COSC 6336
Natural Language Processing

Instructor: Thamar Solorio

Some content in these slides was adapted from Prof. Rada Mihalcea

UNIVERSITY of **HOUSTON**

Your Instructor

Thamar Solorio, PhD

Associate Professor

Email: thamar.solorio@gmail.com

Personal webpage: http://solorio.uh.edu/

Office hours: MW 5:30-6:30, in PGH 584

Your TA

Niloofar Safi Samghabadi

Email: niloofar.safi@gmail.com

Personal webpage: https://sites.google.com/view/nsafi

Office hours: W 1:00-3:00, in PGH 550 A

Logistics and Administrivia

Piazza website (this is where you go to post questions):
piazza.com/uh/spring2020/cosc6336

We will also try out Perusall.com. Create an account at https://app.perusall.com

Use code: SOLORIO-S5MPV

More Administrivia

The course will mostly draw the readings from two freely available online books:

[3rd Edition Book from Jurafsky and Martin](#) (J&M)

[Natural Language Processing by Eisenstein](#) (E)

Tentative schedule

Lecture		Topic	Readings	Assignment
	1	ADMINISTRIVIA, INTRODUCTION, Text Processing		
	2	Classification: binary and multilabel (Logistic Regression)	J&M 5.0-5.6	
	3	HMMs 1 and POS Tagging	J&M 8, E 7.1-7.4	
	4	HMMs 2 and POS Tagging		
	5	CRFs	E 7.5 & 8.3	CRFs for NEs due 2/10
	6	Grammars and Syntactic parsing	J&M 12	
	7			
	8	Dependency parsing	J&M 15.0-15.4, 15.6	
	9	Neural Networks 1	E 3.0-3.3, G 4.0-4.5	
	10	Neural Networks 2	https://arxiv.org/pdf/1708.00214.	Release specs for Final Project (proposal due 3/4)
	11	RNNs	J&M 9.0-9.6	
	12	RNNs		RNN Project due 03/01/2020
	13	Vector Semantics & Embeddings	J&M 6, G 5, E 14.0-14.6	
	14	Word Embeddings 2		
	15	Final Project Pitches		Final project proposals due 03/17/2020
	16	Language Models (count-based)	J&M, E 6.0-6.2, 6.4	
	17	Neural Language Models 1	E 6.3	LM Project due 03/29/2020
	18	NLM 2 & ELMo	Elmo paper	
	19	Seq2Seq, Attention	Vaswani et al., 2017	
	20	Seq2seq cont		Seq2seq project due 01/07/2020
	21	Pretrained transformers: GPT, BERT, RoBERTa	BERT	
	22	Pretrained transformers: GPT, BERT, RoBERTa	GPT	
	23	NLP Applications	RoBERTa	
	24			
	25	Last Class		
		Final Project Presentations		
		Final Project Presentations		

Grading Criteria

★ 50% - Assignments (mini projects, 3-4 total)
★ 40% - Final Project
★ 10% - In-class participation and quizzes.

Natural?

Natural Language?

language spoken by people, e.g. English, Japanese, Swahili, as opposed to artificial languages, like C++, Java, etc.
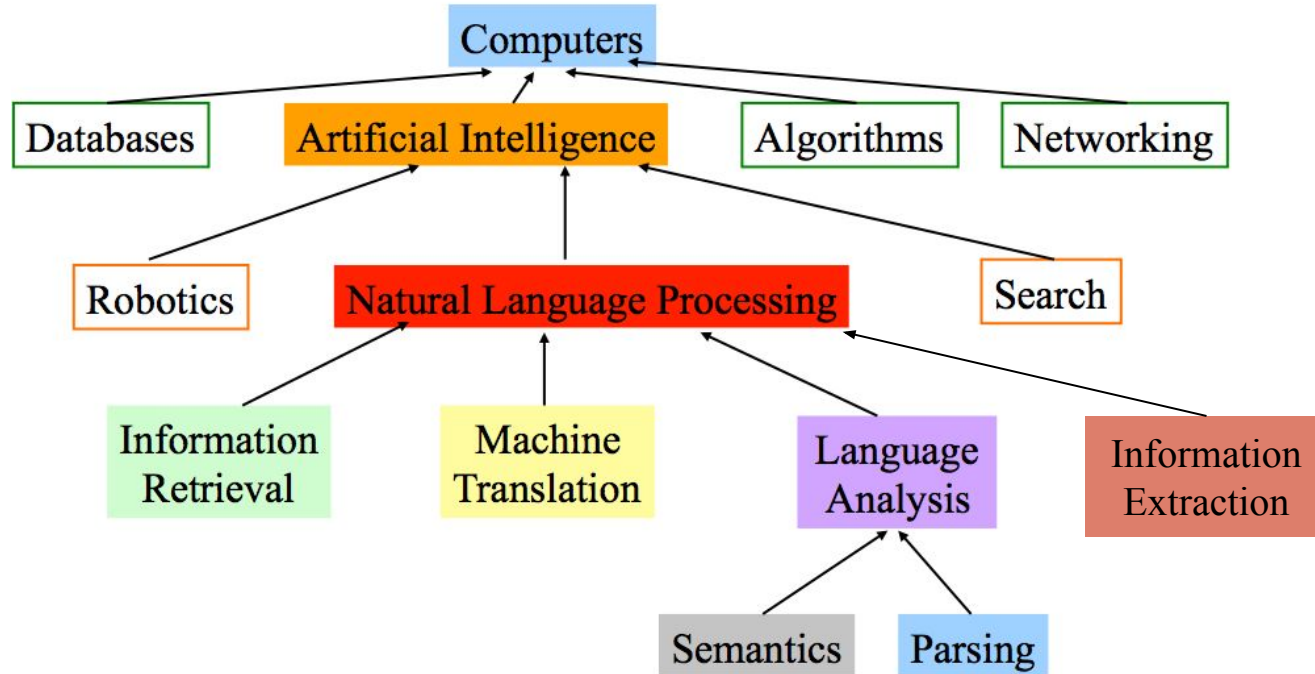
Natural Language Processing

Applications that deal with natural language in a way or another

Computational Linguistics

Doing linguistics on computers, more on the linguistic side than NLP, but closely related

Where does NLP fit in the COSC taxonomy?

Why Natural Language Processing?

★ " … language is what made us human" (Guy Deutscher)

★ Through language humans:
   ○ Pass on knowledge
   ○ Create new thoughts and ideas
   ○ Express deep (and not so deep) reflections

Why Natural Language Processing?

★ We want to know what's happening in a timely manner!

    ○ To react quickly and effectively

★ People use social media every day
    ○ Around 6,000 tweets are tweeted every second!

★ People post about almost anything they can

    ○ Behavioral aspects, current events, trending brands…

Why is this important?

★ What about the Hurricane Harvey?

Rae (Roca) Pickett
@RaeRoca

Follow

Anyone in North Houston have a boat and can rescue a 3 &6 yo, mom, gramma and grandpa and 2 dogs? 1960 Ella street #houstonflood #Rescue

3:44 PM - 27 Aug 2017 from Annandale, VA

662 Retweets 854 Likes

38   662   854

Why is this important?

★ What about the Hurricane Harvey?

Rae (Roca) Pickett
@RaeRoca
Follow

Anyone in North Houston have a boat and can rescue a 3 &6 yo, mom, gramma and grandpa and 2 dogs? 1960 Ella street #houstonflood #Rescue

MELO
@4GP5
Follow

5048 Shreveport needs assistance right now in Port Arthur! 6ft of water in the house!

5:25 PM - 30 Aug 2017 from Port Neches, TX

69 Retweets 16 Likes

5   69   16

Why is this important?

★ What about the Hurricane Harvey?

Rae (Roca) Pickett
@RaeRoca
Follow

MELO
@4GP5
Follow

5048 Shreveport needs assistance right now in Port Arthur! 6ft of water in the house!

crimsonboudoir
@crimsonboudoir
Follow

#HELP @uscoastguard ELDERLY WOMAN ALONE NEEDS #IMMEDIATE ASSISTANCE 3 ft of water in house trying to get phone # 729 LINKWOOD Port Aurthur

Jekera hebert @Jekeralenaeee
Replying to @Jekeralenaeee
Port Arthur Texas 77642 ..I'm in Houston she's alone

12:20 AM - 30 Aug 2017

1 Retweet 1 Like

1   1   1

Why is this important?

★ What about the Hurricane Harvey?

Rae (Roca) Pickett
@RaeRoca
Follow

MELO
@4GP5
Follow

crimsonboudoir
Follow

ht now
e!

AN
CE 3
729

David Sloves
@Johnny_Champagn
Follow

Replying to @realDonaldTrump

My family is still waiting for a water rescue in the attic, they can't get to roof! 6606 Reamer St
Houston, TX 77074
Please RT

12:17 PM - 27 Aug 2017

112 Retweets **71** Likes

💬 24   ⟲ 112   ♡ 71

Why is this important?

★ What about the Hurricane Harvey?
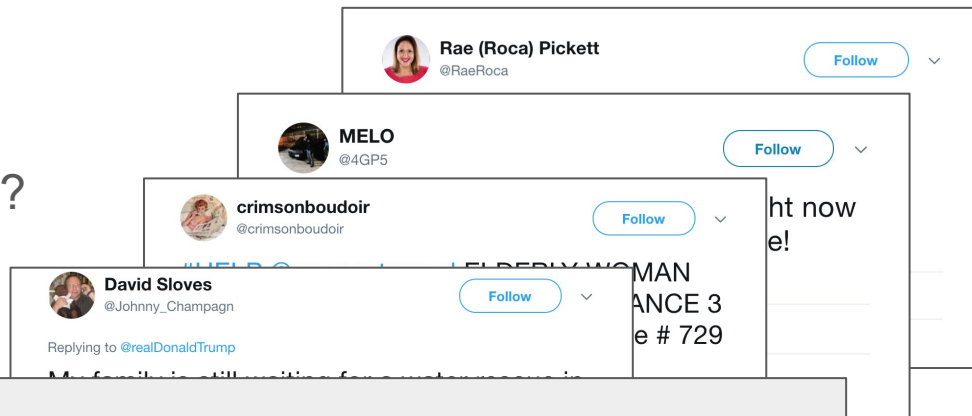
How valuable would have been to systematically detect *the who*, *the when* and *the what* of those urgent messages?

Why Natural Language Processing?

★ Practical value:
  ○ Companies want to know what consumers are saying
  ○ Intelligence communities want to know what persons of interest are planning
  ○ New products that use language as the interface with humans

★ Scientific value:
  ○ Gain a deeper understanding of how the human brain is able to process language

★ Computer programs that can process this data require NLP expertise

What's difficult about Natural Language Processing?

★   kJfmmfj mmmvvv nnnffn333

★   Uj iheale eleee mnster vensi credur

★   Baboi oi cestnitze

★   Coovoel2^ ekk; ldsllk lkdf vnnjfj?

★   Fgmflmllk mlfm kfre xnnn!

Computers lack knowledge!

★ People have no trouble understanding language
  ○ Common sense knowledge
  ○ Reasoning capacity
  ○ Experience
★ Computers have
  ○ No common sense knowledge
  ○ No reasoning capacity

Levels of Analysis

★ Speech
  ○ Phonology
★ Text
  ○ Morphology
  ○ Syntax
  ○ Semantics
  ○ Pragmatics
★ Interaction between levels

Some NLP applications

★ Speech recognition
  ○ Voicemail transcription

★ Dialogue systems
  ○ Alexa, Google home, Siri, Cortana

★ Information extraction
  ○ Named Entity Recognition and Linking
  ○ Event detection

★ Machine translation
  ○ Text to text
  ○ Speech to speech

Challenges in NLP

Main issue is ambiguity ….

Ambiguity in Speech

★ 264 Lane Street vs. 26 four-lane street

★ For invoices vs foreign voices

★ Colorectal cancer risks vs co-director cancel risks

★ Frapuccino vs Fred Paccino
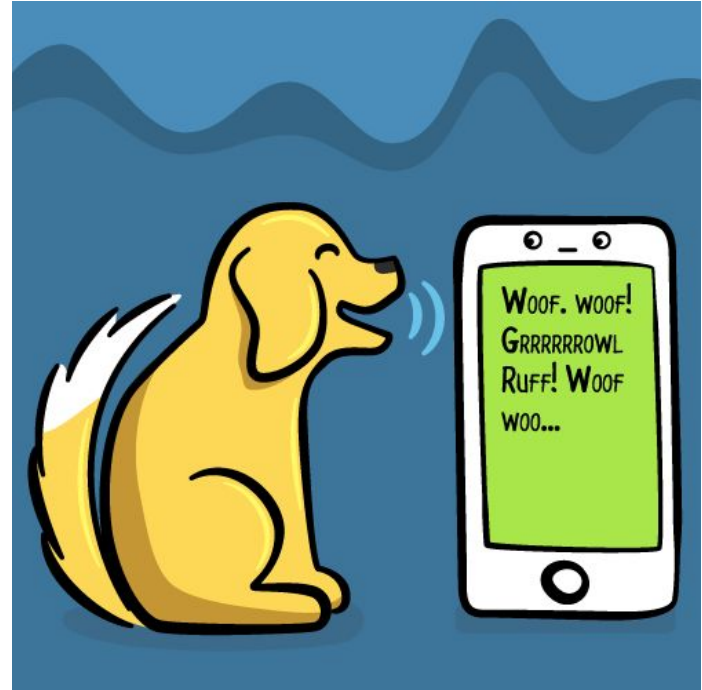
Ambiguity in Morphological Analysis

ride	rideable
do	doable
like	likeable

★ Pattern: Verb + "able" → Adjective (able to do/be Verb-ed)

Ambiguity in Morphological Analysis

happy	unhappy
cool	uncool
stable	unstable

★ Pattern: "un" + Adjective → Adjective (not Adjective)

UNIVERSITY of HOUSTON

Ambiguity in Morphological Analysis

do	undo
zip	unzip
dress	undress

★ Pattern: "un" + Verb → Adjective (to reverse Verb-ing)

Ambiguity in Morphological Analysis

What about the word unlockable?

UNIVERSITY of **HOUSTON**
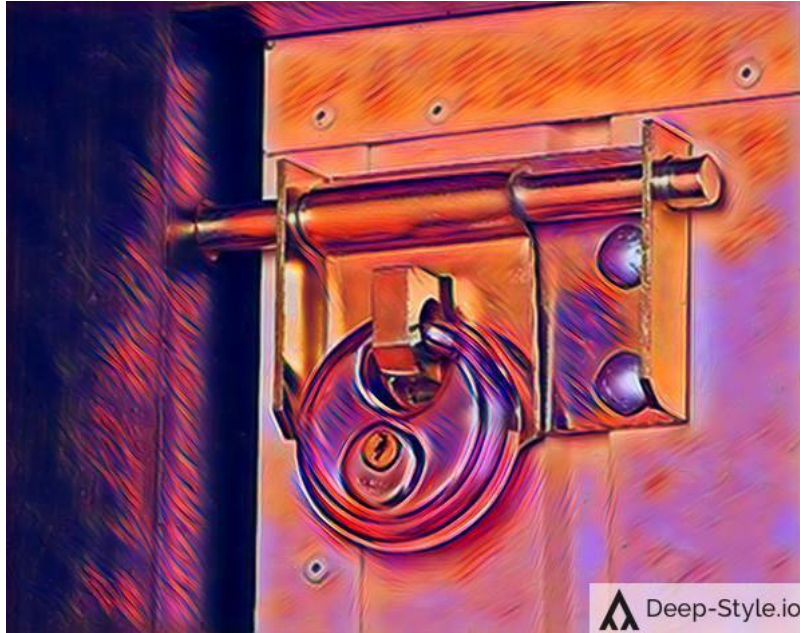
Ambiguity in Morphological Analysis

unlockable

Option 1:

"un" + lock (Verb) → unlock (Verb) (to reverse locking)

unlock + "able" →   unlockable (Adjective) (able to unlock)

Option 2:

lock + "able" → lockable (Adjective) (able to lock)

"un" + lockable → unlockable (Adjective) (not able to lock)

Ambiguity in Syntax

Ambiguity in Syntax

★ Jake told Mike he has cancer

★ Eat spaghetti with meatballs vs eat spaghetti with chopsticks

★ We saw the Eiffel Tower flying to Paris

★ Old men and women

Small surface changes → Big changes in interpretation

Only he told her that he loved her

Small surface changes → Big changes in interpretation

Only he told her that he loved her
He only told her that he loved her

Small surface changes → Big changes in interpretation

Only he told her that he loved her
He only told her that he loved her
He told only her that he loved her

Small surface changes → Big changes in interpretation

Only he told her that he loved her
He only told her that he loved her
He told only her that he loved her

Small surface changes → Big changes in interpretation

Only he told her that he loved her
He only told her that he loved her
He told only her that he loved her
He told her only that he loved her

Small surface changes → Big changes in interpretation

Only he told her that he loved her
He only told her that he loved her
He told only her that he loved her
He told her only that he loved her
He told her that only he loved her

Small surface changes → Big changes in interpretation

Only he told her that he loved her
He only told her that he loved her
He told only her that he loved her
He told her only that he loved her
He told her that only he loved her
He told her that he only loved her

Small surface changes → Big changes in interpretation

Only he told her that he loved her
He only told her that he loved her
He told only her that he loved her
He told her only that he loved her
He told her that only he loved her
He told her that he only loved her
He told her that he loved only her

Small surface changes → Big changes in interpretation

Only he told her that he loved her
He only told her that he loved her
He told only her that he loved her
He told her only that he loved her
He told her that only he loved her
He told her that he only loved her
He told her that he loved only her
He told her that he loved her only

More Issues in Syntax

Anaphora resolution:

"The <u>dog</u> entered my room. <u>It</u> scared me"

Preposition attachment:

"I saw the man in the park <u>with </u>a telescope"

Issues in Semantics

Understand language, but how?

★ "plant" = industrial plant
★ "plant" = living organism

Words are ambiguous by design

What is the relevance of getting the semantics right?

★ Machine translation (wrong translations)
★ Information retrieval (wrong information)

Challenges in Information Extraction

The closure of California's main coastal road demonstrates just how badly these mudslides have damaged this picturesque seaside town, which is simultaneously reeling from the flooding-related deaths of at least **20** residents following the storm early Tuesday morning.

The storm destroyed at least **65** homes and damaged at least **460** more, authorities said. Firefighters are continuing their painstaking work of combing through the debris with heavy equipment and hand tools, aware more bodies are likely buried beneath. At least four people remain missing.

How many deaths? 20? 65? 460? Where? California's coastal road

How many houses destroyed? 20? 65? 460?

UNIVERSITY of **HOUSTON**

Challenges in Information Extraction

Detect new patterns:

★   Detect hacking / hidden information / etc./

★   Gov. mil, puts lots of money into IE research

Challenges in Information Retrieval

★ General model:
  ○ A huge collection of texts
  ○ A query
★ Task: find documents that are relevant to the given query
★ How? Create an index, like the index in a book
★ More …
  ○ Vector-space models
  ○ Boolean models
★ Examples: Google, Yahoo, Baidu, etc.

Challenges in Information Retrieval

★ Retrieve specific information

★ Question Answering:
   ○ What's the age of the Earth?
   ○ Approx. ~ 4.5 billion years

★ Cross Language Information Retrieval

★ What's the minimum age requirement for car rental in Italy?

★ Integrate large number of languages

★ Integrate into performant IR engines

So Far...

Lots of interesting challenges that require NLP to address them!

First Reading Assignment

★ J&M 5.0-5.6

   Due 01/15/2020 Before coming to class

Next Class

★ Review of classification, binary and multiclass