CSCE 771: Computer Processing of Natural Language
Lecture 8: (NLP) Evaluation, Semantics

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

12TH SEPTEMBER, 2024

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

Acknowledgement: Used materials by Jurafsky & Martin,

Organization of Lecture 8

- Opening Segment
  - Announcements
- Main Lecture
- Concluding Segment
  - About Next Lecture – Lecture 9

Main Section

- Review quiz
- Review last class/ parsing
- Introduce evaluation metrics in NLP context
- Semantics
- Review projects

Announcements

GUEST LECTURES ON LANGUAGE MODELS

Sep 24 (Tu)	Language Model – PyTorch, BERT, {Resume data, two tasks} **– Guest Lecture**
Sep 26 (Th)	Language Model – Finetuning, Mamba - **Guest Lecture**
Oct 1 (Tu)	Language model – comparing arch, finetuning - **Guest Lecture**
Oct 3 (Th)	Language model – comparison of results, discussion, ongoing trends– **Guest Lecture**

QUIZ

- Submit using Black board
- Includes resume exercise
- Due by next Monday, Sep 16, 2024

Recap of Lecture 7

- We discussed statistical parsing - Probabilistic grammars
  - assign a probability to a sentence or string of words
  - In a probabilistic context-free grammar (PCFG), every rule is annotated with the probability of that rule being chosen assuming conditional independence.
  - The probability of a sentence is computed by multiplying the probabilities of each rule in the parse of the sentence.
  - We looked at Stanford parser
- We had Quiz 1

Review of Quiz 1

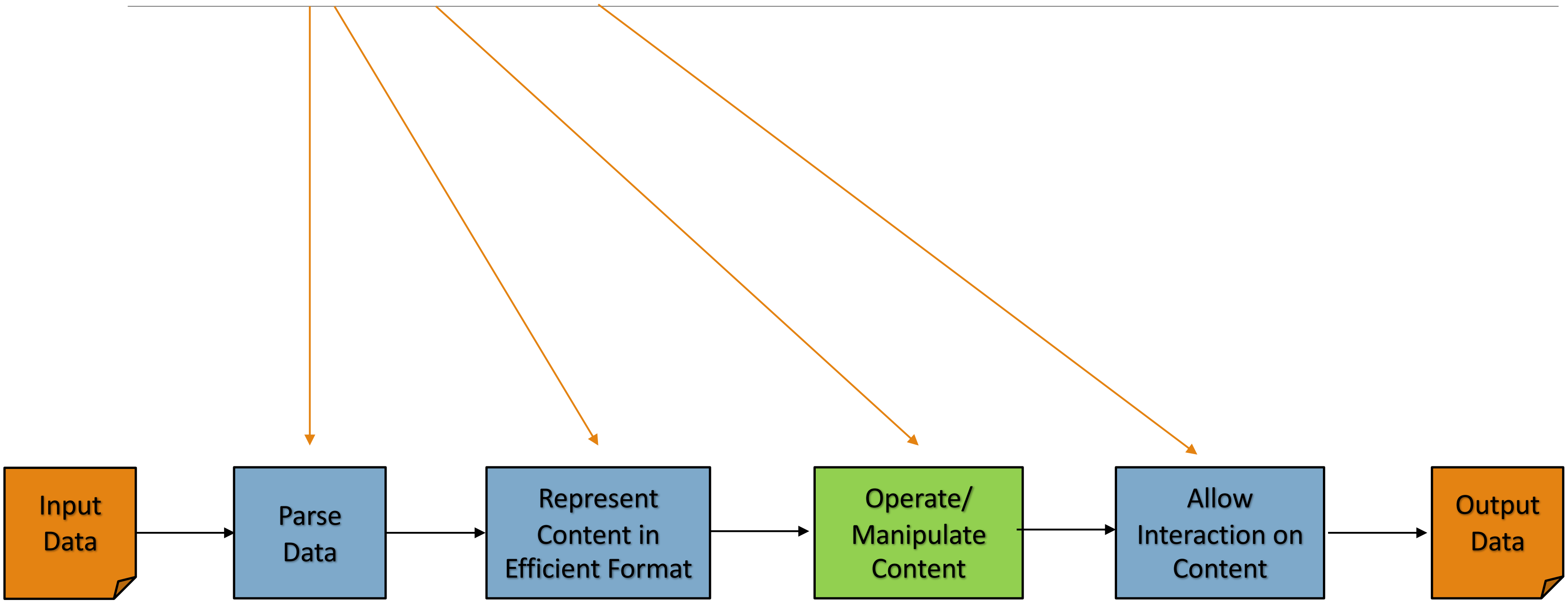
- Submit using Black board
- Includes resume exercise
- Due by next Monday, Sep 16, 2024

Bonus question for Quiz 1 [10 points] - as discussed in class.

Represent resumes as vectors. Now show how the following will be done:
Select a query **q** (document/resume) and then identify the document most similar to it from a corpus **C** containing **d** resumes, using **TF-IDF scores**?

Main Lecture

NLP Evaluation



Metric Types

- **Effectiveness**: what the **user** of a system sees, primarily cares about

Extrinsic evaluation (esp. downstream applications)

- **Efficiency**: what the **executor** in a system sees, primarily cares about

Intrinsic evaluation

Effectiveness ***Metrics***

Efficiency ***Metrics***

Example: Detecting Spam in Email

- **Effectiveness**: what the **user** of a system sees, primarily cares about
  - *How many spams identified?*
  - *How many spams missed?*
- **Efficiency**: what the **executor** in a system sees, primarily cares about
  - *How fast were spams detected?*
  - *How much memory was used per million emails processed ?*

Metrics: Accuracy, Precision, Recall

	Predicted class		
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

Accuracy =
(TP+TN)/
(TP+FP+FN+TN)

Precision =
(TP)/
(TP+FP)

Recall =
(TP)/
(TP+FN)

F1 Score: *Harmonic Mean*

1/F1 = 1/Precision + 1/Recall

F1 = 2*(Recall * Precision) /
(Recall + Precision)

Evaluating Parsers - PARSEVAL

Degree to which the constituents in the hypothesis parse tree look like the constituents in a hand-labeled, gold-reference parse like PENN TreeBank

Overall measure is by F1 score

$$F_1 = \frac{2PR}{P+R}$$

$$\textbf{labeled recall:} = \frac{\text{\# of correct constituents in hypothesis parse of } s}{\text{\# of correct constituents in reference parse of } s}$$

$$\textbf{labeled precision:} = \frac{\text{\# of correct constituents in hypothesis parse of } s}{\text{\# of total constituents in hypothesis parse of } s}$$

From Jurafsky & Martin

Average Performance With Multiple Classes

- Setting
  - Class A: 1 TP and 1 FP
  - Class B: 10 TP and 90 FP
  - Class C: 1 TP and 1 FP
  - Class D: 1 TP and 1 FP

Precision = (TP)/ (TP+FP)

- Average precision = ?

- Macro and micro average
  - A macro-average will compute the metric independently for each class and then take the average (hence treating all classes equally)
  - A micro-average will aggregate the contributions of all classes to compute the average metric.

A macro-average will then compute: $Pr = \frac{0.5+0.1+0.5+0.5}{4} = 0.4$

A micro-average will compute: $Pr = \frac{1+10+1+1}{2+100+2+2} = 0.123$

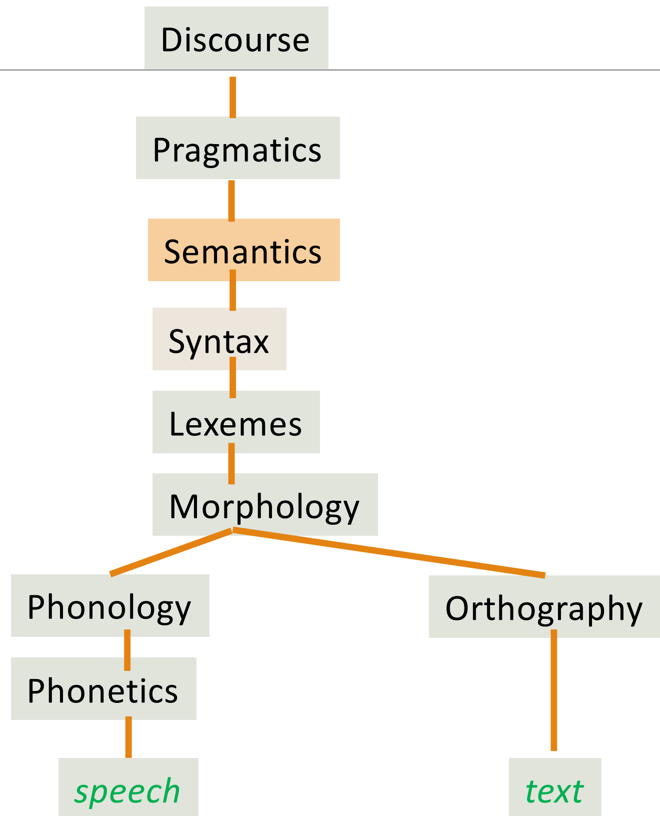
Source and credit: https://datascience.stackexchange.com/questions/15989/micro-average-vs-macro-average-performance-in-a-multiclass-classification-settin

Code Sample – Metrics Calculation

Notebook:

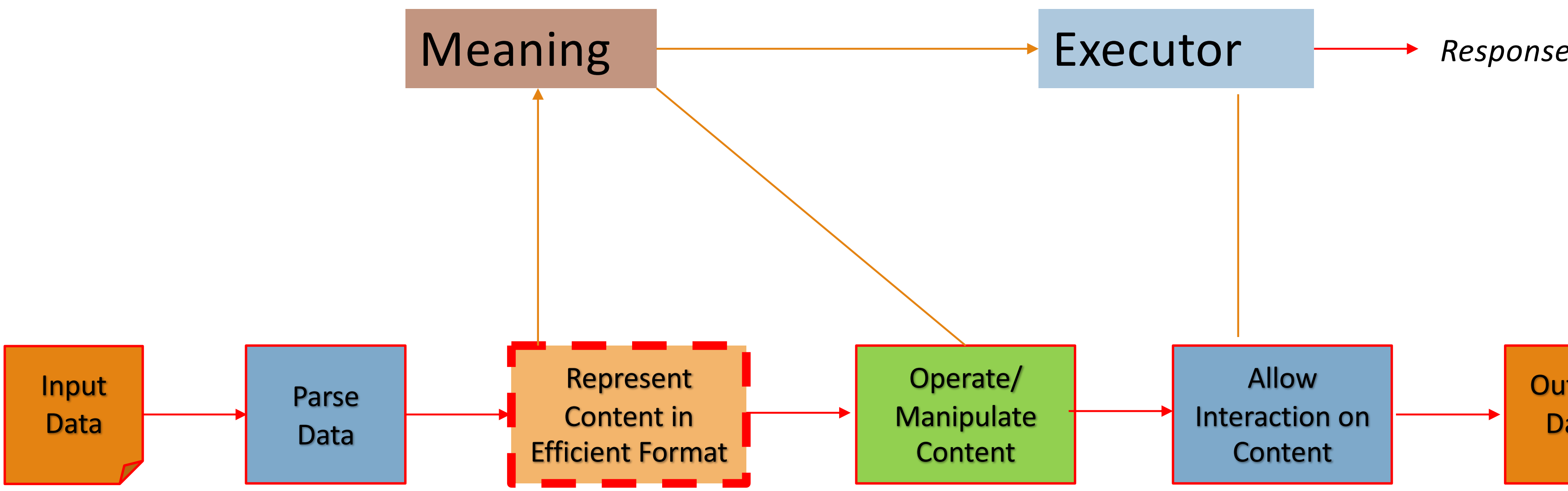
https://github.com/biplav-s/course-nl-f22/blob/main/sample-code/l8-review-evalmetrics/Metric%20Calculations.ipynb

Semantics



- **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

Semantics, Parsing and Representation



Semantics

- ***lexical semantics**:* studies word meanings and word relations, and
- ***formal semantics**:* studies the logical aspects of meaning, such as sense, reference, implication, and logical form
- ***conceptual semantics**:* studies the cognitive structure of meaning

Source: Jurafsky & Martin,
Wikipedia (https://en.wikipedia.org/wiki/Semantics)

From Text to Meaning

- Shallow semantics
  - Input: text
  - Output: ***lexical semantics***
- Deep semantics
  - Input: text
  - Output: ***formal semantics***

Source: Abstract Meaning Representation for Sembanking, https://amr.isi.edu/a.pdf

LOGIC format:

$\exists$ w, b, g:
instance(w, want-01) $\wedge$ instance(g, go-01) $\wedge$
instance(b, boy) $\wedge$ arg0(w, b) $\wedge$
arg1(w, g) $\wedge$ arg0(g, b)

AMR format (based on PENMAN):

```
(w / want-01
  :arg0 (b / boy)
  :arg1 (g / go-01
          :arg0 b))
```

GRAPH format:

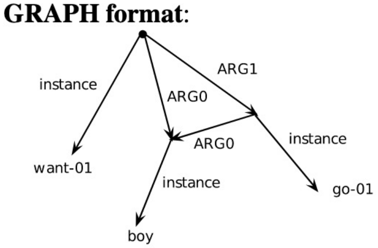


Figure 1: Equivalent formats for representating the meaning of “The boy wants to go”.

First Order Predicate Logic (FOPL)

Quick Review

Concepts

Quick Review

Constants: a, b, student123, teacher94
- Name of a specific object.

Variables: X, Y.
- Refer to an object without naming it.

Predicates: Father, Before
- Relationships between objects. May be many and may not be unique. Objects are specified as arguments (arity of a predicate).

Functions: father-of
- Mapping from objects to objects. Mapping must be present and be unique. Objects are specified as arguments (arity of a predicate).

Terms: dad-of(organism33), leftLeg(John)
- A logical expression that refers to an object

Atomic Sentences: in(dad-of(dog33), food6)
- Can be true or false
- Correspond to propositional symbols P, Q

Objects
Relations
Functions

Adapted from:
a) Dan Weld's AI course (CSE 573, Univ. of Washington)
b) Russell & Norvig, AI: A Modern Approach

FOPL - Syntax

Sentence	—	*AtomicSentence*
	|	*Sentence Connective Sentence*
	|	*Quantifier Variable,. . . Sentence*
	|	¬ *Sentence*
	|	*(Sentence)*
AtomicSentence	—	*Predicate(Term, . . .)* | *Term* = *Term*
Term	→	*Function(Term, . . .)*
	|	*Constant*
	\	*Variable*
Connective	→	⇒ | A V | ⇔
Quantifier	→	V | 3
Constant	→	*A* \ *X*\ \ *John* | · · ·
Variable	→	a | *x* *s* • • •
Predicate	→	*Before* \ *HasColor* \ *Raining* \ · · ·
Function	→	*Mother* \ *LeftLegOf* \ · · ·

Quick Review

BNF (Backus-Naur Form) grammar
of sentences in FOPL

Source: Russell & Norvig, AI: A Modern Approach

Quick Review

Connectives and Quantifiers

Logical connectives: and, or, not, =>

Quantifiers:
- ∀ : Forall
- ∃ : There exists

Examples:
1. All students: ∀ **students**
2. All students are university members:
   ∀x Student(x) => UniversityMember(x)
   (For all x, if x is a student, then x is a UniversityMember)
3. A phone: **∃x Phone(x)**
4. John has a phone:
   ∃x Phone(x) ∧ Owns(John,x)
   (There exists a phone such that John owns it.)

From Text to Meaning

- Deep semantics
  - Input: text
  - Output: ***formal semantics***

LOGIC format:

$\exists$ w, b, g:
instance(w, want-01) $\wedge$ instance(g, go-01) $\wedge$
instance(b, boy) $\wedge$ arg0(w, b) $\wedge$
arg1(w, g) $\wedge$ arg0(g, b)

AMR format (based on PENMAN):

```
(w / want-01
   :arg0 (b / boy)
   :arg1 (g / go-01
            :arg0 b))
```

GRAPH format:

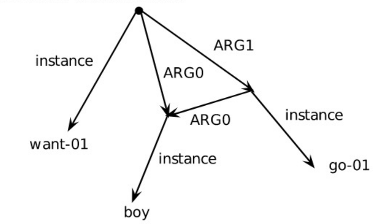


Figure 1: Equivalent formats for representating the meaning of "The boy wants to go".

Source: Abstract Meaning Representation for Sembanking, https://amr.isi.edu/a.pdf

Exercise: Asking LLM for Semantics

- shallow
- deep

Review: Common Definitions

- **Corpus** (plural corpora): a computer-readable corpora collection of text or speech.
- **Lemma**: A lemma is a set of lexical forms having the same stem, the same major part-of-speech, and the same word sense. Example: Cat and cats have same lemma.
- **Word form**: The word form is the full inflected or derived form of the word. Example: Cat and cats have different word forms.
- **Word type**: Types are the number of distinct words in a corpus. if the set of words is V, the number of types is the word token vocabulary size |V|.
- **Word tokens**: The total number N of running words in the sentence / document of interest.
- **Code switching**: use multiple languages in a code switching single communicative act – Example: Hindlish (Hindi English), Spanish (Spanish English)

"They picnicked by the pool, then lay back on the grass and looked at the stars."

- 16 tokens, 14 word types

Source: Jurafsky & Martin

Lexical Semantics

- Lemma
  - Sing, Mouse
- Word form
  - Sing, sang, sung
  - Mouse, mice
- Word sense
  - Mouse: a rodent
  - Mouse: an electronic pointing device

A lemma having many senses is called **Polysemous**

Synonymous and Similar Words

- **Synonym** - one word has a sense whose meaning is identical to a sense of another word
  - Two words are **synonymous** if they are substitutable one for the other in any sentence without changing the truth conditions of the sentence, the situations in which the sentence would be true
  - **Propositional meaning** – synonym words have the same propositional meaning (truth preserving)
- **Principle of contrast** – An assumption in linguistics is that difference in linguistic form (e.g., word form) is always associated with at least some difference in meaning
  - Water and H20 are truth preserving but used in different context
  - Synonym words are used for approximate synonymy. Then, how similar are the words?

Source: Jurafsky & Martin

Word Similarity - SimLex-999

- Captures similarity between word pairs, mining the opinions of 500 annotators via Amazon Mechanical Turk on a scale of 1 to 10

Note: *similarity,* rather than *relatedness* or *association*

- Contains
  - 666 Noun-Noun pairs,
  - 222 Verb-Verb pairs
  - 111 Adjective-Adjective pairs

vanish	disappear	9.8
behave	obey	7.3
belief	impression	5.95
muscle	bone	3.65
modest	flexible	0.98
hole	agreement	0.3

Source: Jurafsky & Martin

- **Usage**: Evaluation of learning based approaches for finding word similarity by correlation

SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation. 2014. Felix Hill, Roi Reichart and Anna Korhonen. *Computational Linguistics.* 2015
Website: https://fh295.github.io/simlex.html

Meaning (Semantics) versus Structure (Lexical)

Pair	Simlex-999 rating	WordSim-353 rating
coast - shore	9.00	9.10
clothes - closet	1.96	8.00

Example courtesy: https://fh295.github.io/simlex.html

Word Relatedness/ Association

- **Semantic Field:** related words from the same particular domain and bear structured relations with each other.
  - Example 1: cup, coffee
  - Example 2: scalpel, surgeon
  - Usually determined by experts in a field

Source: Jurafsky & Martin

- **Word Association Test/ Task:** how word meaning is stored in memory
  - Have people respond to word associations as a game; e.g., say the first word that comes to mind when one says "Doctor"
  - Applications
    - Used in marketing
    - Also evaluation of learning procedures discovering meaning (e.g., word embedding)

Sources:
- https://psychology.jrank.org/pages/656/Word-Association-Test.html,
- Establishing the Reliability of Word Association Data for Investigating Individual and Group Differences , Tess Fitzpatrick, David Playfoot, Alison Wray, Margaret J. Wright *Applied Linguistics*, Volume 36, Issue 1, February 2015, Pages 23–50, https://doi.org/10.1093/applin/amt020

Discovering Word Relatedness

- **Topic model**: a statistical notion of related words in a document. Hope is that meaningful topics will be from the same semantic field, but there is no guarantee
- Key idea
  - Topic: group of words
  - Counting words and grouping similar word patterns to infer topics within unstructured data.
  - Assumptions
    - Distributional hypothesis: similar topics make use of similar words
    - Statistical mixture hypothesis: documents talk about several topics
  - Perform unsupervised analysis/ clustering: given a corpus and number of topics (k), find k topics that are representative of key ideas in the corpus

References:
- Blog: https://monkeylearn.com/blog/introduction-to-topic-modeling/
- Paper: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf
- Tool: Gensim
  - Docs: https://radimrehurek.com/gensim/auto_examples/index.html#documentation
  - Sample usage - https://github.com/biplav-s/course-nl-f22/blob/07f1b1d68a044a8558894ee96611e27d3fc242a9/sample-code/l19-topic/ExploreTopics.ipynb

Course Project

Discussion: Course Project

Theme: Analyze quality of official information available for elections in 2024 [in a state]

- Take information available from
  - Official site: State Election Commissions
  - Respected non-profits: League of Women Voters
- Analyze information
  - State-level: Analyze quality of questions, answers, answers-to-questions
  - Comparatively: above along all states (being done by students)
- Benchmark and report
  - Compare analysis with LLM
  - Prepare report

- Process and analyze using NLP
  - Extract entities
  - Assess quality – metrics
    - Content – *Englishness*
    - Content – *Domain* -- election
  - … other NLP tasks
  - Analyze and communicate overall

Major dates for project check

- Sep 10: written – project outline
- Oct 8: in class
- Oct 31: in class // LLM
- Dec 5: in class // Comparative

Review current states chosen by others

Project Discussion

1. Go to Google spreadsheet against your name
2. Enter the **state** you will focus on for course project

1. Create a private Github repository called "CSCE771-Fall2024-<studentname>-Repo". Share with Instructor (biplav-s) and TA (vr25)
2. Create Google folder called "CSCE771-Fall2024-<studentname>-SharedInfo". Share with Instructor (prof.biplav@gmail.com) and TA (rawtevipula25@gmail.com)
3. Create a Google doc in your Google repo called "Project Plan" and have the following by Friday (Aug 30, 2024)

Timeline

1. Title: Analyze quality of official information available for elections in 2024 in <state>
2. Data need:
   1. Official: state's election commission
   2. LWV: https://www.vote411.org/
3. Methods:
4. Evaluation:
5. Milestones
   - Sep 10: written and feedback
   - Oct 8: in class
   - Oct 31: in class
   - Dec 5: in class

Obtaining Election Data

Here are a few things to do:
A) **Official data** backed by laws: state election commission
a) Find the state's election commission
b) Find the Q/As they provide. They may be as FAQs or on different web pages.
c) Collect the Q/A programmatically

B) **Secondary data** sources: non-profit
a) Find Q/As from Vote 411 which is supported by the non-profit: LWV.

For reference, for SC,

A) Official - https://scvotes.gov/voters/voter-faq/

B) Secondary - https://www.vote411.org/south-carolina

For extraction, one or more approaches:
- Manually annotating
- BeautifulSoup,
- Tika
- or other open source libraries.

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

Review current states chosen by others

Course Project – Deadlines and Penalty Rubric

- Penalty
  - Missing milestones: **[-10%]**
  - Maximum: **[-40%]**
- Bonus possible
  - if two or more states considered
  -

Timeline

1. Title: Analyze quality of official information available for elections in 2024 in <state>
2. Data need:
   1. Official: state's election commission
   2. LWV: https://www.vote411.org/
3. Methods:
4. Evaluation:
5. Milestones
   - Sep 10: written and feedback
   - Oct 8: in class
   - Oct 31: in class
   - Dec 5: in class

Lecture 8: Concluding Comments

- We looked at evaluation measures
  - accuracy, precision, recall, F1
  - Macro and micro averages
- We also started to look at semantics

Concluding Segment

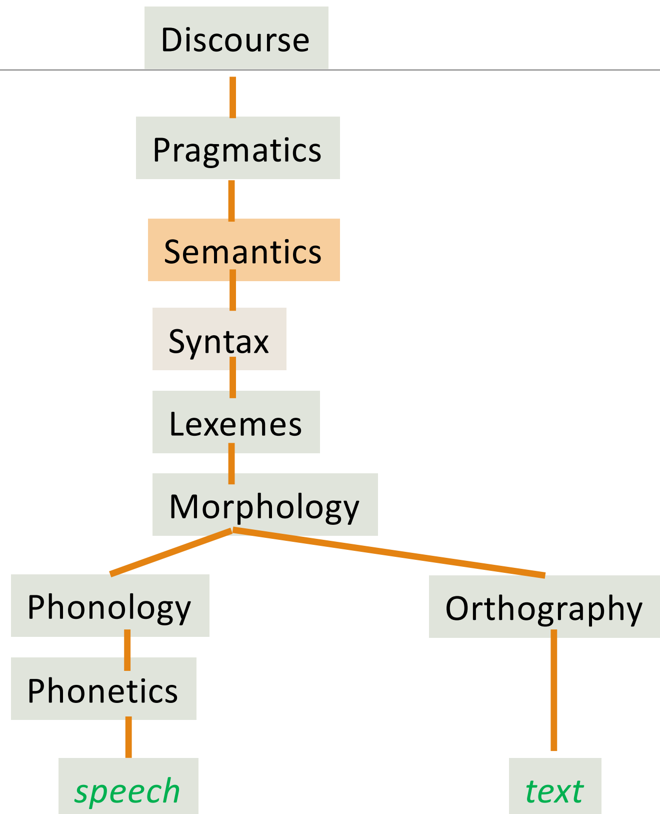
About Next Lecture – Lecture 9

Lecture 9: Semantics, ML

- Complete semantics discussion
- Discuss ML for NLP

4	Aug 29 (Th)	NLP Tasks, Case Study – Business Application		
5	Sep 3 (Tu)	Parsing, Paper 1 discussion; project topics review		Practice exercise
6	Sep 5 (Th)	Project topics review, statistic Parsing		
7	Sep 10 (Tu)	Statistical parsing, QUIZ		Quiz 1, **Project Check**
8	Sep 12 (Th)	Evaluation, Semantics		Coding running example
9	Sep 17 (Tu)	Semantics, Machine Learning for NLP		Code: scikit f1 score package, Code: ConceptIO
10	Sep 19 (Th)	Towards Language Model: Vector embeddings, Embeddings, CNN/ RNN		Code: embedding, genism word vector, tf-idf

Summary of NLP Analysis



- **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