BBM 495
INTRODUCTION TO NATURAL LANGUAGE PROCESSING (NLP)

LECTURER: BURCU CAN

2019-2020 SPRING

Natural language

- Languages that are used by human beings are called ‘natural languages’.

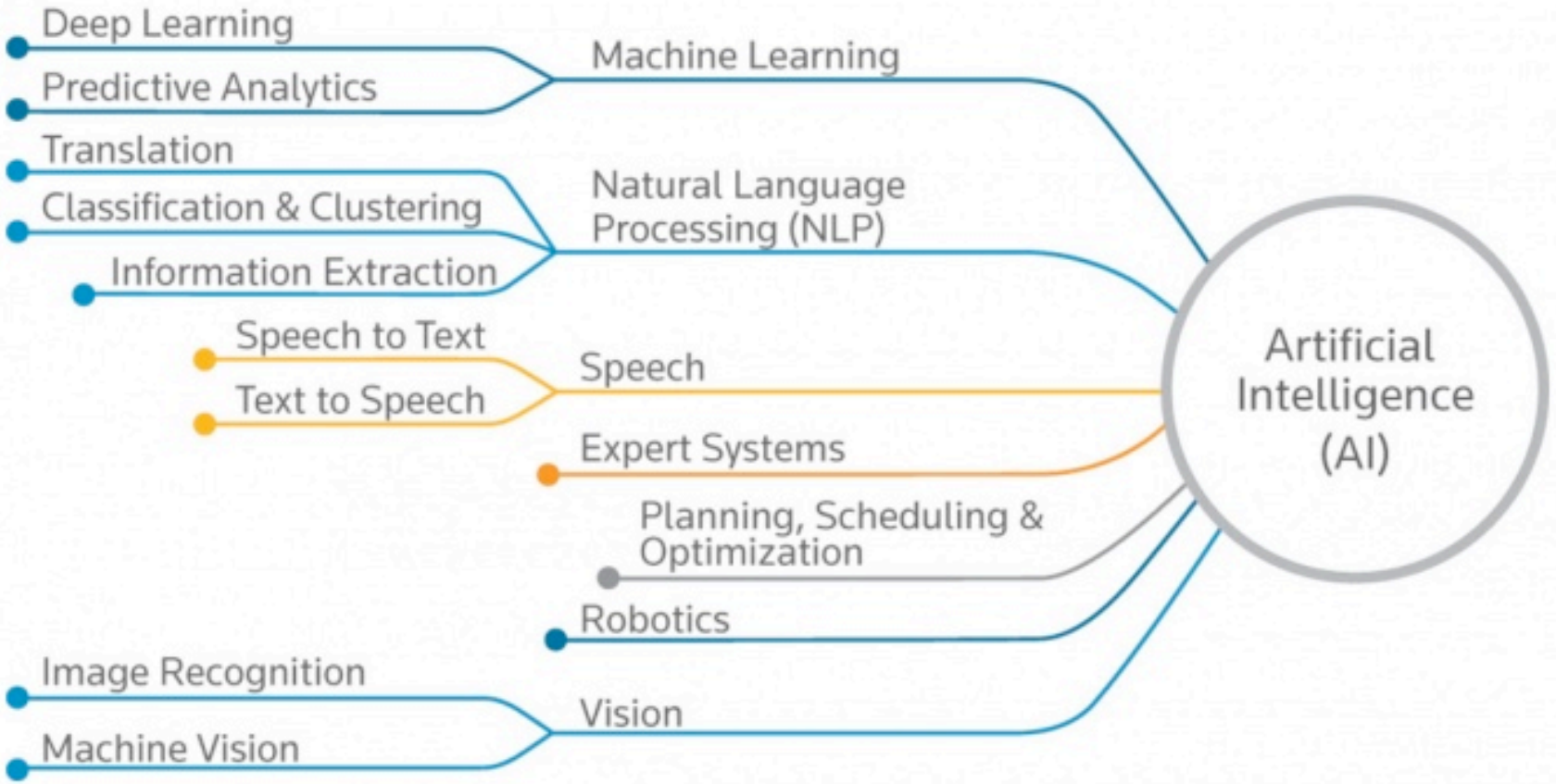
- What languages do you know apart from the natural languages?

Natural Language Processing

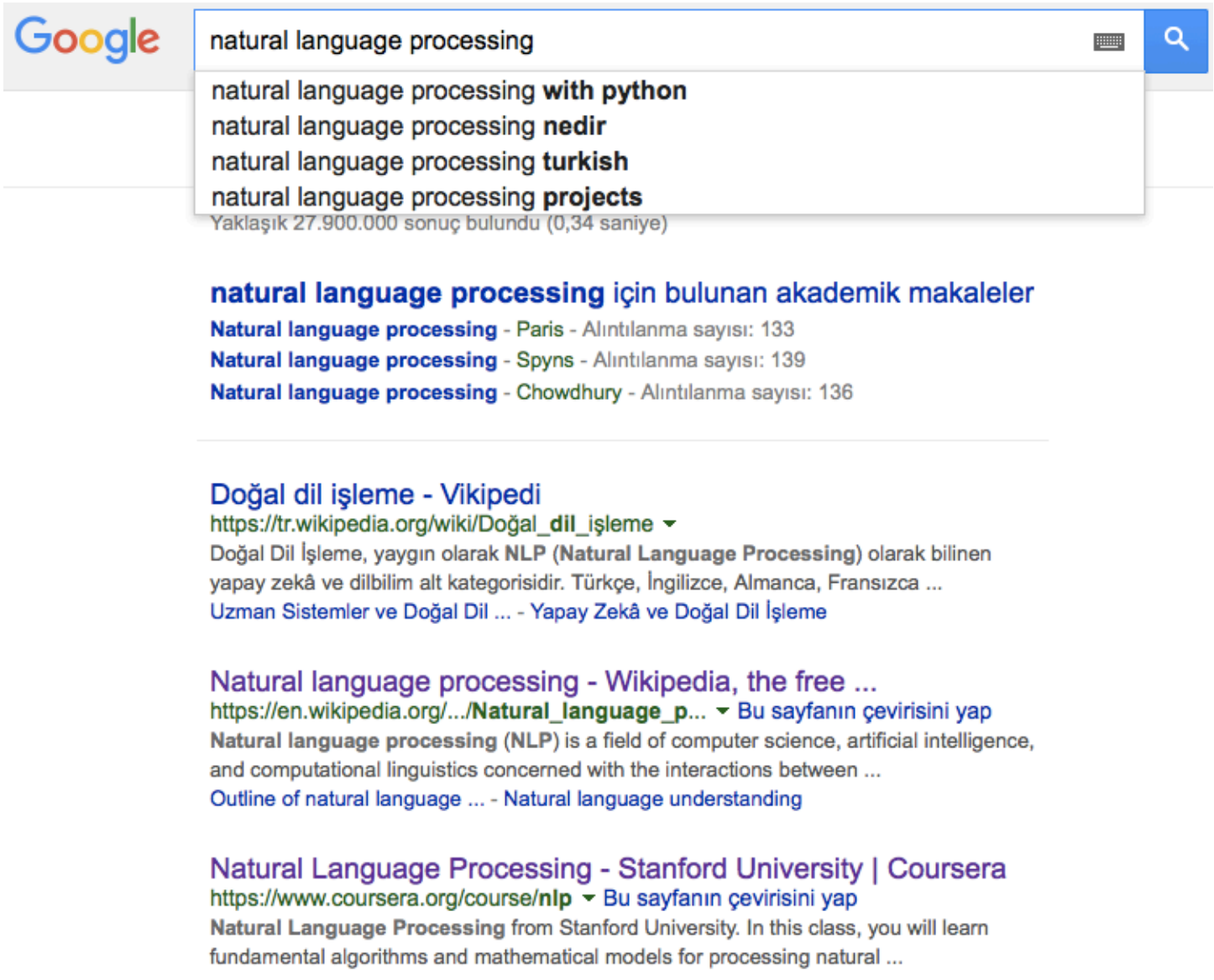
- Natural language processing (NLP) is a field of computer science and artificial intelligence, and it is concerned with the interactions between computers and human (natural) languages. (Wikipedia)

- **What kinds of tasks do computers perform on human languages?**

NLP as an Artificial Intelligence Field



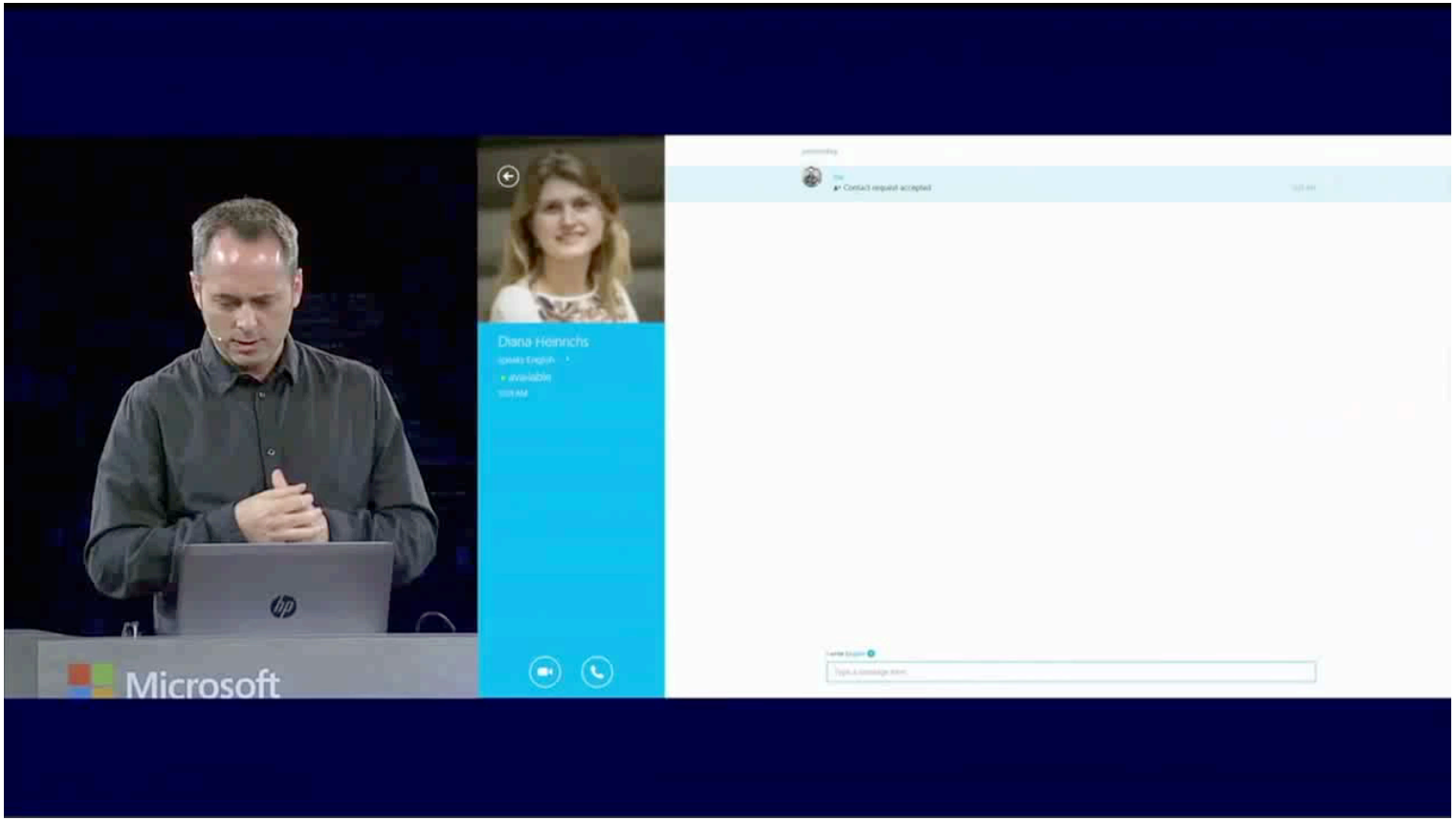
When you google sth.



Machine translation

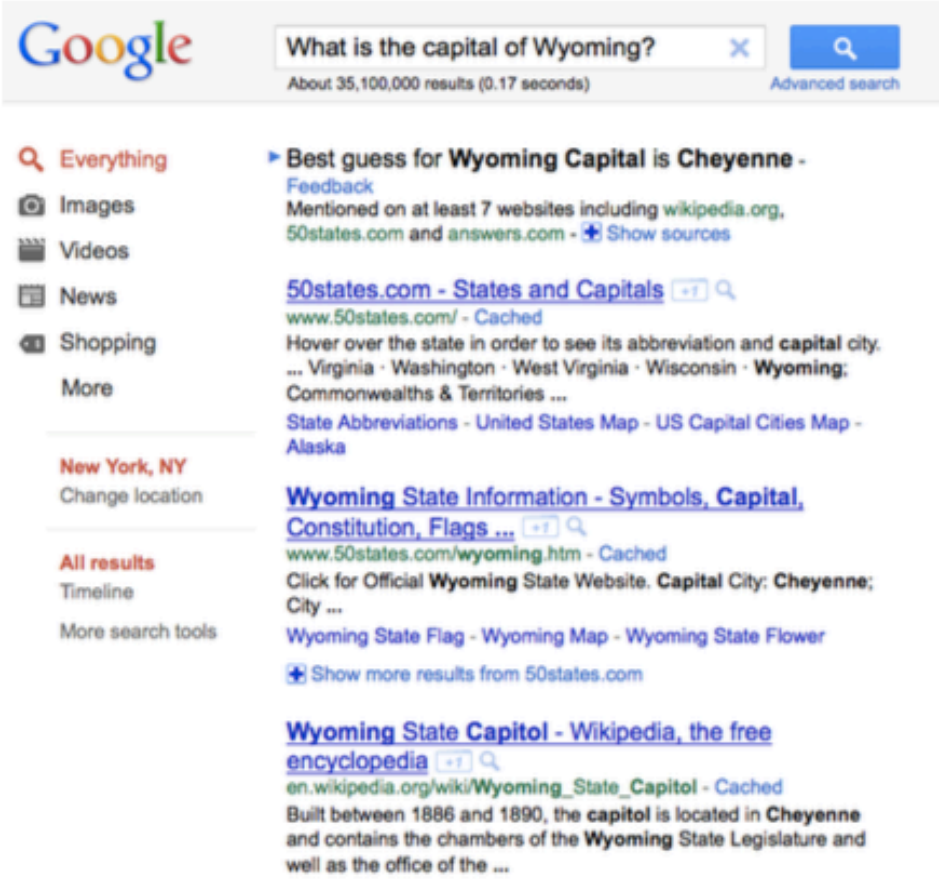
- Translate text from one language to another
- Recombines fragments of example translations
- Challenges:
  - What fragments? [learning to translate]
  - How to make efficient? [fast translation search]

Skype translator



Question Answering

- Question Answering:
  - More than search
  - Ask general comprehension questions of a document collection
  - Can be really easy: "What's the capital of Wyoming?"
  - Can be harder: "How many US states' capitals are also their largest cities?"
  - Can be open ended: "What are the main issues in the global warming debate?"
- SOTA: Can do factoids, even when text isn't a perfect match



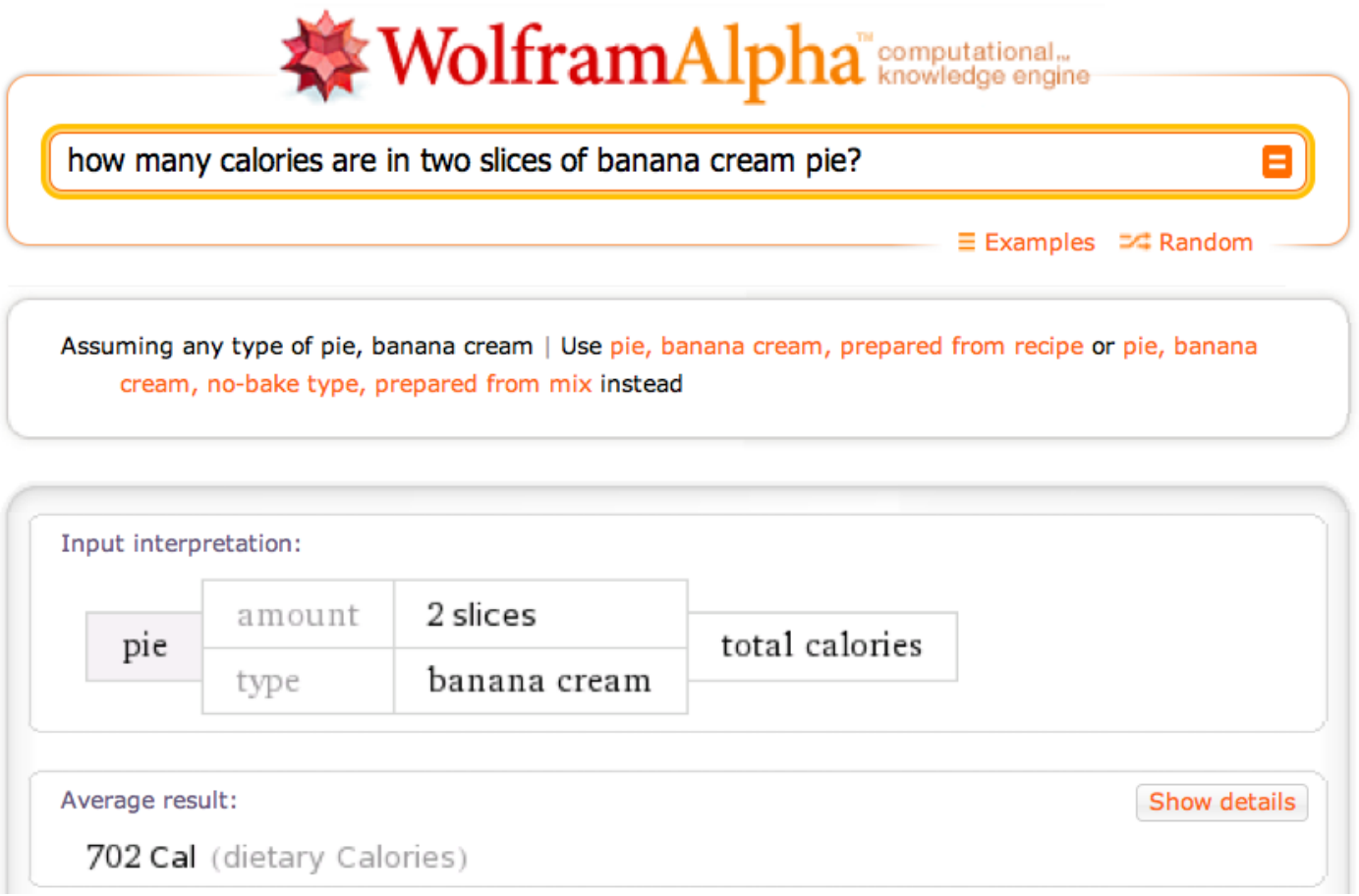
Question Answering: IBM's Watson

- Won Jeopardy on February 16, 2011!

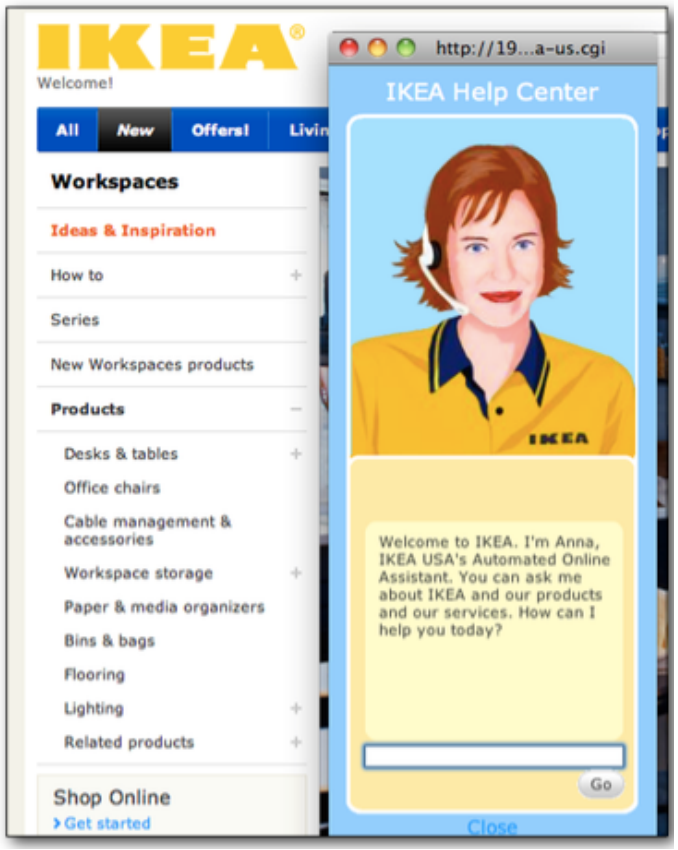
WILLIAM WILKINSON'S
"AN ACCOUNT OF THE PRINCIPALITIES OF WALLACHIA AND MOLDOVIA"
INSPIRED THIS AUTHOR'S
MOST FAMOUS NOVEL

Bram Stoker

Question Answering:



Dialog systems



Eliza

(computer program)

- ELIZA is rule-based. It has gained popularity in the area of psychotherapy.
  - ELIZA was written at MIT by Joseph Weizenbaum between 1964 and 1966.

```
 USER: Men are all alike
ELIZA: In what way?
 USER: They're always bugging us about something or other
ELIZA: Can you think of a specific example?
 USER: My boyfriend made me come here
ELIZA: Your boyfriend made you come here
 USER: He says I'm depressed much of the time
ELIZA: I'm sorry to hear you are depressed
...
```

Text Classification

- Male or female author:

1. By 1925 present-day Vietnam was divided into three parts under French colonial rule. The southern region embracing Saigon and the Mekong delta was the colony of Cochin-China; the central area with its imperial capital at Hue was the protectorate of Annam…
2. Clara never failed to be astonished by the extraordinary felicity of her own name. She found it hard to trust herself to the mercy of fate, which had managed over the years to convert her greatest shame into one of her greatest assets…

Argamon, M. Koppel, J. Fine, A. R. Shimoni, 2003. "Gender, Genre, and Writing Style in Formal Written Texts," Text, volume 23, number 3, pp. 1–346

Text Classification

- Positive or negative movie review

- unbelievably disappointing

- Full of zany characters and richly applied satire, and some great plot twists

- this is the greatest screwball comedy ever filmed

- It was pathetic. The worst part about it was the boxing scenes.

Sentiment analysis

- Extraction of emotions from a text.

Source:
www.amazon.com

Sentiment analysis

Sentiment Analysis – Definition

"Sentiment analysis is the task of identifying positive and negative opinions, emotions and evaluations in text"

Text Classification

Subject: **Important notice!**
From: Stanford University <newsforum@stanford.edu>
Date: October 28, 2011 12:34:16 PM PDT
To: undisclosed-recipients:;

Greats News!

You can now access the latest news by using the link below to login to Stanford University News Forum.

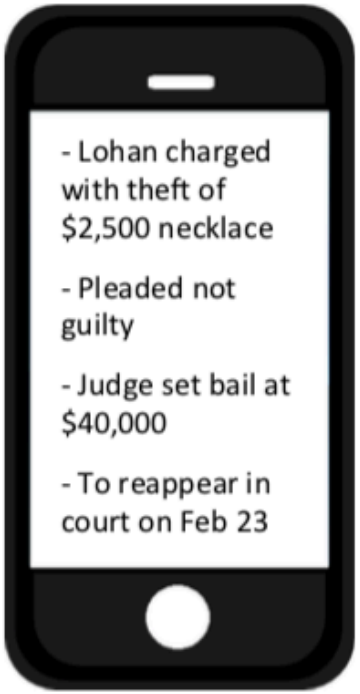
http://www.123contactform.com/contact-form-StanfordNew1-236335.html

Click on the above link to login for more information about this new exciting forum. You can also copy the above link to your browser bar and login for more information
about the new services.

Text summarization

- Generation of a small text from a given long document:
  - Article: With a split decision in the final two primaries and a flurry of superdelegate endorsements, Sen. Barack Obama sealed the Democratic presidential nomination last night after a grueling and history-making campaign against Sen. Hillary Rodham Clinton that will make him the first African American to head a major-party ticket. Before a chanting and cheering audience in St. Paul, Minn., the first-term senator from Illinois savored what once seemed an unlikely outcome to the Democratic race with a nod to the marathon that was ending and to what will be another hard-fought battle, against Sen. John McCain, the presumptive Republican nominee….
  - Summary: Senator Barack Obama was declared the presumptive Democratic presidential nominee.

Text summarization



Information Extraction

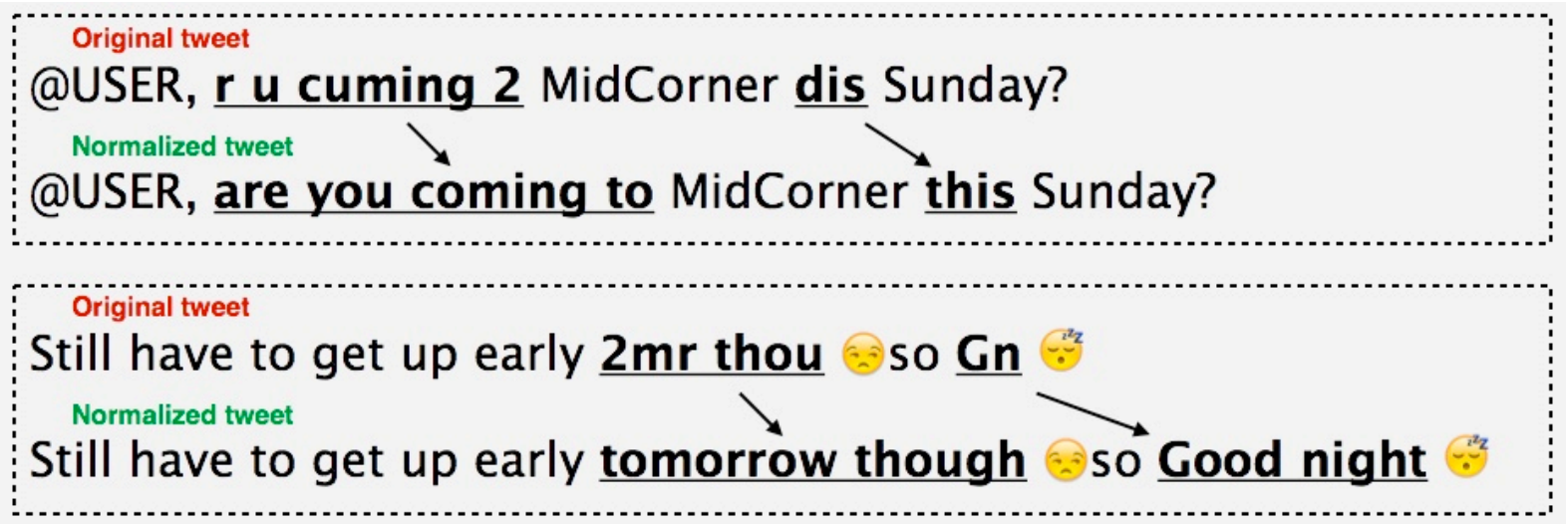
- Unstructured text to database entries

New York Times Co. named Russell T. Lewis, 45, president and general manager of its flagship New York Times newspaper, responsible for all business-side activities. He was executive vice president and deputy general manager. He succeeds Lance R. Primis, who in September was named president and chief operating officer of the parent.

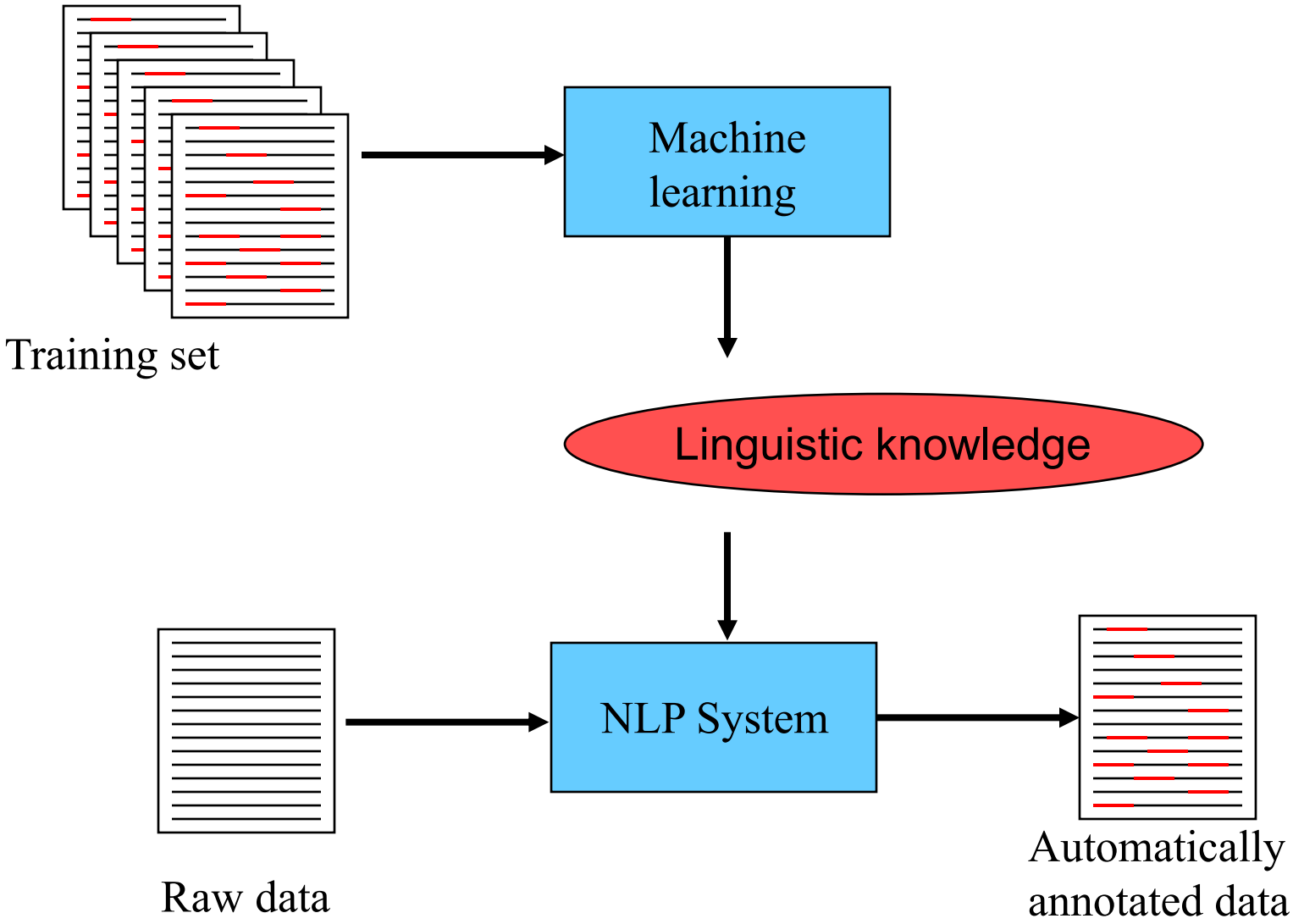
Person	Company	Post	State
Russell T. Lewis	New York Times newspaper	president and general manager	start
Russell T. Lewis	New York Times newspaper	executive vice president	end
Lance R. Primis	New York Times Co.	president and CEO	start

- SOTA: perhaps 80% accuracy for multi-sentence templates, 90% + for single easy fields
- But remember: information is redundant!

Text normalization



Learning


Machine learning
Training set
Linguistic knowledge
NLP System
Raw data
Automatically annotated data

What is NLP?

- Fundamental goal: *deep* understand of *broad* language
  - Not just string processing or keyword matching!
- End systems that we want to build:
  - Simple: spelling correction, text categorization…
  - Complex: speech recognition, machine translation, information extraction, dialog interfaces, question answering…
  - Unknown: human-level comprehension (is this just NLP?)

Language Comprehension

"The rock was still wet. The animal was glistening, like it was still swimming," recalls Hou Xianguang. Hou discovered the unusual fossil while surveying rocks as a paleontology graduate student in 1984, near the Chinese town of Chengjiang. "My teachers always talked about the Burgess Shale animals. It looked like one of them. My hands began to shake." Hou had indeed found a Naraoia like those from Canada. However, Hou's animal was 15 million years older than its Canadian relatives.

It can be inferred that Hou Xianguang's "hands began to shake", because he was:

(A) afraid that he might lose the fossil
(B) worried about the implications of his finding
(C) concerned that he might not get credit for his work
(D) uncertain about the authenticity of the fossil
(E) excited about the magnitude of his discovery

Why Study NLP?

- Text is the largest repository of human knowledge and is growing quickly.
  - emails, news articles, web pages, IM, scientific articles, insurance claims, customer complaint letters, transcripts of phone calls, technical documents, government documents, patent portfolios, court decisions, contracts, ……
  - Are we reading any faster than before?

IT IS THE LARGEST environmental enforcement recovery by the Department of Justice, exceeding even the $4bn paid by BP in 2012 to resolve criminal proceedings over the 2010 Deepwater Horizon spill. Police at the Ministry of Public Security yesterday said 46 suspects at GSK's Chinese subsidiary had been identified as part of a "complete bribery chain" that funnelled money to hospitals, doctors and government officials between 2009 and 2012. Mark Reilly, a Briton who was head of the unit, ordered subordinates to offer the illegal payments, they said. The allegations will almost certainly lead to charges, which could strain China's relations with Britain. David Cameron, UK prime minister, had sought to any to the London-based company. Regulators in the US and UK are closely monitoring events in China and could launch their own investigations, although legal experts said they were likely to wait for Beijing to complete its probe first. A senior official at the Chinese ministry's economic crime investigation bureau said "all departments" of GSK's China subsidiary were "fully engaged" in the alleged bribery. The company paid for the bribes by inflating the price of medicines in China, some of which were seven times more expensive than in other markets, the official said. According to the official, the company set up several internal units with code names such as "operation Great Wall" and "operation soaring dragon" to pay the bribes. The developments will send a chill through the international business community. So far very few expatriates have faced corruption charges. GSK has previously denied any systemic corruption. In a statement yesterday it said: "We take the allegations that have been raised veas head of the unit, ordered subordinates to offer the illegal payments, they said. The allegations will almost certainly lead to charges, which could strain China's relations with Britain. David Cameron, UK prime minister, had sought to limit any damage to the London-based company. Affiliated Managers Group could hardly resemble US fund sales titan Vanguard less, except in one respect: no two US complexes have less experience of investor outflows. AMG, a holding company for boutique asset managers, has posted positive sales for 16 straight quarters. Unlike Vanguard, however, AMG specialises in higher-priced actively managed equities and alternative investments, driving average annual profit growth topping 30 per cent for the past four years. Though the bulk of AMG's $594bn in assets are from institutional clients outside the US, it has made US retail growth a priority. It rolled out the AMG Funds retail brand last month, hired a US retail sales chief and prepared to advertise. US retail may not seem its most logical market, given that Vanguard's ruthless discounting and no-frills index products have long dominated, but Sean Healey, chief executive of AMG, says demand for boutique investing is building. "We don't need to convince anyone that passive is going away," Mr Healey says. "Rather, we need to convince investors that we are on the other end, on the alpha-generating end of the barbell." The long-anticipated rotation out of fixed income and general risk aversion bodes well for AMG,

Twitter Usage Statistics

675,886,424

Tweets sent today

Microsoft TweetBot

NLP can do...

- Many companies would make a lot of money if they could use computer programmes that understood text or speech. Just imagine if a computer could be used for:
  - answering the phone, and replying to a question
  - understanding the text on a Web page to decide who it might be of interest to
  - translating a daily newspaper from Japanese to English
  - understanding text in journals / books and building an expert system based on that understanding

NLP applications (cont')

- Information Retrieval
- Information Extraction
- Document Classification
- Automatic Summarization
- Text Proofreading – Spelling & Grammar
- Machine Translation
- Plagiarism detection

- Can you think of anything else???

New Trends

Contextualized Sarcasm Detection on Twitter

David Bamman and Noah A. Smith
School of Computer Science
Carnegie Mellon University
{dbamman,nasmith}@cs.cmu.edu

Abstract

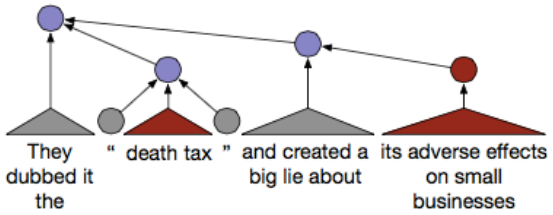
Sarcasm requires some shared knowledge between speaker and audience; it is a profoundly *contextual* phenomenon. Most computational approaches to sarcasm detection, however, treat it as a purely linguistic matter, using information such as lexical cues and their corresponding sentiment as predictive features. We show that by including extra-linguistic information from the context of an utterance on Twitter – such as properties of the

Political Ideology Detection Using Recursive Neural Networks

Mohit Iyyer[1], Peter Enns[2], Jordan Boyd-Graber[3,4], Philip Resnik[2,4]
[1]Computer Science, [2]Linguistics, [3]iSchool, and [4]UMIACS
University of Maryland
{miyyer,peter,jbg}@umiacs.umd.edu, resnik@umd.edu

Abstract

An individual's words often reveal their political ideology. Existing automated techniques to identify ideology from text focus on bags of words or wordlists, ignoring syntax. Taking inspiration from recent work in



What Yelp Fake Review Filter Might Be Doing?

Arjun Mukherjee[†] Vivek Venkataraman[†] Bing Liu[†] Natalie Glance[‡]
[†] University of Illinois at Chicago [‡] Google Inc.
arjun4787@gmail.com; {vvenka6, liub}@uic.edu; nglance@google.com

Abstract

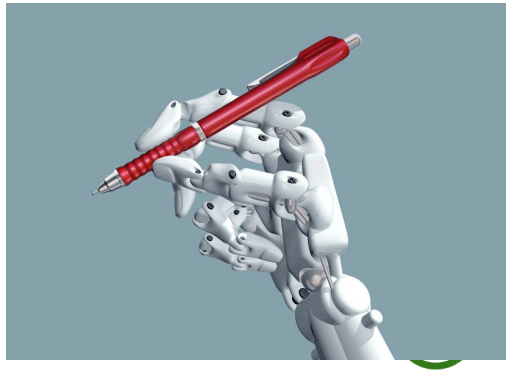
Online reviews have become a valuable resource for decision making. However, its usefulness brings forth a curse – *deceptive opinion spam*. In recent years, fake review detection has attracted significant attention. However, most review sites still do not publicly filter fake reviews. Yelp is an exception which has been filtering reviews over the past

deceptive fake reviews to promote or to discredit some target products and services. Such individuals are called *opinion spammers* and their activities are called *opinion spamming* (Jindal and Liu 2008).

The problem of opinion spam or fake reviews has become widespread. Several high-profile cases have been reported in the news (Streitfeld, 2012a). Consumer sites

Robot Poet

People picking up electric chronic.
The balance like a giant tidal wave,
Never ever feeling supersonic,
Or reaching any very shallow grave.



Why is NLP hard?

- Language = Words + rules + exceptions..
- Ambiguity at all levels..
- We speak different languages..
- And language is a cultural entity..
- Highly systematic but also complex..
- Keeps changing.. New words, new rules and new exceptions..
- Source : Electronic texts / Printed texts / Acoustic Speech Signal.. they are noisy..
- Language looks obvious to us.. But it is a Big Deal ☺!

Course material

- Linguistic topics
  - morphology, syntax, semantics etc.
- Applications
  - spelling correction, machine translation, text classification etc.

- Building a computer that ‘understands’ text!!!

The NLP Pipeline

Levels of language analysis

- Phonology
- Morphology
- Syntax
- Semantics
- Pragmatics
- Discourse

Task: Segmentation

- Text is split into words and sentences.
  - Languages like Chinese do not have spaces between words.

original, un-segmented text

再往远些看，随着汉字识别和语音识别技术的发展，
中文计算机用户将跨越语言差异的鸿沟，
在录入上走向中西文求同的道路。

separated word entities after segmentation

再 往 远 些 看 ， 随着 汉字 识别 和 语音 识别 技术 的 发展 ，
中文 计算机 用户 将 跨越 语言 差异 的 鸿沟 ，
在 录入 上 走 向 中 西文 求 同 的 道路 。

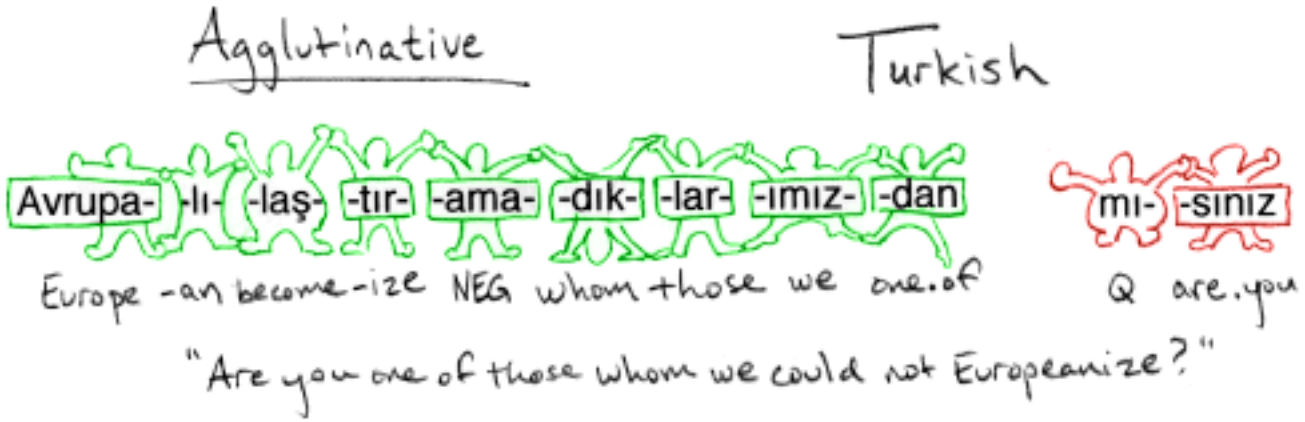
Task: Part-of-speech tagging

- Tagging the words in a sentence with their syntactic roles.

- John saw the saw and decided to take it to the table.
- PN V Det N Con V Part V Pro Prep Det N

Task: Morphological segmentation

- The longest word in Turkish:
  - *muvaffakiyetsizleştiricileştiriveremeyebileceklerimizdenmişsinizcesine*
  - *“As though you are from those whom we may not be able to easily make into a maker of unsuccessful ones”*



Ambiguity

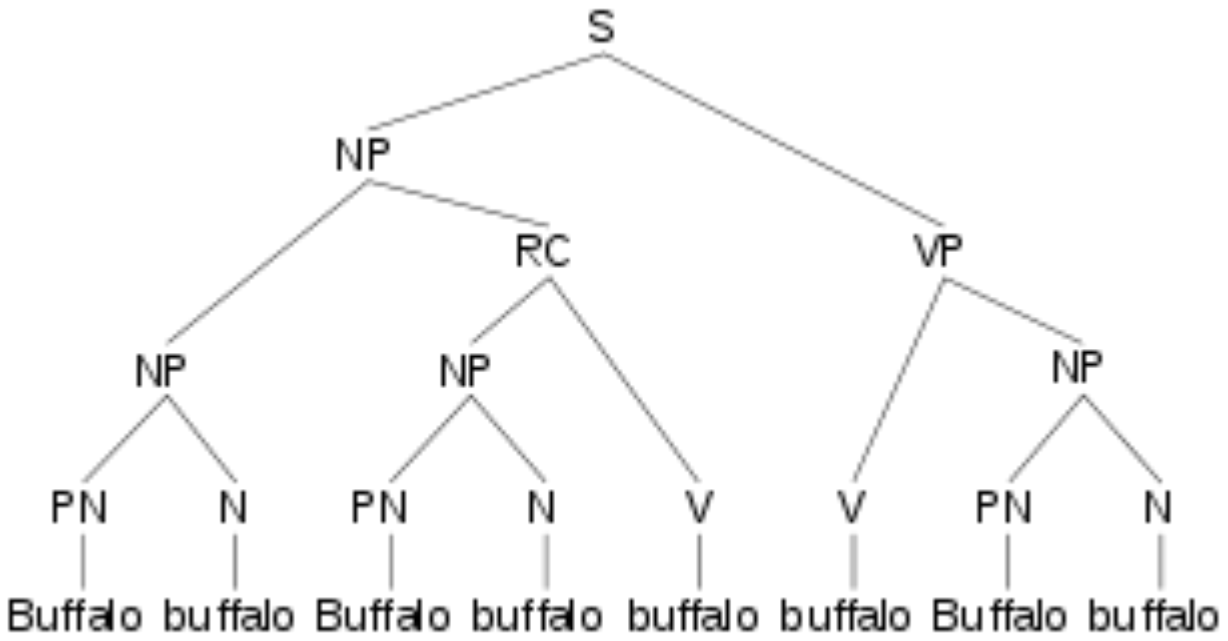
- Natural language is ambiguous!
  - In Turkish: *koyun*
  - In English: I can hear *bass* sounds.
    They like grilled *bass*.

Ambiguity

- "I saw the man with the telescope": 2 meanings
- "I saw the man on the hill with the telescope.": 5 meanings
- "I saw the man on the hill in Texas with the telescope": 14 meanings
- "I saw the man on the hill in Texas with the telescope at noon.": 42 meanings
- "I saw the man on the hill in Texas with the telescope at noon on Monday" 132 meanings

Syntactic parsing

- Örnek: Buffalo buffalo Buffalo buffalo buffalo buffalo Buffalo buffalo



- Diğer buffalo bizonlarının korkuttuğu bufalo bizonları, yine aynı buffalo bizonlarını korkutmaktadır

Semantic Disambiguation Example: "with"

Sentence	Relation
I ate spaghetti with meatballs.	(ingredient of spaghetti)
I ate spaghetti with salad.	(side dish of spaghetti)
I ate spaghetti with abandon.	(manner of eating)
I ate spaghetti with a fork.	(instrument of eating)
I ate spaghetti with a friend.	(accompanier of eating)

Pragmatics

- Uses context of utterance
  - Where, by who, to whom, why, when it was said
  - Intentions: *inform, request, promise*, *criticize,* …
- Handling Pronouns
  - "Mary eats apples. She likes them."
    - She="Mary", them="apples".
- Handling ambiguity
  - Pragmatic ambiguity: "you're late": What's the speaker's intention: informing or criticizing?
- "I saw the man with binoculars"

Discourse Analysis

Text units beyond sentences — examples

- A story (such as a fairy tale, a drama, ...).
- A news item.
- Dialogue.
- Technical text (manual, textbook, documentation).
- A document in a document base (abstract, patent description, ...).

Links between sentences/phrases in a larger text

- Textual ordering.
- Temporal link (for example, an event precedes another event).
  - *Jim saw the bus. He ran to catch it.*
    - "saw" precedes "ran"

Why Study NLP?

- To get a job in industry:
  - Many current job listings are NLP jobs
    - e.g. Facebook, Google, Amazon etc.
    - And many other technology companies in Turkey!

- To get a job in academia
  - As a computational linguist

About the class…

About the class…

- … Lectures:

  Thursdays, 9:00-11:45, D-10

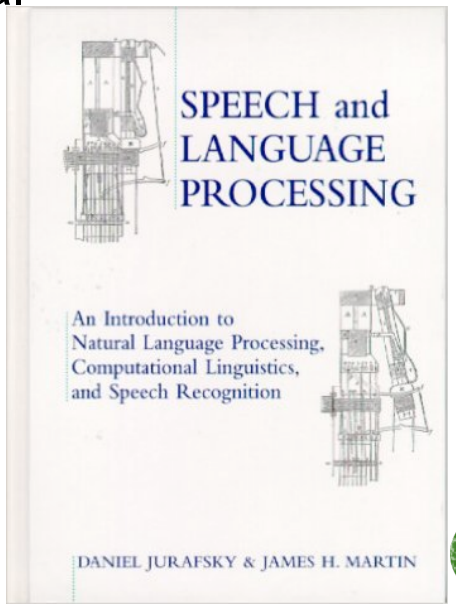
- … E-mail:
- burcucan@cs.hacettepe.edu.tr
- … Website:
- http://web.cs.hacettepe.edu.tr/~bbm495/
- … Readings:
- Textbook + additional readings (on website)
- … Assessment:
- Two midterms, one final exam

- No NLP background needed but some machine learning knowledge will help you in the lectures.
- Statistical knowledge is required!
- Of course, coding experience is compulsory!

Reading

- The textbooks:
- Jurafsky and Martin, Speech and Language Processing
- Manning and Schütze, Foundations of Statistical Natural Language Processing
- Yoav Goldberg, Neural Network Methods in Natural Language Processing

- Other material:
- Posted on the course website



Contact

- My email is:
  - burcucan@cs.hacettepe.edu.tr
- For all announcements follow Piazza:
  - piazza.com/hacettepe.edu.tr/spring2020/bbm495
  - piazza.com/hacettepe.edu.tr/spring2020/bbm497
- For the lecture slides, assignments, and important dates
  - http://web.cs.hacettepe.edu.tr/~burcucan/BBM495.htm

Exams

- What?
- -1st Midterm: 26th March
- - 2nd Midterm: 14th May
- - Final exam: To be announced later.
- Why?
- -To make sure you understand what you learned well enough to explain and apply it.
- How?
- - Closed-book
  - Will be based on lectures

Grading

- 1st Midterm: %30
- 2nd Midterm: %25
- Final Exam: %40
- Attendance: %5

Course outline (tentative)

Week	*Date*	*Topic*
1	Feb 27	Introduction to NLP
2	Mar 5	Language Models and N-grams, A Revision of Probability Theory
3	Mar 12	Regular Expressions, FSAs, FSTs
4	Mar 19	HMMs and PoS tagging
5	Mar 26	Midterm 1
6	Apr 2	Context Free Grammars and Parsing
7	Apr 9	Lexical Semantics and Word Sense Disambiguation
8	April 16	Introduction to Deep Learning
9	Apr 23	Official holiday
10	Apr 30	Word Embeddings
11	May 7	Recurrent Neural Networks
12	May 14	Midterm 2
13	May 21	Encoder-Decoder Models and Machine Translation
14	May 28	Advanced Neural Models

BBM 497
Introduction to NLP LAB.

- What?
  - 4 assignments (all programming) %50
  - Project %40
  - Quizzes %10
- Why?
- - To make sure you can put what you've learned to practice.

- How?
  - You will have 2-3 weeks to complete each assignment.
- - Grades will be based on your write-up, code, and results.

"My mom used to say 'If you can't say anything nice, don't say anything at all.' And that's how I got into sign language."

End of the first lecture…