CSCE 771: Computer Processing of Natural Language
Lecture 5: Representation (Paper), Parsing, Projects

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

1ST SEPTEMBER, 2022

Carolinian Creed: "I will practice personal and academic integrity."

Acknowledgement: Used materials by Profs. Mausam, Jurafsky & Martin, Robert C. Berwick, Graham Neubig

Organization of Lecture 5

- Opening Segment
  - Announcements

- Main Lecture

Main Section
- Paper discussion – Word Representation
- Parsing - introduction

- Concluding Segment
  - Course Project – review topics
  - About Next Lecture – Lecture 6

Recap of Lecture 4

- We looked at a variety of NLP basic tasks
  - Tokenization – getting tokens for processing
  - Normalization - making into canonical form
  - Case folding – handling cases
  - Lemmatization – handling variants (shallow)
  - Stemming – handling variants (deep)

- NLP for business – sentiments for market intelligence

Main Lecture

Paper Discussion

Contextual Word Representations: Putting Words into Computers”,

by Noah Smith, CACM June 2020

Problem

- How to represent words ?
- How to measure similarity, e.g., between words, and texts?
- How to determine different contexts (senses) in which words are used?
- How to handle noise, typos?

S1 - This is an apple
S2 - These are apples

S3 - This is an apples
S4 - There are apply

Option 1 - Characters

- How to represent words?
  - Characters / Unicode / …
- How to measure similarity between words, and texts?
  - Edit distance
  - Hamming distance
- How to determine different contexts (senses) in which words are used?
  - Neighborhood of words: Bi-, tri-, N-gram representations

Option 2 - Vectors

- How to represent words? Vectors
  - But, what scheme in vectors
    - One-hot encoding
    - Arbitrary, principled, …

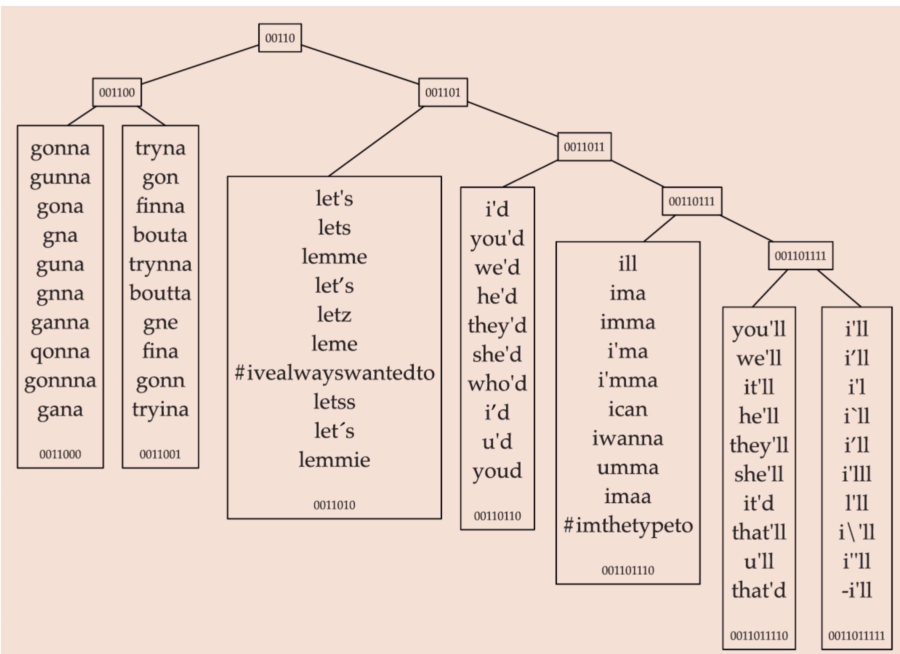
- How to measure similarity between words, and texts?
  - Cosine similarity

- How to determine different contexts in which words are used?
  - Neighborhood of words: Bi-, tri-, N-gram representations
  - Contextual word vectors

Contextual Word Embeddings

- Words as discrete
- Words with distributional assumptions:
  - Context: given a word, its nearby words or sequences of words
  - ***Words used in similar ways are likely to have related meanings***; i.e., words used in the same (similar) context have related meanings
    - No claim about meaning except relative similarity v/s dis-similarity of words

Contextual Representation by Clustering



Main steps

- Cluster words by context (i.e., neighborhood of the word)
- Compare with words in a manually-created taxonomy, e.g., Wordnet

The 10 most frequent words in clusters in the section of the hierarchy with prefix bit string 00110.

Owoputi, O., O’Connor, B., Dyer, C., Gimpel, K., Schneider, N., and Smith, N.A. Improved part-ofspeech tagging for online conversational text with word clusters. In Proceedings of 2013 NAACL.

Credit:
Contextual Word Representations: Putting Words into Computers”, by Noah Smith, CACM June 2020

Contextual Representation by Dimensionality Reduction

- Creating word vectors in which each dimension corresponds to the frequency the word type occurred in some context (here, two words on either side of *astronomers, bodies, objects*)
- Strategy 1: select contexts
  - Examples
    - Words in the neighborhood
    - Words of specific types
  - Build vectors
  - Use vector operations to derive meaning

context words	**v**(astronomers)	**v**(bodies)	**v**(objects)
't			1
,		2	1
.	1		1
1			1
And			1
Belt			1
But	1		
Given			1
Kuiper			1
So	1		
and		1	
are		2	1
between			1
beyond		1	
can			1
contains		1	
from	1		
hypothetical			1
ice		1	
including		1	
is	1		
larger		1	
now	1		
of	1		

$$\text{cosine\_similarity}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \cdot \|\mathbf{v}\|}$$

	astronomers	bodies	objects
astronomers	$\frac{14}{\sqrt{14} \cdot \sqrt{14}} = 1$	$\frac{0}{\sqrt{24} \cdot \sqrt{14}} = 0$	$\frac{1+1}{\sqrt{14} \cdot \sqrt{16}} \approx 0.134$
bodies		$\frac{24}{\sqrt{24} \cdot \sqrt{24}} = 1$	$\frac{2+2+2}{\sqrt{24} \cdot \sqrt{16}} \approx 0.306$
objects			$\frac{16}{\sqrt{16} \cdot \sqrt{16}} = 1$

Bodies and **objects** are **most similar** (0.306) than

- **Bodies** and **astronomers** (0)
- **Objects** and **astronomers** (0.134)

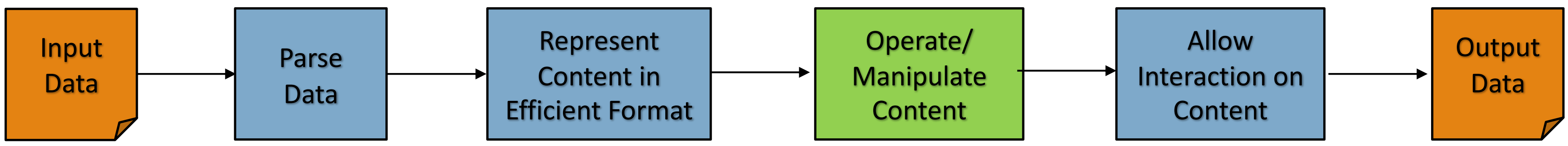
Credit:
Contextual Word Representations: Putting Words into Computers", by Noah Smith, CACM June 2020

Code

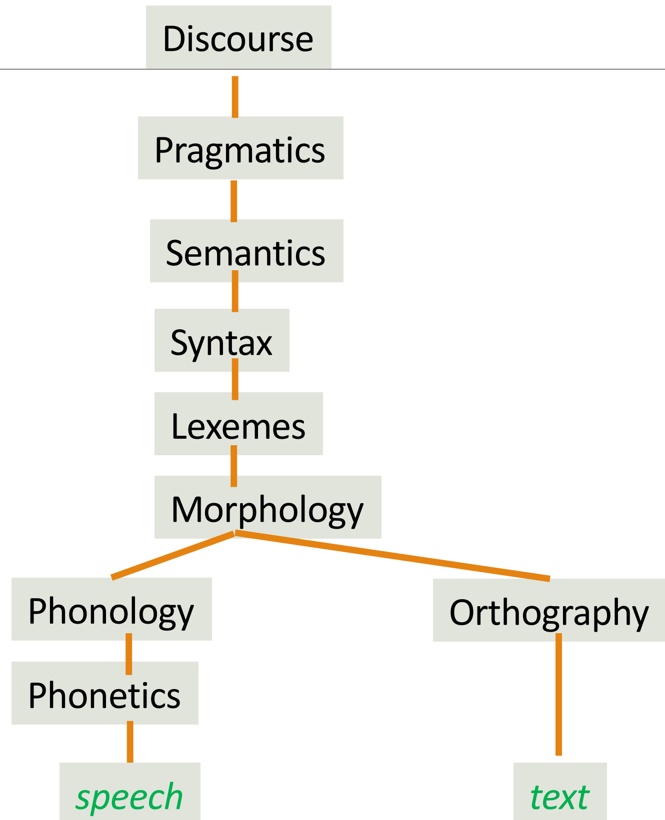
Notebook:

https://github.com/biplav-s/course-nl-f22/blob/main/sample-code/l5-wordrepresent/Word%20Representations%20-%20Vectors.ipynb

Parsing


Input Data
Parse Data
Represent Content in Efficient Format
Operate/ Manipulate Content
Allow Interaction on Content
Output Data

Levels of Linguistic Studies



- **Discourse:** study of group of sentences
- **Pragmatics:** how context contributes to meaning of sentences
- **Semantics:** meaning of words and combinations of words
- **Syntax:** rules for combining and using words/ phonemes.
- **Lexemes:** a set of words that are related through inflection (fly: verb, fly: noun)
- **Morphology**—rules that govern morphemes - the minimal meaningful units of language (lemmas and affixes)
- **Orthography**: convention for writing a language. E.g., spelling
- **Phonology:** organization of speech sound (i.e., phoneme)
- **Phonetics**: study of how sound is made and received

Why Parsing

- Recognizing legal inputs from illegal
- Usage of parse representation - parse tree
  - Grammar checking
  - Semantic analysis
  - Machine translation
  - Question answering
  - Information extraction
  - Speech recognition
  - …

Adapted from material by Robert C. Berwick

Background: Context Free Grammar (CFG)

N a set of **non-terminal symbols** (or **variables**)

Σ a set of **terminal symbols** (disjoint from N)

R a set of **rules** or productions, each of the form $A \rightarrow \beta$,
where A is a non-terminal,
β is a string of symbols from the infinite set of strings $(\Sigma \cup N)*$

S a designated **start symbol** and a member of N

From Jurafsky & Martin

Simple Example Using CFGs

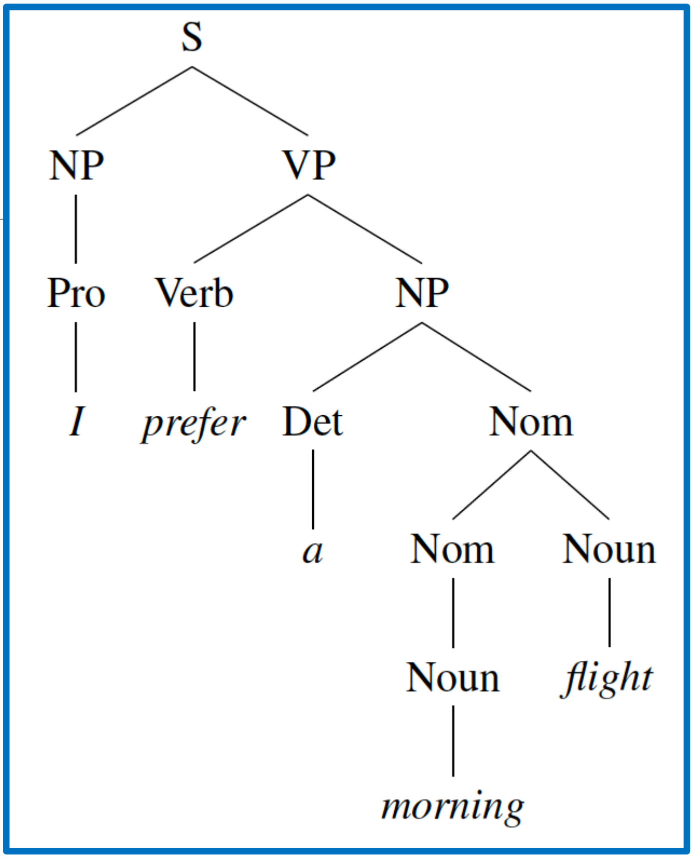
Noun	$\rightarrow$	*flights* | *breeze* | *trip* | *morning*
Verb	$\rightarrow$	*is* | *prefer* | *like* | *need* | *want* | *fly*
Adjective	$\rightarrow$	*cheapest* | *non-stop* | *first* | *latest*
		| *other* | *direct*
Pronoun	$\rightarrow$	*me* | *I* | *you* | *it*
Proper-Noun	$\rightarrow$	*Alaska* | *Baltimore* | *Los Angeles*
		| *Chicago* | *United* | *American*
Determiner	$\rightarrow$	*the* | *a* | *an* | *this* | *these* | *that*
Preposition	$\rightarrow$	*from* | *to* | *on* | *near*
Conjunction	$\rightarrow$	*and* | *or* | *but*

Grammar Rules			Examples
S	$\rightarrow$	*NP VP*	I + want a morning flight
NP	$\rightarrow$	*Pronoun*	I
	|	*Proper-Noun*	Los Angeles
	|	*Det Nominal*	a + flight
Nominal	$\rightarrow$	*Nominal Noun*	morning + flight
	|	*Noun*	flights
VP	$\rightarrow$	*Verb*	do
	|	*Verb NP*	want + a flight
	|	*Verb NP PP*	leave + Boston + in the morning
	|	*Verb PP*	leaving + on Thursday
PP	$\rightarrow$	*Preposition NP*	from + Los Angeles

From Jurafsky & Martin

An Example Using CFGs

Grammar Rules			Examples
S	→	*NP VP*	I + want a morning flight
NP	→	*Pronoun*	I
	|	*Proper-Noun*	Los Angeles
	|	*Det Nominal*	a + flight
Nominal	→	*Nominal Noun*	morning + flight
	|	*Noun*	flights
VP	→	*Verb*	do
	|	*Verb NP*	want + a flight
	|	*Verb NP PP*	leave + Boston + in the morning
	|	*Verb PP*	leaving + on Thursday
PP	→	*Preposition NP*	from + Los Angeles



From Jurafsky & Martin

$[_S\ [_{NP}\ [_{Pro}\ \text{I}]]\ [_{VP}\ [_V\ \text{prefer}]\ [_{NP}\ [_{Det}\ \text{a}]\ [_{Nom}\ [_N\ \text{morning}]\ [_{Nom}\ [_N\ \text{flight}]]]]]]$

Bracketed Notation

Example: Larger English CFG

Grammar
$S \rightarrow NP\ VP\ .$
$S \rightarrow NP\ VP$
$S \rightarrow$ “ S ” , $NP\ VP\ .$
$S \rightarrow$ -NONE-
$NP \rightarrow DT\ NN$
$NP \rightarrow DT\ NNS$
$NP \rightarrow NN\ CC\ NN$
$NP \rightarrow CD\ RB$
$NP \rightarrow DT\ JJ\ ,\ JJ\ NN$
$NP \rightarrow PRP$
$NP \rightarrow$ -NONE-
$VP \rightarrow MD\ VP$
$VP \rightarrow VBD\ ADJP$
$VP \rightarrow VBD\ S$
$VP \rightarrow VBN\ PP$
$VP \rightarrow VB\ S$
$VP \rightarrow VB\ SBAR$
$VP \rightarrow VBP\ VP$
$VP \rightarrow VBN\ PP$
$VP \rightarrow TO\ VP$
$SBAR \rightarrow IN\ S$
$ADJP \rightarrow JJ\ PP$
$PP \rightarrow IN\ NP$

Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential there
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	to
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

Table Source:
https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

Interpretation of Parsing Rules

- generation (production): $S \rightarrow NP\ VP$
- parsing (comprehension): $S \leftarrow NP\ VP$
- verification (checking): $S = NP\ VP$
- CFGs are declarative – tell us *what* the well-formed structures & strings are
- Parsers are procedural – tell us *how* to compute the structure(s) for a given string

From Robert C. Berwick

Types of Parsing

- **Phrase structure** / **Constituency Parsing**: find phrases and their recursive structure.
Constituency - groups of words behaving as single units, or constituents.

  - **Shallow Parsing/ Chunking**: identify the flat, non-overlapping segments of a sentence: noun phrases, verb phrases, adjective phrases, and prepositional phrases.

- **Dependency Parsing**: find relations in sentences

- **Probabilistic Parsing**: given a sentence X, predict the most probable parse tree Y

Lecture 5: Concluding Comments

- We looked at parsing and roles it plays: verifying , generating, recognizing
- Many types of parsing
- Shallow parsing for quick NLP tasks
- Phrase structure parsing
- Dependency parsing

Concluding Segment

Choosing a Project – Some Considerations

- Scope: what is the problem?
- Current-state: what happens in the problem today?
- Who cares: who will benefit with the problem being solved?
- Desired-state: what will be the future situation if your project succeeds?
- Resources/ dataset: do you have reasonable data and compute resources to do the work?
- Evaluation: how will we measure goodness of the work?

Review project spreadsheet

Discussion: Course Project

- Expectations
  - Apply methods learned in class or of interest to a problem of interest
  - Be goal oriented: aim to finish, be proactive, be innovative
  - Do top-class work: code, writeup, presentation
- Typical pitfalls
  - Not detailing out the project, assuming data
  - Not spending enough time
- What will be awarded
  - Results and efforts (balance)
  - Challenge level of problem

Course Project – Deadlines and Penalty Rubric

- Project plan **not** ready by Sep 15, 2020 **[-20%]**
    * Project Title
    * Description: motivation and expected output
    * Illustrative Test cases: i.e., Example input / output
    * Data sources:
    * Technique and tools to use:
    * Metric for measuring output
    * How will you collect results
    * Format of report, presentation
    * Time schedule:
- Project report **not** ready by Nov 10, 2022 **[-20%]**
- Project presentations **not** ready by Nov 15, 2022 **[-10%]**

About Next Lecture – Lecture 6

Lecture 6:

- Shallow/ Deep parsing
- Statistical Parsing