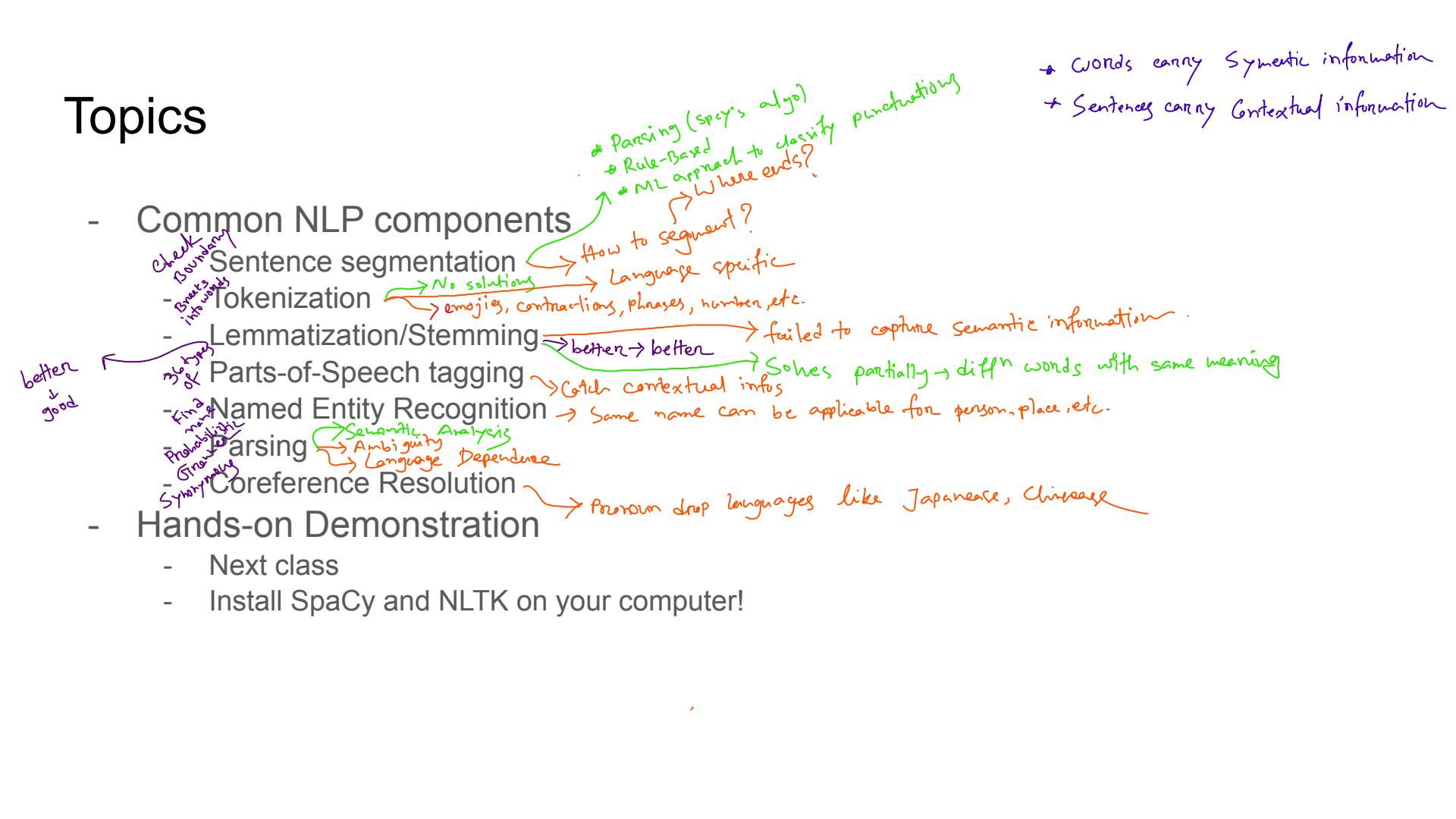
CSE440: Natural Language Processing II

Dr. Farig Sadeque
Associate Professor
Department of Computer Science and Engineering
BRAC University

Lecture 2: Linguistics Essentials

Topics

- Common NLP components
  - Sentence segmentation
  - Tokenization
  - Lemmatization/Stemming
  - Parts-of-Speech tagging
  - Named Entity Recognition
  - Parsing
  - Coreference Resolution
- Hands-on Demonstration
  - Next class
  - Install SpaCy and NLTK on your computer!



NLP Annotations

- Associating extra information to a piece of text

- Example:

  Dr. Jennifer Smith visited China. She liked it very much.

NLP Annotations

- Associating extra information to a piece of text

- Example:

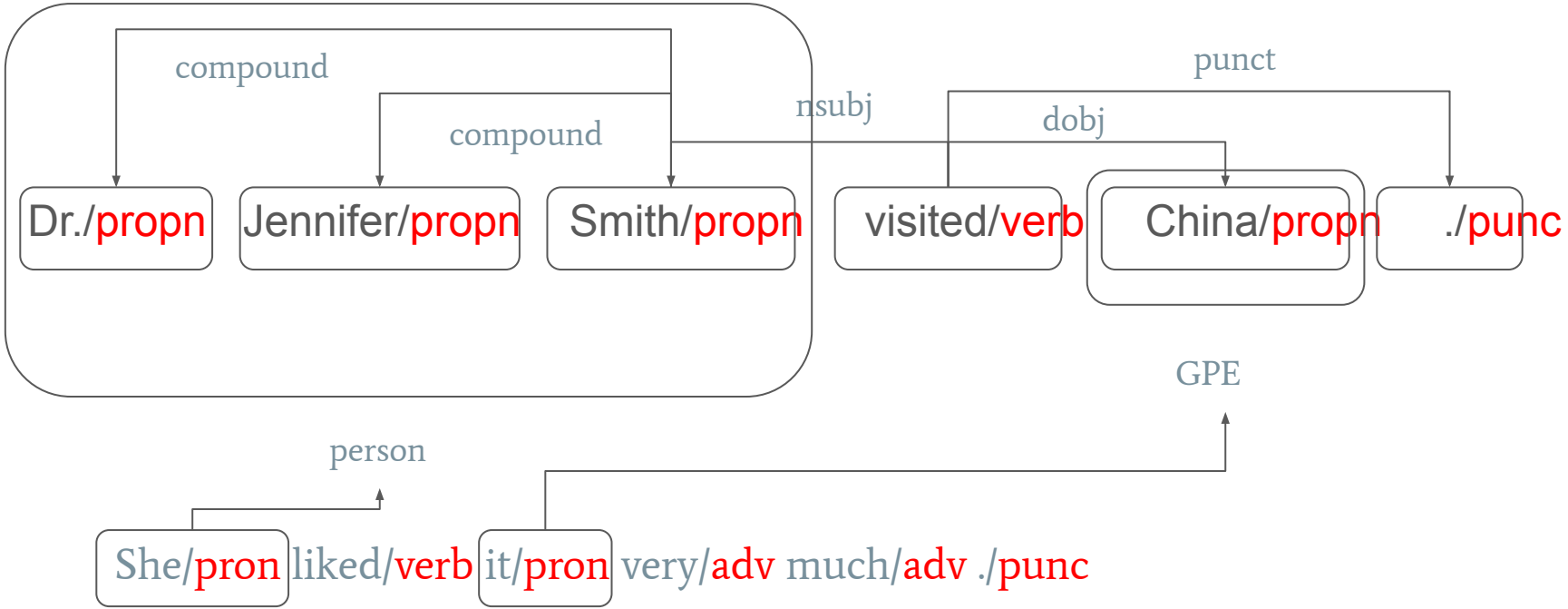
  Dr. Jennifer Smith visited China. She liked it very much.

  Dr./propn Jennifer/propn Smith/propn visited/verb China/propn ./punc

  She/pron liked/verb it/pron very/adv much/adv ./punc

  This is called Parts-of-Speech tagging.

More Annotations: Dependencies, Named Entities, Coreference



Common NLP Components

- Sentence segmentation
- Tokenization
- Lemmatization/Stemming
- Parts-of-Speech tagging
- Named Entity Recognition
- Parsing
- Coreference Resolution

Sentence Segmentation: Challenges

How do you know where an English sentence ends?

Sentence Segmentation: Challenges

How do you know where an English sentence ends?

- Consider: Mr. Smith lives in the U.S.A. He said “I am an American citizen!”

Sentence Segmentation: Challenges

How do you know where an English sentence ends?

- Consider: Mr. Smith lives in the U.S.A. He said “I am an American citizen!”

Many ‘.’ ‘!’ and ‘?’ end sentences but not all:

- Some ‘.’ are in abbreviations
- Some ‘.’ in abbreviations also end sentences
- Quotes after ‘.’ ‘!’ or ‘?’ are in the same sentence
- etc.

Sentence Segmentation: Solution

Rules:

- Easy to write a few rules
- Large rule sets are hard to maintain

Sentence Segmentation: Solution

Rules:

- Easy to write a few rules
- Large rule sets are hard to maintain

Machine learning:

- Classify each punctuation character: sentence final?
- Features: surrounding characters, words
- Around 99% accuracy

Sentence Segmentation: Solution

Rules:

- Easy to write a few rules
- Large rule sets are hard to maintain

Machine learning:

- Classify each punctuation character: sentence final?
- Features: surrounding characters, words
- Around 99% accuracy

Parsing (spacy's algorithm):

- Let the dependency parser figure it out

Tokenization: Brainstorming

Someone has told you that words in English can be separated by simply splitting on whitespace. How many times would that heuristic fail for the following text?

Mr. O’Neill said reaction to Sea Container’s proposal “hasn’t been very positive.” In New York Stock Exchange composite trading yesterday, Sea Containers closed at $62.625, down 62.5 cents.

What could you do to improve the heuristic?

Tokenization Challenges

- Words with punctuation: C++, C#, M*A*S*H, etc.
- Emoticons: =) :) ;-) etc.
- Contractions: I'll, isn't, dog's, etc.
  - Typically split to separate, e.g., noun (I) from verb ('ll)
- Hyphens in words: e-mail, co-operate, etc.
- Hyphens between morphemes: non-lawyer, pro-Arab
- Hyphens between words: once-quiet study,
- take-it-or-leave-it offer, 26-year-old, etc.
- Names: New York vs. York
- Phrasal verbs: make up, work out, etc.
- Phone numbers: +(880) 1756-111111

Tokenization Challenges

How about other languages?

- Chinese: 我正在教一堂課
  - Means “I am teaching a class.”
  - each character is a word, simpler characters build complex ones, and there is no space!
- German: Lebensversicherungsgesellschaftsangestellter (pronounce this!)
  - Means “life insurance company employee”

Tokenization Challenges

How about other languages?

- Chinese: 我正在教一堂課
  - Means “I am teaching a class.”
  - each character is a word, simpler characters build complex ones, and there is no space!
- German: Lebensversicherungsgesellschaftsangestellter (pronounce this!)
  - Means “life insurance company employee”
- How about Bangla?
  - Let me present you the one and only Michael Madhusudan Dutta

    নিকুম্ভিলা যজ্ঞ সাংগ করি, আরম্ভিলে/যুদ্ধ দম্ভি মেঘনাদ, বিষম সঙ্কটে/ঠেকিবে বৈদেহীনাথ, কহিনু তোমারে

Tokenization Solutions

Unfortunately, no general solution. Each language requires it’s own tokenization principles.

How does common tools do it then?

Spacy: recursively split on whitespace, known exceptions, affixes, and punctuation

Problem: Similar Words Look Different

The words *dog* and *dogs* are closely related, but on a computer "dog" != "dogs"

Solutions:

- Cut out common substrings (stemming/lemmatization)
- Replace words with vectors (embeddings)

Stemming and Lemmatization

Stemming:

- Rules strip pieces of words (not morphemes)
- E.g, Porter stemmer: *equivalence* → *equival*
- Fast, but inaccurate, e.g., *organization* → *organ*, *European* !→ *Europe*

Stemming and Lemmatization

Stemming:

- Rules strip pieces of words (not morphemes)
- E.g, Porter stemmer: *equivalence* → *equival*
- Fast, but inaccurate, e.g., *organization* → *organ*, *European* !→ *Europe*

Lemmatization

- Hand-built lexicon for all word forms, walked → walk
- Accurate, but slower, and there is a chicken-egg scenario with parts of speech tagging

Embedding

If dog = [0.5; 0.4; 0.1] and dogs = [0.5; 0.4; 0.2]
then cos dog; dogs = 0:99

Goal: learn an embedding vector for each word such that similar words have similar vectors.

Will be covered in session 3.

Before moving on

Our next overview is going to be on Parts of Speech tagging. Before starting that, please review these two links:

- Penn TreeBank tags:
  - https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html
  - Section 2 and 3
- Universal POS tags:
  - https://universaldependencies.org/u/pos/all.html

NLP Libraries

	SpaCy	NLTK	CoreNLP	Processors
Fast	yes	no	yes	yes
State-of-the-art	yes	no	yes	yes
Large community	yes	yes	yes	no
Simple APIs	yes	yes	no	no
Language	Python	Python	Java	Scala

Before next class, please install SpaCy and NLTK on your computer

Parts-of-Speech (POS) Tagging

Assigning grammatical categories for words

She/pron liked/verb it/pron very/adv much/adv ./punc

closed class

- categories have a fixed set of words
- prepositions, determiners, pronouns, conjunctions, auxiliary verbs, particles, numerals

open class

- categories have a growing set of words
- nouns, verbs, adjectives, adverbs

POS tagging

noun “person, place, or thing”: farmer, Dhaka, dice but also explosion, moment

verb “action or process”: grab, evolve, rain

adjective “property or quality”, modify nouns: green, old

adverb “modify verbs and adjectives”: slowly, very, today

adposition “before/after a noun phrase”: over, before

determiner “express reference of noun”: a, the, that

pronoun “substitute for noun”: you, our, who

conjunction “join two phrases”: and, but, if

particle “associated with other word”: not, maybe rule out

interjection “exclamation”: psst, ouch, hello

POS Tagging Challenges

One word can have different POS tag based on its use

- I painted the room vs. the painted room
- Is *painted* a verb or an adjective?

Annotate the following sentence with POS tags from Penn TreeBank tags:

“Wow! That first post really blew up.”

Named Entity Recognition (NER)

Identify phrases that are named people, locations, organizations, etc.

punct

Dr./propn	Jennifer/propn	Smith/propn	visited/verb	China/propn	./punc
	Person			GPE	

Common named entity types:

- **person** Turing is often considered the. . .
- **organization** The IPCC said it is likely that. . .
- **location** The Mt. Sanitas loop hike. . .
- **geo-political entity** Palo Alto will raise parking fees.
- etc.

NER Challenges

Ambiguity:

- **Washington** was born into slavery.
- **Washington** went up 2 games to 1.
- Blair arrived in **Washington** today.
- **Washington** passed a primary seatbelt law.

NER Challenges

Ambiguity:

- **Washington** was born into slavery. <per>
- **Washington** went up 2 games to 1. <org>
- Blair arrived in **Washington** today. <loc>
- **Washington** passed a primary seatbelt law. <gpe>

Solution

Sequence Tagging

- Will study it in session 4

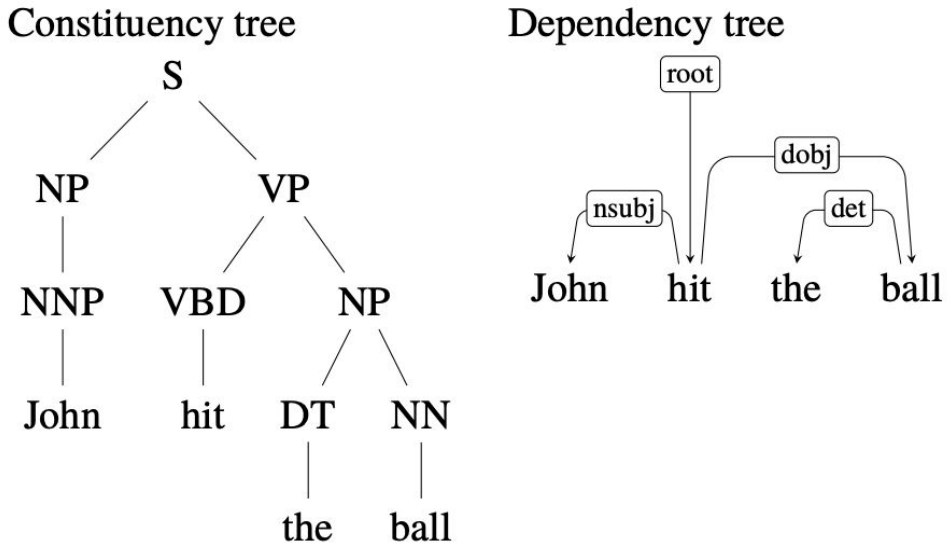
Simple scheme: label each word as (I)nside or (O)utside

More elaborate schemes:

- BIO: begin, inside, outside
- BILOU: begin, inside, last, outside, unit-length

Parsing and Syntactic Representation

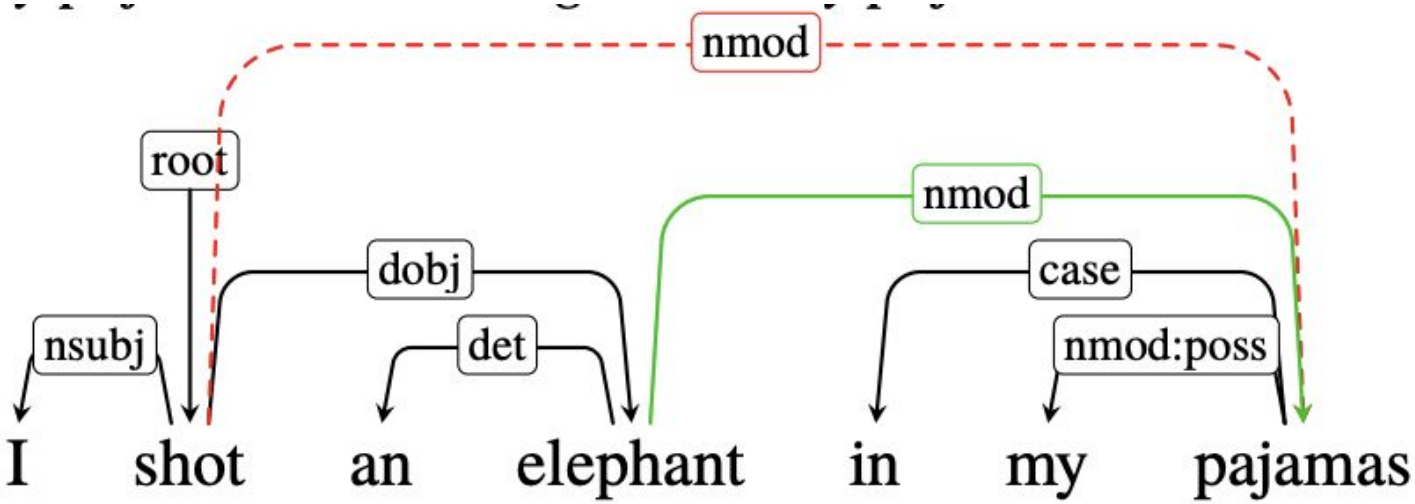
Example: John hit the ball.



Parsing Challenges

Attachment ambiguity: One morning I shot an elephant in my pajamas.

- Who was in my pajamas? Me? The elephant?



Coordination Ambiguity

Old men and women

- Old (men and women)?
- Old (men) and women?

Which one is correct?

Parsing solutions

- Probabilistic grammar based parsing
- Transition based parsing

We will learn theories of parsing in session 5

Demo

We will now check out some of the tools that are available to us.

- SpaCy
- NLTK
- **Stanford CoreNLP**