Introduction to NLP

Dr. Demetrios Glinos

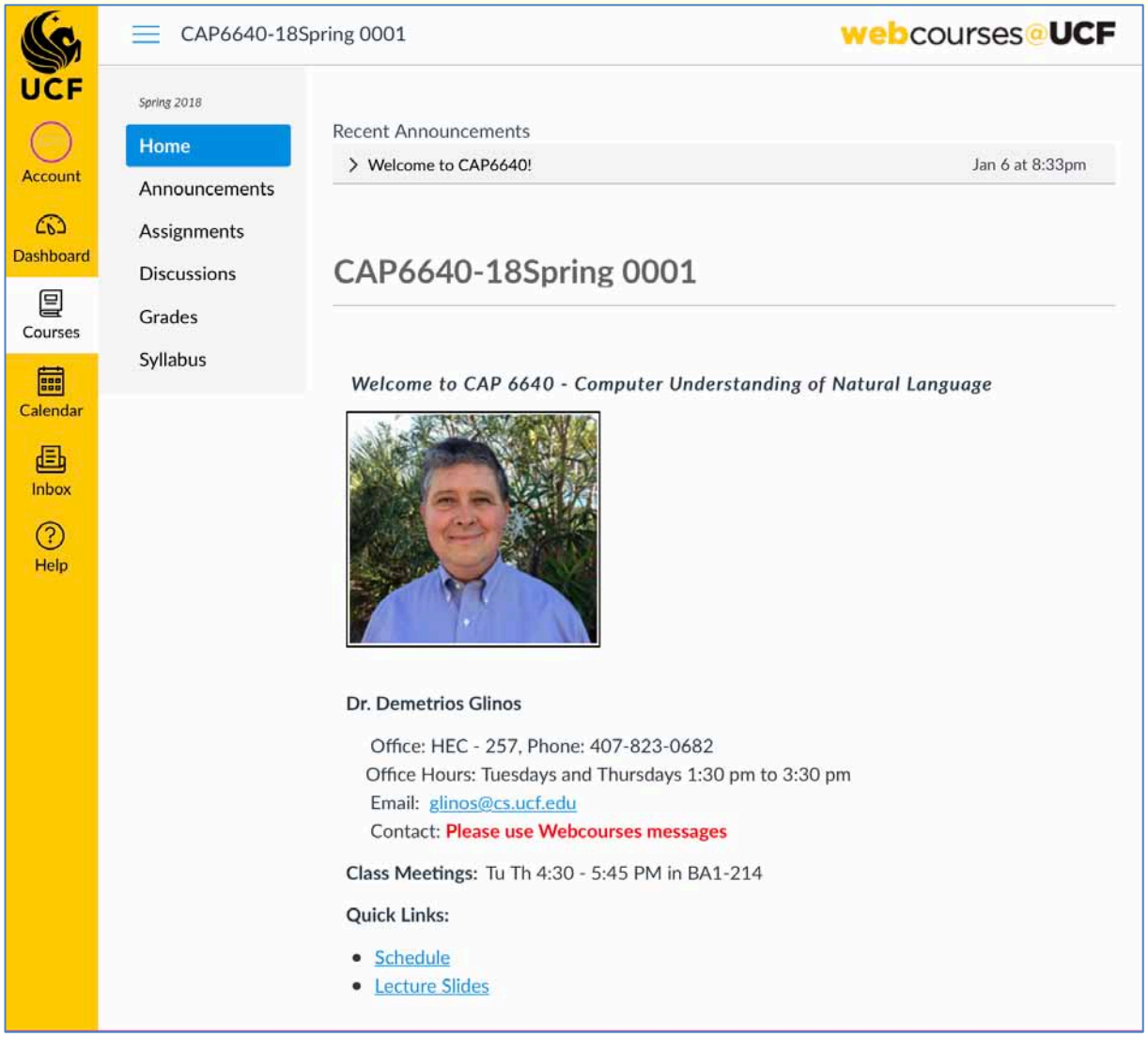
CAP6640 – Computer Understanding of Natural Language

Spring 2018

Today

- Course logistics
- What is natural language processing?
- What can natural language processing do?
- What this course will cover
- Action items for this week

Course Home Page



Course Schedule

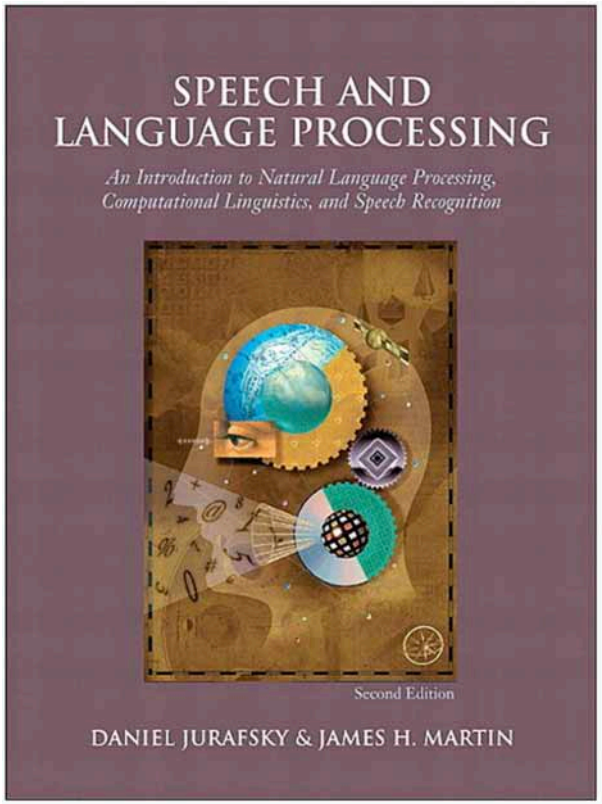
Weeks/Dates	Topics Covered	References/Assignments
Week 1 1/8 – 1/12	Course Introduction Basic Text Processing	J&M Ch. 1, 2 Engagement assignment due Friday (or when join class)
Week 2 1/15 - 1/19	Minimum Edit Distance Language Modeling	J&M Ch. 3, 4
Week 3 1/22 - 1/26	Text Classification and Naive Bayes Sentiment Analysis	KM Ch. 5
Week 4 1/29 - 2/2	Hidden Markov Models and POS Tagging Maximum Entropy Classifiers	J&M Ch. 5, 6 Program 1 due Sunday 2/4
Week 5 2/5 - 2/9	Information Extraction and NER Relation Extraction	J&M Ch. 22
Week 6 2/12 - 2/16	Neural Network Basics Deep Neural Networks	GBC Ch. 6
Week 7 2/19 - 2/23	Statistical Natural Language Parsing Dependency Parsing	J&M Ch. 12, 14 Project Proposal Due in Class 2/22 Program 2 due Sunday 2/25
Week 8 2/26 - 3/2	Information Retrieval Ranked Information Retrieval	J&M Ch. 22
Week 9 3/5 - 3/9	Tuesday: Spelling Correction Thursday: Midterm Exam	J&M Ch. 4 Mid-Term Exam on Thursday
Week 10 3/12 - 3/16	SPRING BREAK	Program 3 due Sunday 3/18
Week 11 3/19 - 3/23	Vector Semantics Dense Vectors	J&M Ch. 20 *(Withdrawal deadline Monday 3/21)*
Week 12 3/26 - 3/30	Question Answering Summarization	J&M Ch. 23 Project Report Due Sunday 4/1
Week 13 4/2 - 4/6	Project Presentations	
Week 14 4/9 - 4/13	Project Presentations	
Week 15 4/16 - 4/20	Project Presentations	
Week 16 4/23 - 4/27	*No classes this week*	

J&M: Jurafsky & Martin
KM : Kevin Murphy
GBC : Goodfellow, Bengio & Courville

Recommended Textbooks

Speech and Language Processing (2nd Ed.), Daniel Jurafsky and James H. Martin, Pearson, 2009

- **ISBN-13:** 978-0131873216
- **ISBN-10:** 0131873210



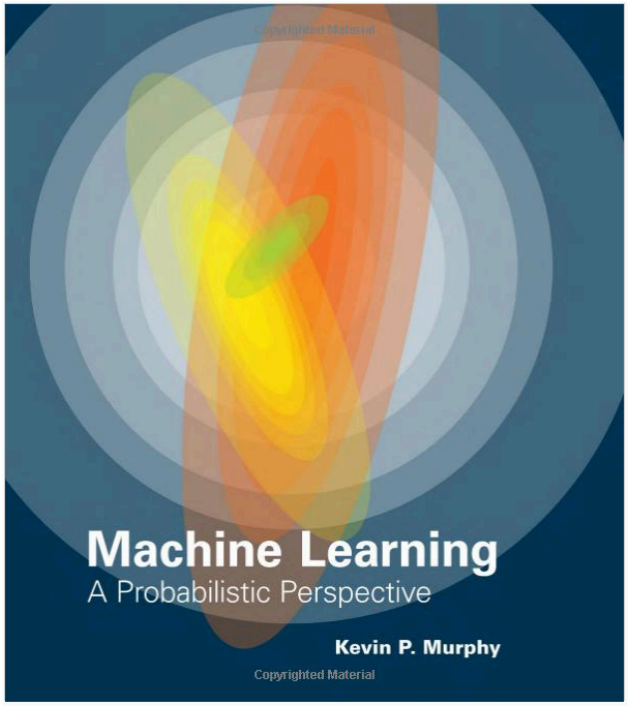
- The primary reference for this course
- Draft sections from upcoming 3rd edition may be found at

https://web.stanford.edu/~jurafsky/slp3/

Recommended Textbooks

Machine Learning: A Probabilistic Perspective, Kevin Murphy, The MIT Press, 2012

- **ISBN-13:** 978-0262018029
- **ISBN-10:** 0262018020



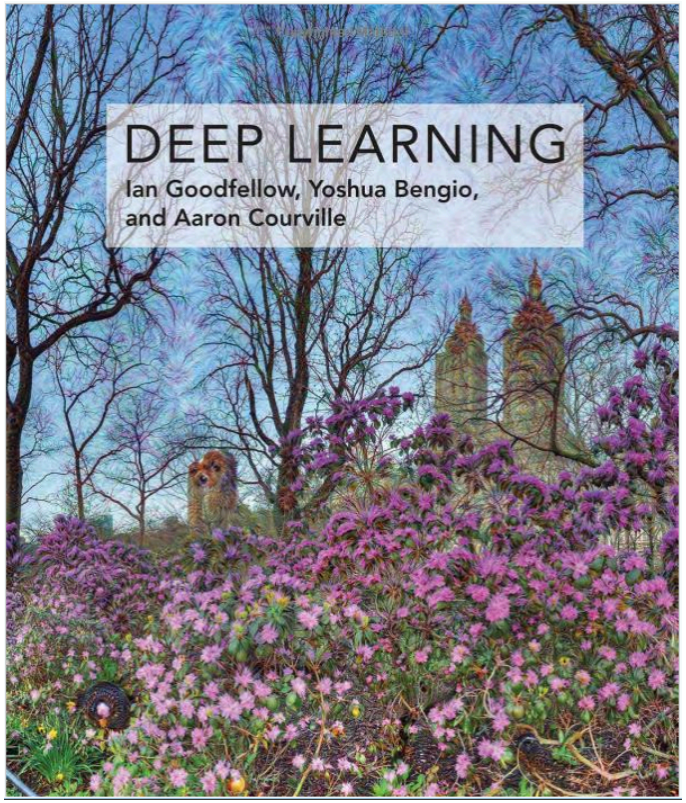
- Contains a comprehensive review of probability theory, regression models, and other relevant topics

Recommended Textbooks

Deep Learning, Ian Goodfellow, Yoshua Bengio, and Aaron Courville, The MIT Press, 2016

- **ISBN-13:** 978-0262035613
- **ISBN-10:** 0262035618

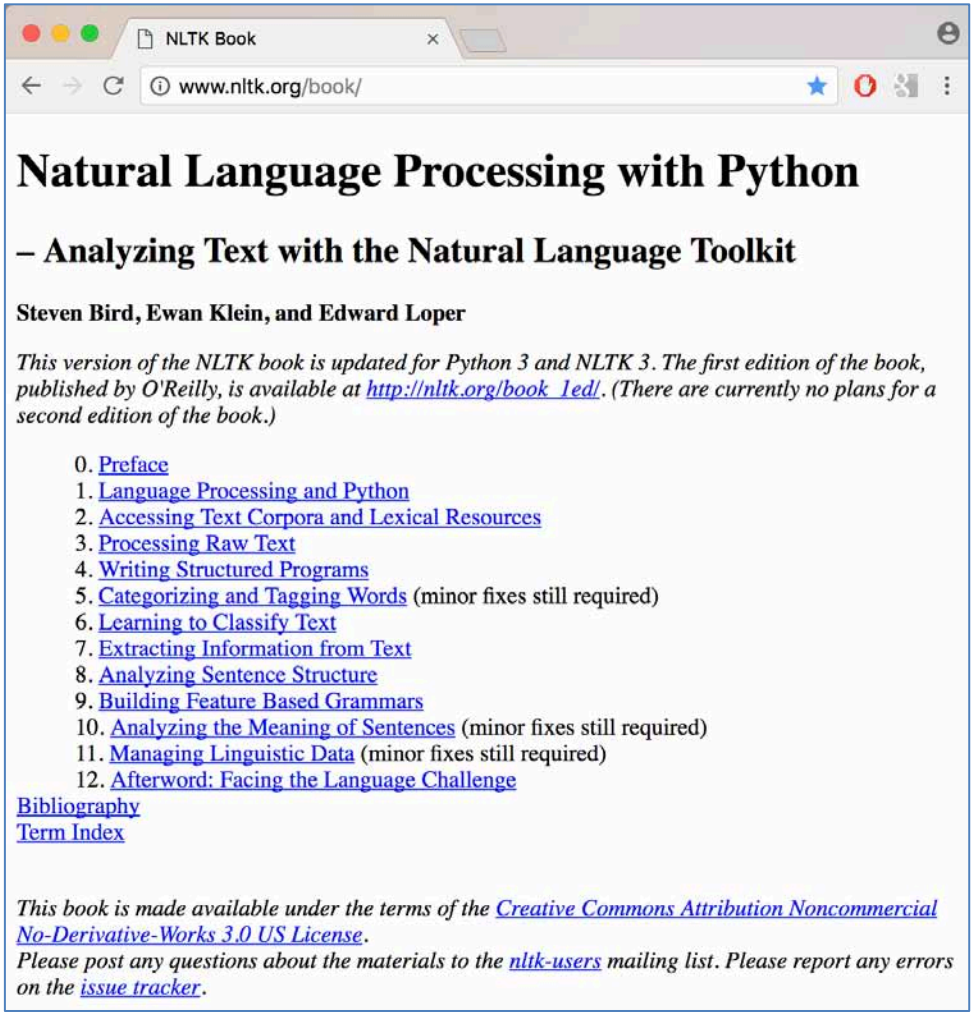
- Good reference for deep learning concepts



Recommended Resource

- Natural Language Toolkit (NLTK)
  - www.nltk.org
  - suite of libraries and programs for symbolic and statistical NLP
  - written in Python
  - includes Gutenberg, Brown, WordNet, and CONLL2000 corpora

  - useful companion text:
    - Natural Language Processing with Python, S. Bird, E. Klein, and E. Loper, O'Reilly, 2009
    - www.nltk.org/book



Additional Resources

- Stanford Core NLP software
  - https://stanfordnlp.github.io/CoreNLP/
- Python for scientific computing
  - http://cs231n.github.io/python-numpy-tutorial/
- Linear algebra review
  - http://cs229.stanford.edu/section/cs229-linalg.pdf
- Probability theory review
  - Chapter 2 of Kevin Murphy text, plus many others

Grading Policy

(45%) Programming assignments
(20%) Midterm Exam
(20%) Project Report
(10%) Project Presentation
(5%) Class Participation

Letter Grade	Range
A	90 and above
B	80 to 89
C	70 to 79
D	60 to 69
F	below 60

Notes:

- each category will use 100-point scale
- compute weighted sum
- add 1 point for engagement assignment
- round to nearest integer
- convert to letter grade, per chart above
- plus/minus grading will not be used

Program Assignments and Midterm

- Programming Assignments (45%)
  - Three programming assignments, 15% each
  - Programs must be done individually, not in groups
  - Must be written in C, C++, Java, or Python
    - must be able to run on instructor's machine
      - macOS 10.13.2 (High Sierra),
      - Java 1.8.0_152,
      - Python 3.6.2,
      - gcc/g++ 4.2.1
    - must not require installation of additional software
  - Include instructions for running the program in source file header
  - Submit on Webcourses
- Midterm Exam (20%)
  - On Thursday, 3/8/18 (right before Spring break)

Term Project and Participation

- Term Project (30%)
  - In lieu of final exam
  - Topic must be related to NLP
  - Research must involve development of a program (any language)
- Teams of 2-3 students
- Project Proposal due 2/22
- Project Report due 4/1
  - academic paper format: introduction, related work, problem formulation, experimental results, conclusion, and references
- Presentations during last 3 weeks of class
  - slides due on date of presentation

- Class Participation (5%)
  - Will review presentations of other teams
  - Review forms will be provided

Late Submissions

- Late submission policy
  - Late up to 24 hours: 50% deduction
  - More than 24 hours: 0 credit
- Policy applies to
  - Program assignments
  - Project report
  - Project presentation slides

Academic Integrity

- You are expected to adhere to the highest standards of academic honesty and integrity
- Academic misconduct by students in any form, including cheating and plagiarism, will not be tolerated
- Refer to the Syllabus and UCF's Golden Rule for details

Today

- Course logistics
- What is natural language processing?
- What can natural language processing do?
- What this course will cover
- Action items for this week

What is NLP?

- natural language
  - any language used for everyday communications by humans
  - constantly evolving
  - typically with complex features, nuances, and ambiguities

- natural language processing (NLP)
  - any computer processing of natural language
  - applications that require knowledge of natural language in some manner

Today

- Course logistics
- What is natural language processing?
- What can natural language processing do?
- What this course will cover
- Action items for this week

Question Answering

- IBM's Watson system won "Jeopardy" against human opponents (2011)

- Game situation
  - contestants are given the "answer" and must determine the appropriate question
  - example answer:

    William Wilkinson's "An Account of the Principalities of Wallachia and Moldovia" inpired this author's most famous novel

  - correct question: "Who was Bram Stoker?"

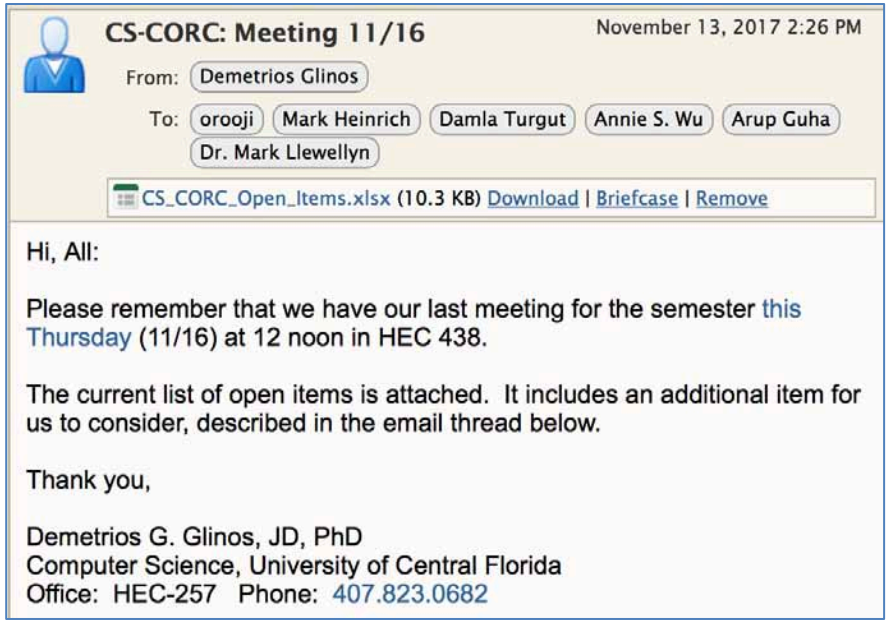
Note:

- *the novel in question is "Dracula"*
- *both human contestants also got the correct answer*

Information Extraction

- Herein of extracting structured information from free text

- Example: extracting calendar events from emails



CS-CORC: Meeting 11/16

November 13, 2017 2:26 PM

From: Demetrios Glinos

To: orooji, Mark Heinrich, Damla Turgut, Annie S. Wu, Arup Guha, Dr. Mark Llewellyn

CS_CORC_Open_Items.xlsx (10.3 KB) Download | Briefcase | Remove

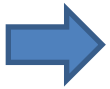
Hi, All:

Please remember that we have our last meeting for the semester this Thursday (11/16) at 12 noon in HEC 438.

The current list of open items is attached. It includes an additional item for us to consider, described in the email thread below.

Thank you,

Demetrios G. Glinos, JD, PhD
Computer Science, University of Central Florida
Office: HEC-257 Phone: 407.823.0682



Event:	CS-CORC meeting
Date:	Nov. 16, 2017
Start:	12:00 pm
End:	12:30 pm (default)
Location:	HEC-438

Information Extraction & Sentiment Analysis

- Herein of categorizing extracted information according to sentiment

Charles Perry

Great!

July 1, 2017

Style: 18-55mm | Configuration: Base | Verified Purchase

I bought it as a gift for my Mom who had a 35 mm rebel. This camera uses the same lenses she already had and that was a very important selling point. It it easy to use, relatively simple to use, but has many great features. I am happy I got it and my mom, is enjoying it very much.

Comment | Was this review helpful to you? Yes No Report abuse

Ryan

BUY BUY BUY

February 23, 2015

Style: 18-55mm | Configuration: Base | Verified Purchase

You're getting a great camera have had it for 2 weeks now. Amazing shots and once you get some practice in upgrade the lens and you'll be blown away. I'm pretty sure the lens will fit the t5i also so when you upgrade you'll have those lenses but don't quote me on that I'm still new to this. I'm happy I have a new hobby and can now take great picture on my family trips. All in all yes buy this over a Nikon the sensor is more advanced and can pick up light better from what my friend that's a pro photographer has told me.

Comment | 7 people found this helpful. Was this review helpful to you? Yes No Report abuse

Rick

Good Buy

May 1, 2016

Style: 18-55mm | Configuration: Base | Verified Purchase

Just starting the learning process and love the functionality of the camera Picture quality is excellent so far as I have tried.
Very happy with this purchase.

Comment | One person found this helpful. Was this review helpful to you? Yes No Report abuse

Kat Heckenbach VINE VOICE

Easy to use for beginners

November 21, 2014

Style: 18-55mm | Configuration: Base | Vine Customer Review of Free Product (What's this?)

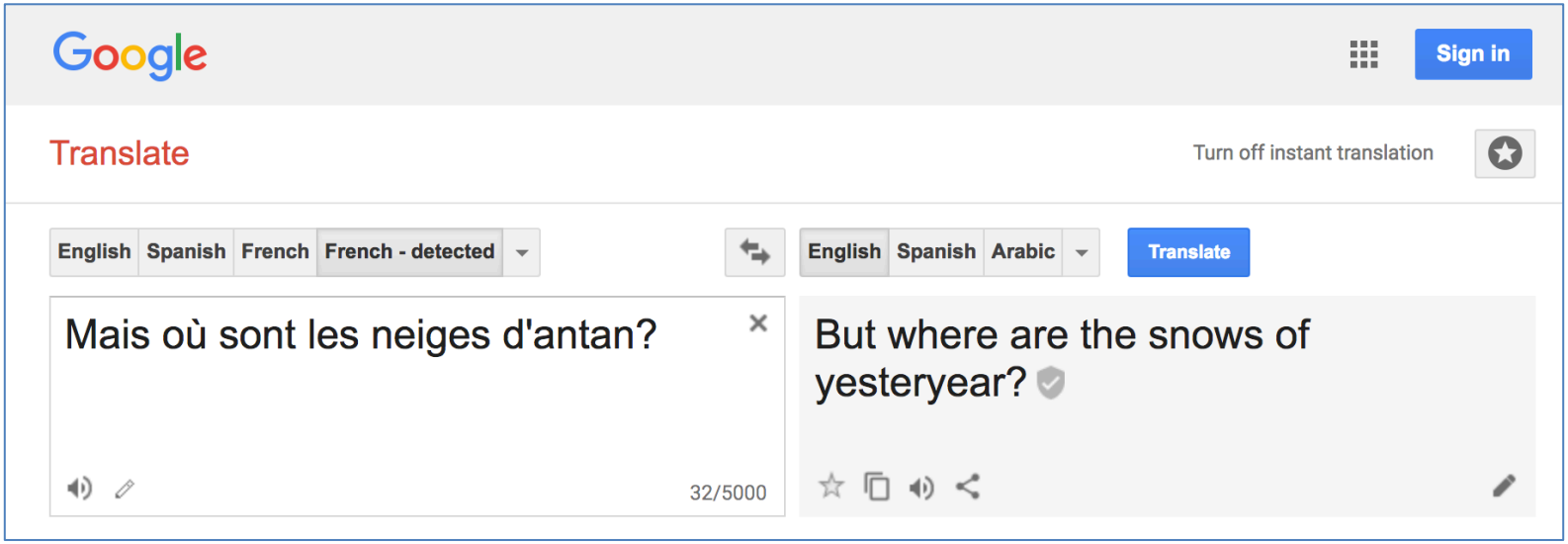
Unlike many of the reviewers on here, I don't have another DSLR camera to compare this one to. It's my first, but what I want to do is approach this review from the pov of a first-time DLSR user. I kinda have no choice, eh? But what I mean is, instead of going on about specifics, I am looking at how easy was it for a beginner like me to use this camera.

Size and weight:

- nice and compact to carry!
- since the camera is small and light, ...
- the camera feels flimsy, is plastic and very light in weight, ...

Machine Translation

- Herein of translating from one natural language to another
- Fully automatic systems



- Systems may also have humans in the loop

Language Technologies

mostly solved

making good progress

still really hard

Spam detection

Let's go to Agra

Buy VIAGRA

Part-of-speech (POS) tagging

NN NN MD VB

Time travel will work

Named entity recognition (NER)

PERSON ORG LOC

Einstein met with UN officials in Princeton

Sentiment analysis

Best pastrami in Orlando

The waiter ignored us

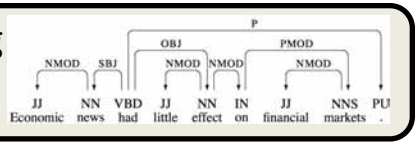
Coreference resolution

Carter told Mubarak he shouldn't run again.

Word sense disambiguation

I need a new battery for my ***mouse***

Parsing



Machine translation (MT)

Time flies ➡ 時光飛逝

Information extraction (IE)

You're invited to dinner on Friday, May 18th at 8:30 pm

Party May 18 add

Question answering (QA)

Q: How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

Summarization

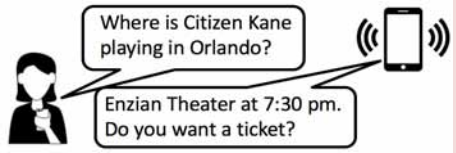
The Dow Jones is up

The S&P 500 jumped

Housing prices rose

➡ The economy is doing well

Dialog



Ambiguity makes NLP hard

- News story headlines
  - Hospitals Are Sued by 7 Foot Doctors
  - Enraged Cow Injures Farmer With Ax
  - Dealers Will Hear Car Talk at Noon
  - Miners Refuse to Work After Death
  - Drunk Gets Nine Months in Violin Case
  - Iraqi Head Seeks Arms
  - Kids Make Nutritious Snacks
  - Hershey Bars Protest

Q: What makes these headlines funny?

(source: http://www.fun-with-words.com/ambiguous_headlines.html)

Why else is natural language understanding difficult?

non-standard English

U kno u inspired me thru this whole process. How to be strong! I thank u bro. And bringgg it home for the LAND. I seee u 3 👌

segmentation issues

the New York-New Haven Railroad
the New York-New Haven Railroad

idioms

dark horse
get cold feet
lose face
throw in the towel

neologisms

unfriend
retweet
bromance

world knowledge

Mary and Sue are sisters
Mary and Sue are mothers

tricky entity names

Where is *A Bug's Life* playing ...
Let It Be was recorded ...
... a mutation on the *for* gene ...

Note: this is not an exclusive list

Making progress on the problem

- The task is difficult!
- Tools we need
  - Knowledge about language
  - Knowledge about the world
  - Mechanisms for combining knowledge sources
- How we generally do this
  - Probabilistic models built from language data
    - P("maison" ⟶ "house") = high
    - P("L'avocat général" ⟶ "the general avocado") = low
- Fortunately, rough text features can often do half the job

Today

- Course logistics
- What is natural language processing?
- What can natural language processing do?
- What this course will cover
- Action items for this week

What this course will cover

- Text processing basics
- Text classification and naive Bayes classification
- Logistic regression
- Sequence labeling (POS tagging and NER)
- Hidden Markov models
- Neural networks and backpropagation
- Constituency and dependency parsing
- Question answering
- Summarization
- Advanced topics: recurrent and convolutional networks, deep learning

Skills you will need

- Simple linear algebra (vectors, matrices)
- Basic probability theory
- Good programmiing skills

Today

- Course logistics
- What is natural language processing?
- What can natural language processing do?
- What this course will cover
- Action items for this week

Action Items for this week

- Engagement assignment
  - due 1/12 at 11:59 pm, or as soon as join class
- Make sure you have a solid software development environment
- Consider installing
  - Python and NLTK (highly recommended)
    - use online NLTK book as tutorial
  - Keras – open source Python neural network library
    - needs TensorFlow, CNTK, or Theano