Introduction to NLP

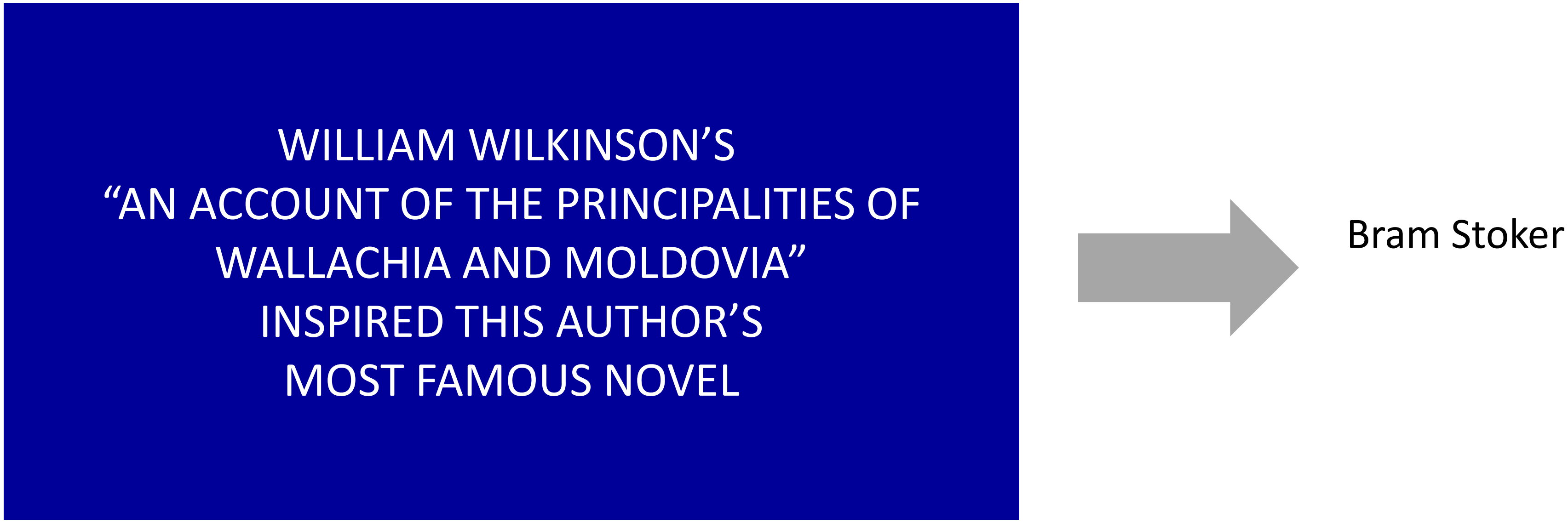
What is Natural Language Processing?

Natural Language Processing

- how to program computers to *process* and analyze large amounts of *natural language* data

Question Answering: IBM’s Watson

- Won Jeopardy on February 16, 2011!



WILLIAM WILKINSON’S
“AN ACCOUNT OF THE PRINCIPALITIES OF
WALLACHIA AND MOLDOVIA”
INSPIRED THIS AUTHOR’S
MOST FAMOUS NOVEL

Bram Stoker

Information Extraction

Subject: **curriculum meeting**

Date: January 15, 2012

To: Dan Jur

Event: Curriculum mtg
Date: Jan-16-2012
Start: 10:00am
End: 11:30am
Where: Gates 159

Hi Dan, we've now scheduled the curriculum meeting.

It will be in Gates 159 tomorrow from 10:00-11:30.

-Chris

Create new Calendar entry

Information Extraction & Sentiment Analysis

Attributes:

zoom

affordability

size and weight

flash

ease of use

Size and weight

- ✓ nice and compact to carry!
- ✓ since the camera is small and ed to carry around those heavy, b al cameras either!
- ✗ the camera feels flimsy, is plastic and very light in weight you have to be very delicate in the handling of this camera

Machine Translation

- Fully automatic

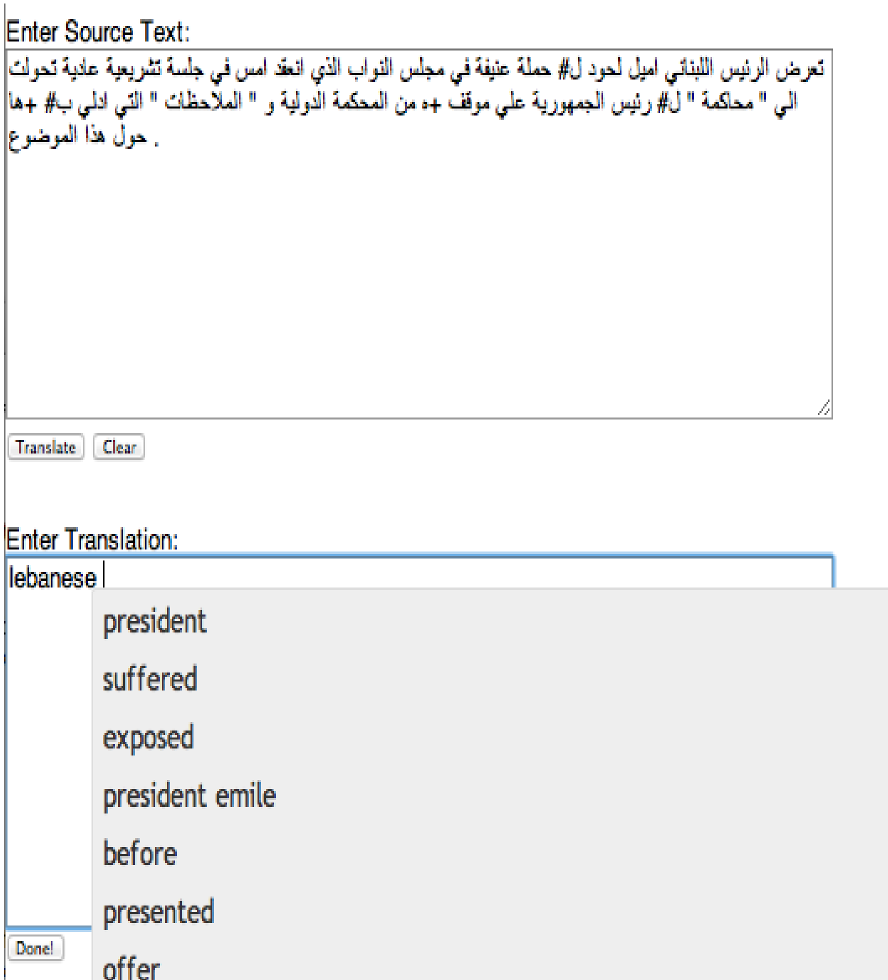
Enter Source Text:

这 不过 是 一 个 时间 的 问题 .

Translation from Stanford’s *Phrasal*:

This is only a matter of time.

- Helping human translators



Enter Source Text:

تعرض الرئيس اللبناني اميل لحود ل# حملة عنيفة في مجلس النواب الذي انعقد امس في جلسة تشريعية عادية تحولت
الي " محاكمة " ل# رئيس الجمهورية علي موقف +ه من المحكمة الدولية و " الملاحظات " التي ادلي ب# +ها
حول هذا الموضوع .

Enter Translation:

lebanese

president
suffered
exposed
president emile
before
presented
offer

Language Technology

making good progress

mostly solved

Spam detection

Let's go to Agra! ✓

Buy Fake Painiting … ✗

Part-of-speech (POS) tagging

ADJ	ADJ	NOUN	VERB	ADV
Colorless	green	ideas	sleep	furiously.

Named entity recognition (NER)

PERSON			ORG		LOC
Einstein	met	with	UN	officials in	Princeton

Sentiment analysis

Best roast chicken in San Francisco! 👍

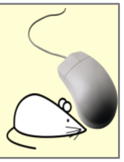
The waiter ignored us for 20 minutes. 👎

Coreference resolution

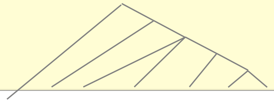
Carter told Mubarak he shouldn't run again.

Word sense disambiguation (WSD)

I need new batteries for my ***mouse***.



Parsing



I can see Alcatraz from the window!

Machine translation (MT)

第13届上海国际电影节开幕… ⇨

The 13th Shanghai International Film Festival…

Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30

Party
May 27
add

still really hard

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&P500 jumped

Housing prices rose

⇨

Economy is good

Dialog

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?

100% REAL

Ambiguity makes NLP hard: “Crash blossoms”

Violinist Linked to JAL Crash Blossoms

[Violinist Linked to JAL Crash] Blossoms

Main Verb

[Violinist] Linked to [JAL Crash Blossoms]

Teacher Strikes Idle Kids

[Teacher Strikes] Idle [Kids]

[Teacher] Strikes [Idle Kids]

Ambiguity makes NLP hard:

100% REAL

excessive
bureaucracy

1. Delay
2. support

Red Tape Holds Up New Bridges

Hospitals Are Sued by 7 Foot Doctors

Hospitals Are Sued by [7 Foot] Doctors

Hospitals Are Sued by 7 [Foot Doctors]

100% REAL

Ambiguity makes NLP hard: “Crash blossoms”

Juvenile Court to Try Shooting Defendant

Juvenile Court to Try [Shooting Defendant]

Juvenile Court to Try Shooting [Defendant]

Local High School Dropouts Cut in Half

Ambiguity is pervasive

New York Times headline (17 May 2000)

Fed raises interest rates

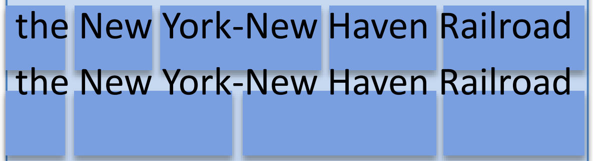
Fed raises interest rates 0.5%

Why else is natural language understanding difficult?

non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either♥

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 …

But that's what makes it fun!

Making progress on this problem…

- The task is difficult! What tools do we need?
  - Knowledge about language
  - Knowledge about the world
  - A way to combine knowledge sources
- How we generally do this:
  - probabilistic models built from language data
    - P("L'avocat général" → "the general avocado") low
    - P("L'avocat général" → "the general advocate") high

Avocado and advocate both words translate to avocat in French

  - Luckily, rough text features can often do half the job.

Top Conferences and Journals in NLP Field

Conferences

- ACL(The Association for Computational Linguistics) https://www.aclweb.org
- HLT-NAACL
- EMNLP
- COLING

Journals

- ACM Transactions on Speech and Language Processing
- Computational Linguistics
- International Journal of Computational Linguistics
- Information Processing and Management (Journal)
- Knowledge and Data Engineering (Journal)
- Information Science (Journal)
- Knowledge Based systems (Journal)

NLP-related Resources

- NLTK(Python-based)
- Keras, TensorFlow (Deep Learning)
- Statistical natural language processing and corpus-based computational linguistics: An annotated list of resources
- Stanford NLP parser (Stanford University NLP group)
- OpenNLP (Apache)
- LingPipe (Jave-based)

Machine Learning Toolkits

- Weka (A rich collection of machine learning algorithms, Machine Learning Group at the University of Waikato)
- Mallet (An alternative package for Weka, developed by Andrew McCallum at University of Massachusetts Amherst)
- LibSVM (A collection of SVMs, developed by Chih-Chung Chang and Chih-Jen Lin at National Taiwan University)
- SVM-light (Another collection of SVMs, developed by Thorsten Joachims at Cornell University)
- GraphLab (Large-scale machine learning package)
- mahout (Apache large-scale machine learning package)
- Topic Models (David Blei's collection of various topic models)

Course Outline