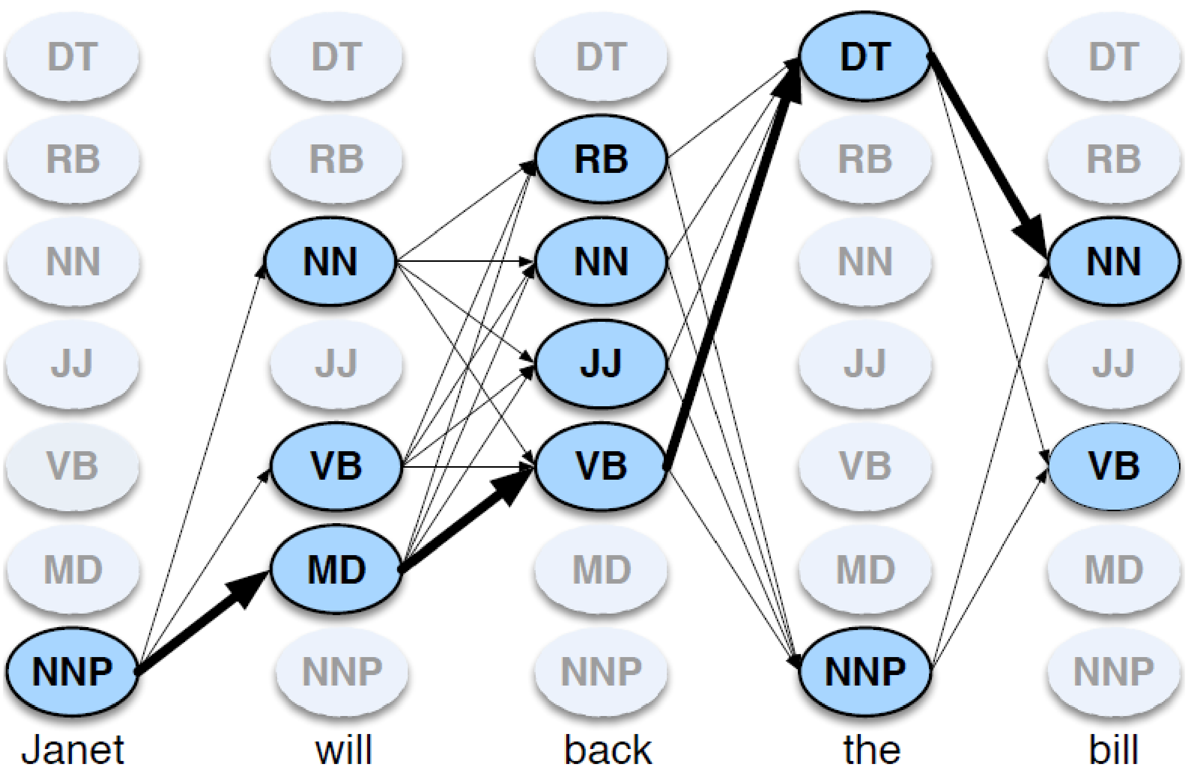
HMM

- HMM: a probabilistic sequence model/classifier, trained from tagged corpus.
  - A transition probabilities – the probability of a tag occurring given the previous tag.
  - B observation likelihoods – the probability that a given tag will be associated with a given word



HMM Decoding

- Given a sequence of words (observations), and an HMM, compute a probability distribution over possible sequences of labels (states) and chooses the best label sequence.



Shallow Parsing / Chunking

- To identify phrases in a text (noun phrases, verb phrases, and prepositional phrases, etc.)
- Largely stochastic techniques based on probabilities derived from an annotated corpus – segmenting and labeling
- Faster, more robust than full parsing

[NP About six and a half hours] [ADVP later] , [NP Mr. Armstrong] [VP opened] [NP the landing craft] [NP 's hatch] , [VP stepped] [ADVP slowly] [PP down] [NP the ladder] and [VP declared] [SBAR as] [NP he] [VP planted] [NP the first human footprint] [PP on] [NP the lunar crust] : "[NP That] [VP 's] [NP one small step] [PP for] [NP man] , [NP one giant leap] [PP for] [NP mankind] ."

Generated by UIUC chunker

Name Entity Recognition

- Recognition of particular types of proper noun phrases, specifically *persons*, *organizations*, *locations*, and sometimes *money*, *dates*, *times*, and *percentages*.
- Very useful in text analytics applications, by turning verbose text data into a more compact structural form
- **More details in another module**

[**LOC Houston**] , Monday, July 21 -- Men have landed and walked on the moon. Two [**MISC Americans**] , astronauts of [**ORG Apollo**] 11, steered their fragile four-legged lunar module safely and smoothly to the historic landing yesterday at 4:17:40 P.M., Eastern daylight time. [**PER Neil A. Armstrong**] , the 38-year-old civilian commander, radioed to earth and the mission control room here: "[**LOC Houston**] , [**ORG Tranquility Base**] here; the Eagle has landed."

Generated by UIUC NER system

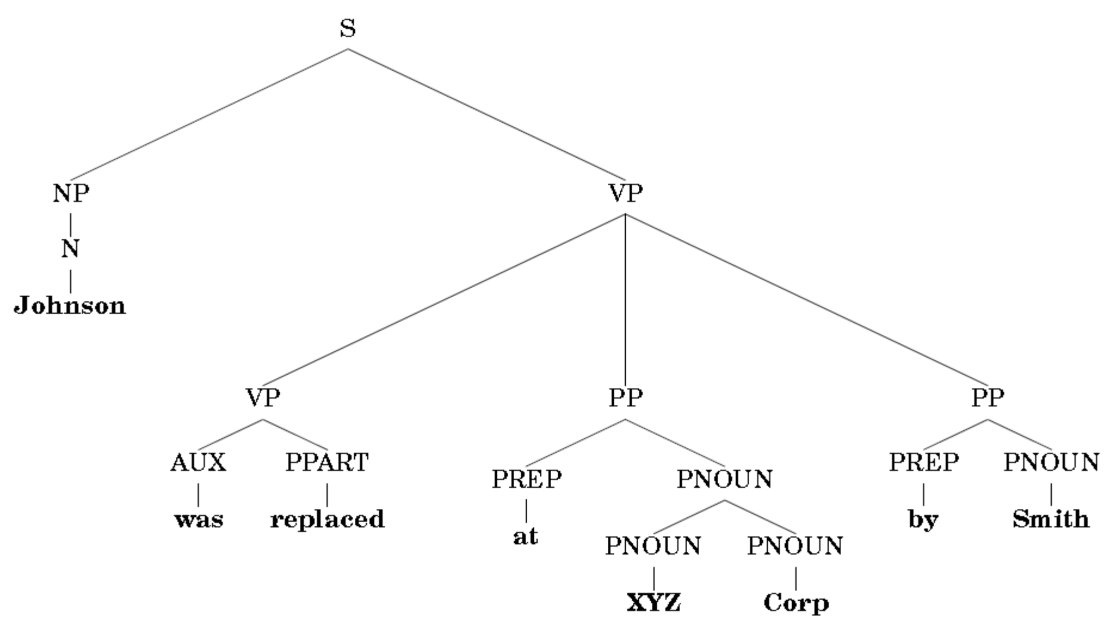
Word Sense Disambiguation

- Words are also ambiguous as to their meaning or reference (polysemous)
  - E.g. *table*: 1. a piece of furniture with a flat top supported by legs
    2. A list of numbers, facts, or information arranged in rows across and down a page
- Disambiguation of meanings in context has not been well solved, partly due to the lack of semantic concordances, corpora of disambiguated text to serve as training corpus for machine learning algorithms
  - E.g. You will find^{9}_{v} that avocado^{1}_{n} is^{1}_{v} unlike^{1}_{j} other^{1}_{j} fruit^{1}_{n} you have ever^{1}_{r} tasted^{2}_{v}
- Usually not applied in a typical text analytics application

Parsing

- Or *Syntactic Analysis*, the more sophisticated kind of text processing
- To produce a full parse of a sentence, typically as a tree, with syntactic functions of each word (e.g. subject, object, etc.)
- Comparatively expensive process, but can provide information that shallow parsing can not provide.

Johnson was replaced at XYZ Corp by Smith



Trees

- Trees can be represented in bracketed forms:

Tagging

```
John/NNP was/VBD replaced/VBN at/IN XYZ/NNP Corp/NNP by/IN Smith/NNP ./.
```

Parse

```
(ROOT
  (S
    (NP (NNP John))
    (VP (VBD was)
      (VP (VBN replaced)
        (PP (IN at)
          (NP (NNP XYZ) (NNP Corp)))
        (PP (IN by)
          (NP (NNP Smith)))))
    (. .)))
```

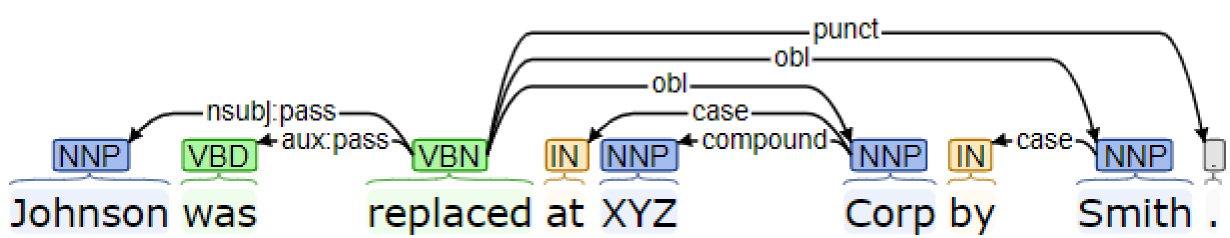
From Stanford Parser

Dependencies

- Typed dependencies

Typed dependencies, collapsed

```
nsubjpass(replaced-3, John-1)
auxpass(replaced-3, was-2)
root(ROOT-0, replaced-3)
nn(Corp-6, XYZ-5)
prep_at(replaced-3, Corp-6)
agent(replaced-3, Smith-8)
```



From Stanford Parser

From Syntax to Semantics

- Semantic analysis can be applied on top of parsing result to help identify the right entities for the text mining task.

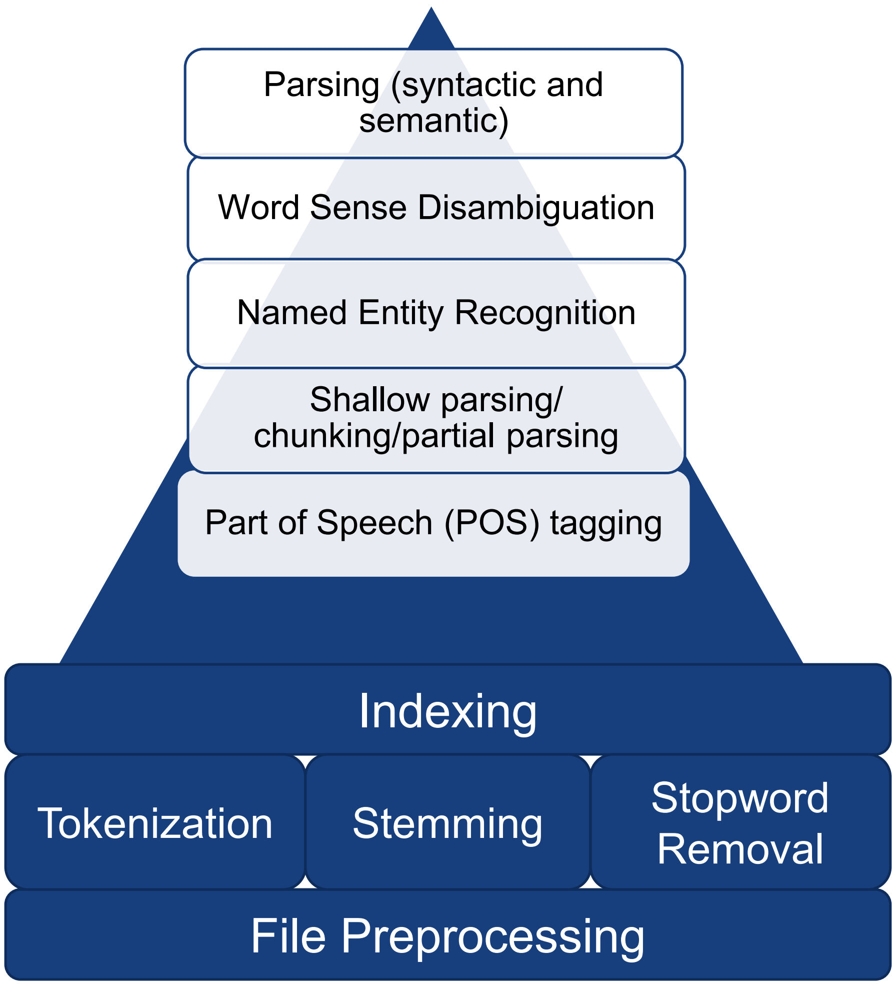
	SRL	Charniak
John	old thing [A1]	(S1 (S (NP (NNP John))
was		(VP (AUX was)
replaced	V: replace	(VP (VBN replaced)
at	location [AM-LOC]	(PP (IN at)
XYZ		(NP (NNP XYZ)
Corp		(NNP Corp))
by	replacer [A0]	(PP (IN by)
Smith		(NP (NNP Smith)))))
.		(. .)))

Generated by UIUC Semantic Role Labeling system

Challenges in Parsing

- Robustness – graceful degradation
  - The input may not conform to what is normally expected
  - Ill-formed input or lack of coverage of grammars
  - To recover as much meaningful information as possible
- Disambiguation
  - Ambiguity accumulated from earlier steps can result in combinatorial increase of possible parses
  - Return the *n* best analyses, if not one, to the next level of processing
- Efficiency
  - Theoretical time complexity of most formalisms are polynomial

Summary



Parsing (syntactic and semantic)
Word Sense Disambiguation
Named Entity Recognition
Shallow parsing/ chunking/partial parsing
Part of Speech (POS) tagging
Indexing
Tokenization
Stemming
Stopword Removal
File Preprocessing

Reference & Resources

- Jurafsky, Dan. *Speech & language processing*. Pearson Education India, 2000. (continuously updated)
- *Introduction to Linguistics for Natural Language Processing*, by Ted Brisco (https://www.cl.cam.ac.uk/teaching/1314/L100/introling.pdf)
- Marcus, Mitchell, Beatrice Santorini, and Mary Ann Marcinkiewicz. "Building a large annotated corpus of English: The Penn Treebank." (1993). (https://catalog.ldc.upenn.edu/docs/LDC95T7/cl93.html)
- UIUC POS Tagger, Chunker, etc.
  - http://cogcomp.cs.illinois.edu/page/demos
- NLP resources: http://nlp.stanford.edu/links/statnlp.html