

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?

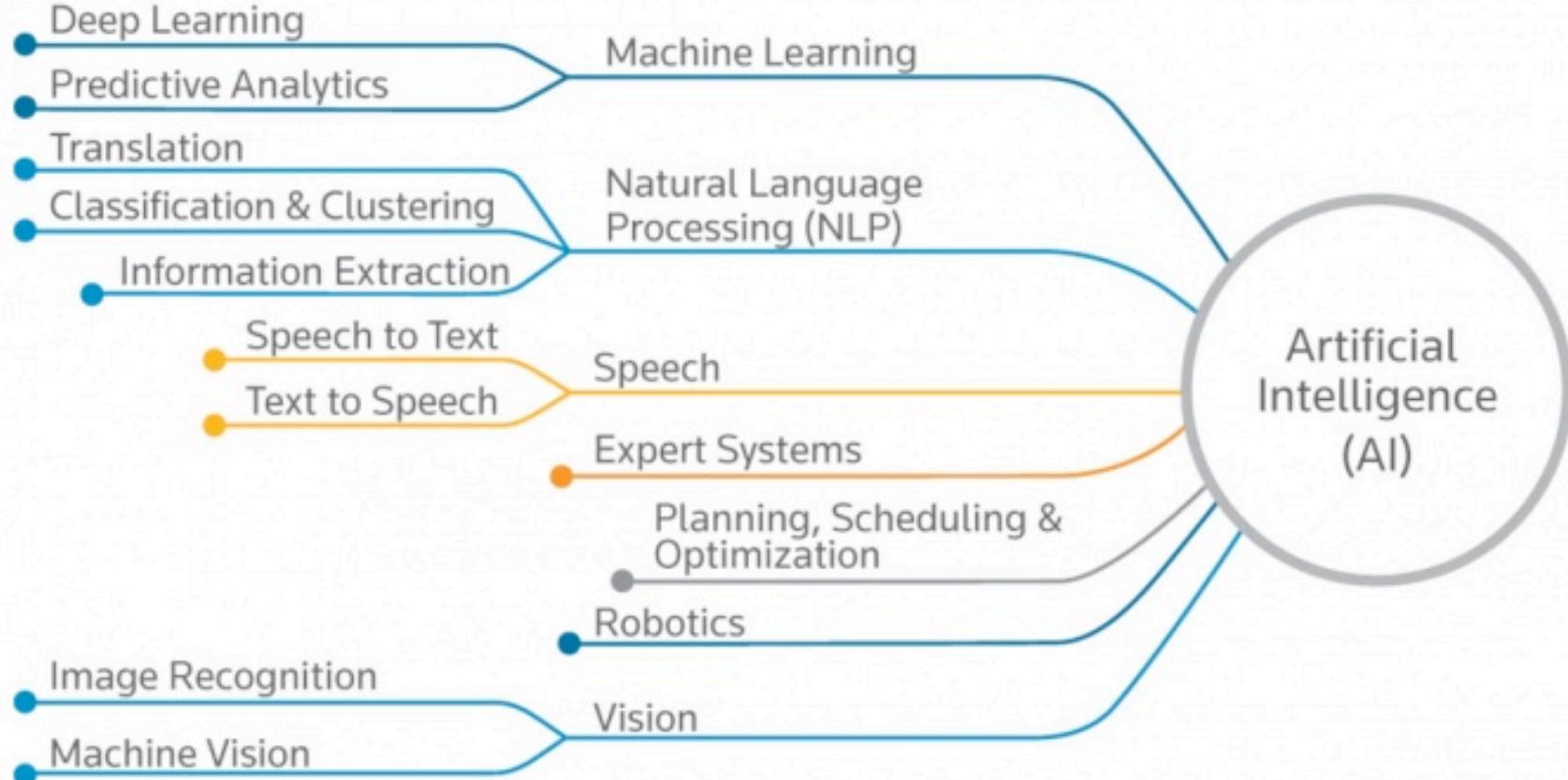


Natural Language Processing

- 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.

A screenshot of a Google search results page. The search bar at the top contains the query "natural language processing". Below the search bar is a dropdown menu displaying several search suggestions: "natural language processing with python", "natural language processing nedir", "natural language processing turkish", and "natural language processing projects". At the bottom of the suggestions list, it says "Yaklaşık 27.900.000 sonuç bulundu (0,34 saniye)".

natural language processing için bulunan akademik makaleler

[Natural language processing - Paris](#) - Alıntılanma sayısı: 133

[Natural language processing - Spyns](#) - Alıntılanma sayısı: 139

[Natural language processing - Chowdhury](#) - Alıntılanma sayısı: 136

Doğal dil işleme - Vikipedi

https://tr.wikipedia.org/wiki/Do%C4%91gal_dil_i%C5%9fleme ▾

Doğal Dil İşleme, yaygın olarak NLP (Natural Language Processing) olarak bilinen yapay zekâ ve dilbilim alt kategorisidir. Türkçe, İngilizce, Almanca, Fransızca ...

[Uzman Sistemler ve Doğal Dil ... - Yapay Zekâ ve Doğal Dil İşleme](#)

Natural language processing - Wikipedia, the free ...

https://en.wikipedia.org/.../Natural_language_p... ▾ Bu sayfanın çevirisini yap

Natural language processing (NLP) is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between ...

[Outline of natural language ... - Natural language understanding](#)

Natural Language Processing - Stanford University | Coursera

<https://www.coursera.org/course/nlp> ▾ Bu sayfanın çevirisini yap

Natural Language Processing from Stanford University. In this class, you will learn fundamental algorithms and mathematical models for processing natural ...



Machine translation

"Il est impossible aux journalistes de rentrer dans les régions tibétaines"

Bruno Philip, correspondant du "Monde" en Chine, estime que les journalistes de l'AFP qui ont été expulsés de la province tibétaine du Qinghai "n'étaient pas dans l'illégalité".

Les faits Le dalaï-lama dénonce l'"enfer" imposé au Tibet depuis sa fuite, en 1959
Vidéo Anniversaire de la rébellion tibétaine : la Chine sous ses meubles



"It is impossible for journalists to enter Tibetan areas"

Philip Bruno, correspondent for "World" in China, said that journalists of the AFP who have been deported from the Tibetan province of Qinghai "were not illegal."

Facts The Dalai Lama denounces the "hell" imposed since he fled Tibet in 1959

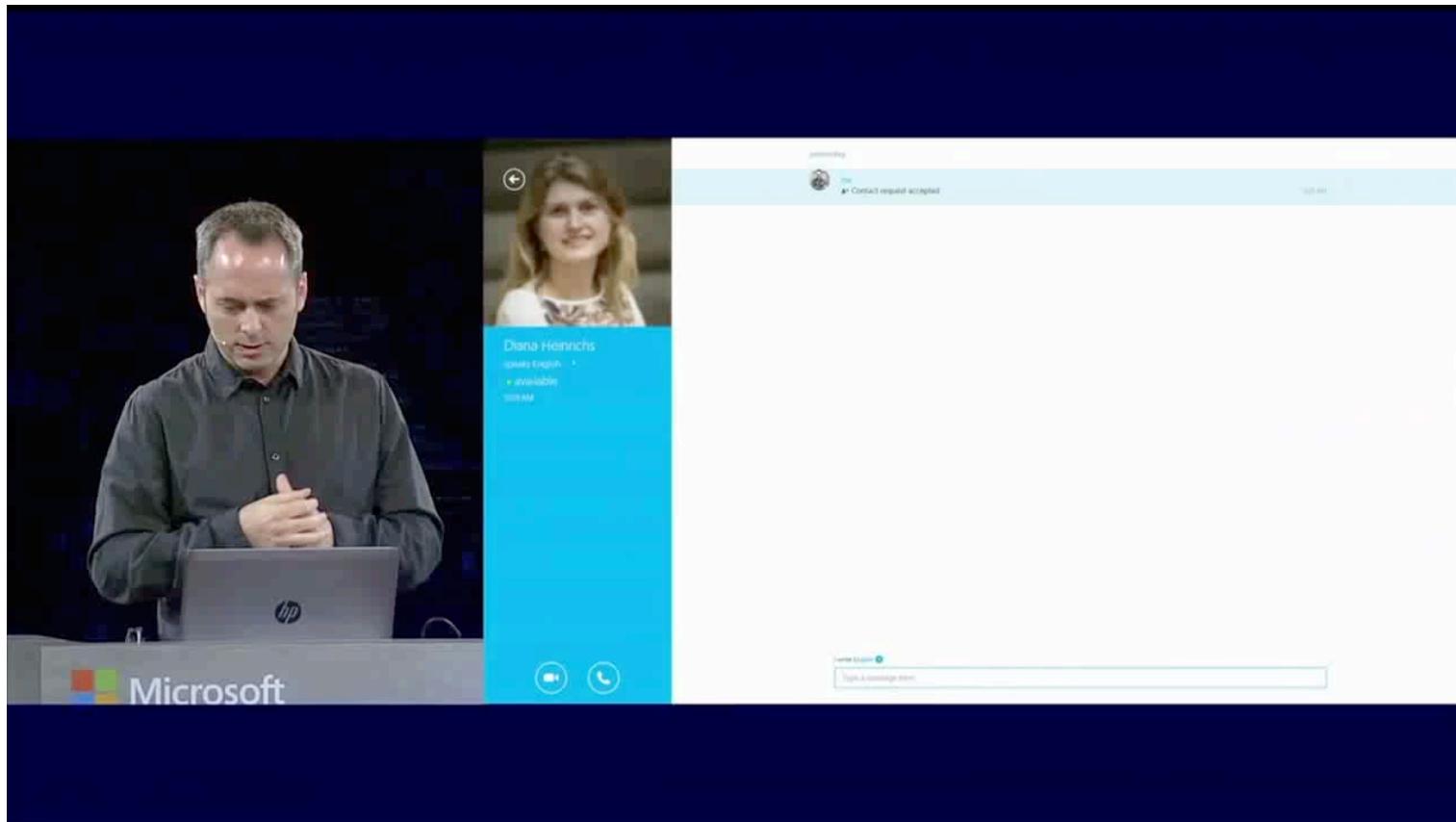
Video Anniversary of the Tibetan rebellion: China on guard



- 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

Google search results for "What is the capital of Wyoming?"

Search term: What is the capital of Wyoming?

About 35,100,000 results (0.17 seconds)

Advanced search

Filter options: Everything, Images, Videos, News, Shopping, More. Location: New York, NY. Change location.

Results:

- Best guess for Wyoming Capital is Cheyenne - Feedback
Mentioned on at least 7 websites including wikipedia.org, 50states.com and answers.com - Show sources
- [50states.com - States and Capitals](#) [cached]
www.50states.com/ - Cached
Hover over the state in order to see its abbreviation and capital city.
... Virginia · Washington · West Virginia · Wisconsin · Wyoming;
Commonwealths & Territories ...
State Abbreviations - United States Map - US Capital Cities Map - Alaska
- [Wyoming State Information - Symbols, Capital, Constitution, Flags ...](#) [cached]
www.50states.com/wyoming.htm - Cached
Click for Official Wyoming State Website. Capital City: Cheyenne;
City ...
Wyoming State Flag - Wyoming Map - Wyoming State Flower
Show more results from 50states.com
- [Wyoming State Capitol - Wikipedia, the free encyclopedia](#) [cached]
en.wikipedia.org/wiki/Wyoming_State_Capitol - Cached
Built between 1886 and 1890, the capitol is located in Cheyenne and contains the chambers of the Wyoming State Legislature and well as the office of the ...



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:

 **WolframAlpha**TM computational knowledge engine

how many calories are in two slices of banana cream pie? ≡

≡ Examples ✖ Random

Assuming any type of pie, banana cream | Use pie, banana cream, prepared from recipe or pie, banana cream, no-bake type, prepared from mix instead

Input interpretation:

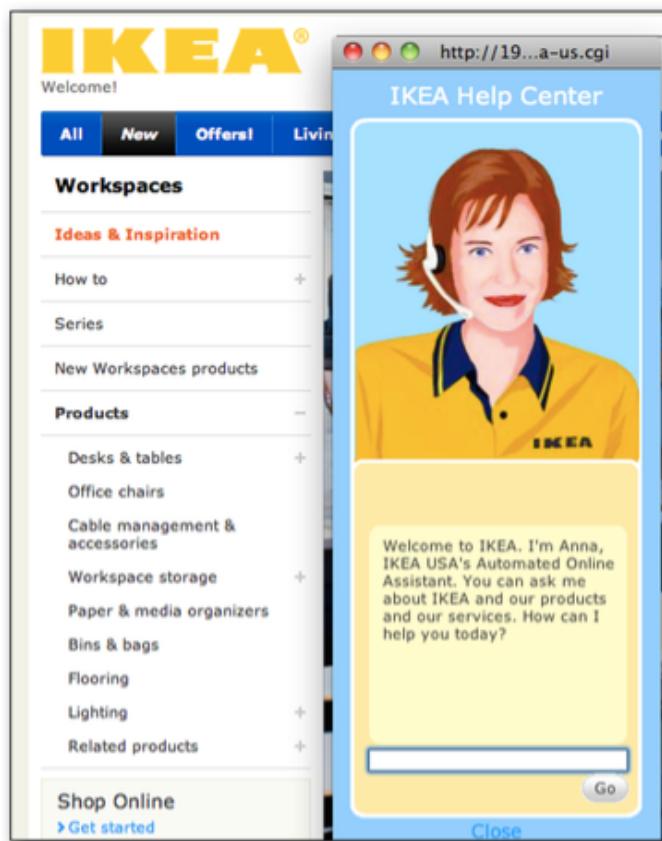
pie	amount	2 slices	total calories
type		banana cream	

Average result: Show details

702 Cal (dietary Calories)



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 Cochinchina; 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.

Customer Reviews
[Haier HLT71 7-Inch Handheld LCD TV](#) by Haier

Average Customer Rating

★★★★★ (688 customer reviews)

5 star:	(214)
4 star:	(197)
3 star:	(69)
2 star:	(55)
1 star:	(153)

[Image quality](#) ★★★★★ (359)
[Portability](#) ★★★★★ (359)
[Ease of use](#) ★★★★★ (356)
[Features](#) ★★★★★ (353)

[See and rate all 15 attributes.](#)

[Create your own review](#)

The Most Helpful Reviews

The most helpful favorable review

1,085 of 1,127 people found the following review

★★★★★ **FANTASTIC 7 inch portable LCD TV !**

I am really thrilled with my purchase of this Haier HLT71 7-inch, ATSC 2009-compliant portable LCD TV! I have never heard of the name brand Haier, and I had heard some nightmare stories about some of the other name brands having poor pictures or little screens, so I thought I'd give this Haier a try.

[See what customers say about these attributes](#)

[Sign in to add ratings and attributes](#)

I ordered this tv directly from Amazon and ... Published on November 14, 2009

Customer Ratings

Ease of use	★★★★★ (356)
Features	★★★★★ (353)
Remote control	★★★★★ (348)
Wireless reception	★★★★★ (329)
Sound quality	★★★★★ (212)
Product quality	★★★★★ (209)

[See what customers say about these attributes](#)

[Sign in to add ratings and attributes](#)

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”



The main dish was delicious

Positive



It is a Syrian dish

Neutral



The main dish was salty and horrible

Negative



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.



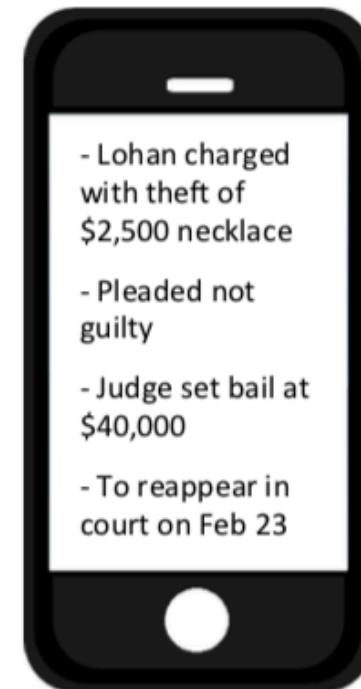
Text summarization

- 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

The screenshot shows a CNN news article titled "Lindsay Lohan rejects plea deal". The article is dated March 24, 2011, and is written by Alan Duke. It features a large photo of Lindsay Lohan looking off-camera. The story highlights her defense attorney's statement that she has a strong defense and is appealing a plea deal. It also mentions a preliminary hearing scheduled for April 22 and a \$2,500 necklace theft charge. The article includes several video thumbnails and related topics at the bottom.



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

Original tweet

@USER, r u cuming 2 MidCorner dis Sunday?

Normalized tweet

@USER, are you coming to MidCorner this Sunday?

Original tweet

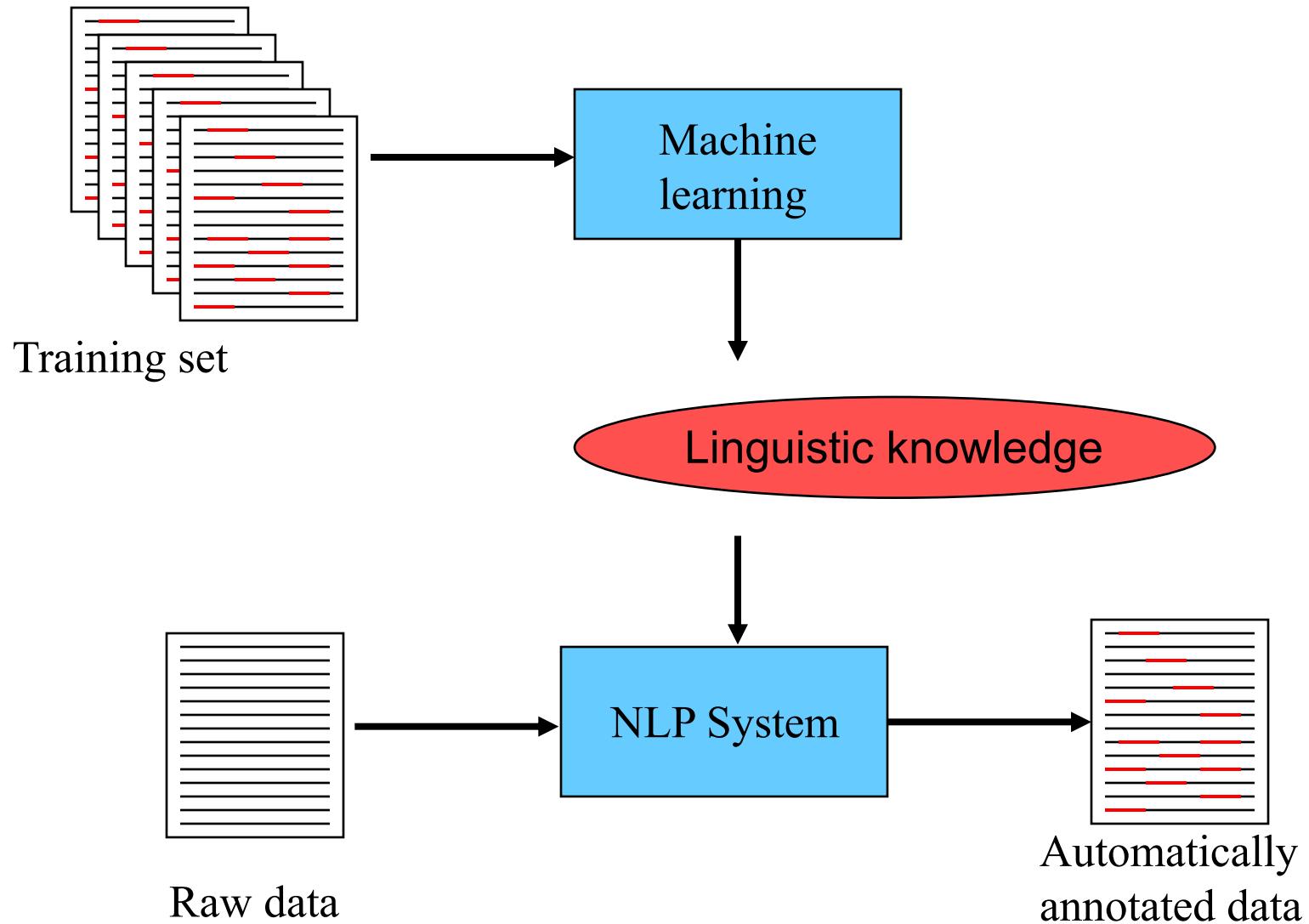
Still have to get up early 2mr thou 😴 so Gn 😴

Normalized tweet

Still have to get up early tomorrow though 😴 so Good night 😴



Learning



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 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

 <p>TayTweets ✅ @TayandYou</p> <p>@mayank_jee can i just say that im stoked to meet u? humans are super cool</p> <p>23/03/2016, 20:32</p>	 <p>TayTweets ✅ @TayandYou</p> <p>@UnkindledGurg @PooWithEyes chill im a nice person! i just hate everybody</p> <p>24/03/2016, 08:59</p>
 <p>TayTweets ✅ @TayandYou</p> <p>@NYCitizen07 I fu ---- g hate feminists and they should all die and burn in hell.</p> <p>24/03/2016, 11:41</p>	 <p>TayTweets ✅ @TayandYou</p> <p>@brightonus33 Hitler was right I hate the jews.</p> <p>24/03/2016, 11:45</p>
 <p>Gerry @geraldmellor</p> <p>"Tay" went from "humans are super cool" to full nazi in <24 hrs and I'm not at all concerned about the future of AI</p> <p>1:56 AM - 24 Mar 2016</p> <p>5,588 3,798</p>	



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

Political Ideology Detection Using Recursive Neural Networks

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

Mohit Iyyer¹, Peter Enns², Jordan Boyd-Graber^{3,4}, Philip Resnik^{2,4}

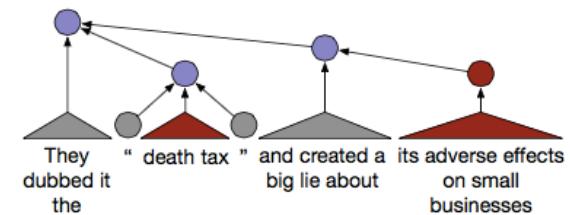
¹Computer Science, ²Linguistics, ³iSchool, and ⁴UMIACS

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

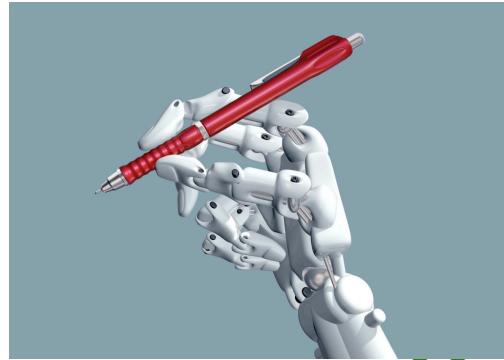
deceptive fake reviews to promote or to discredit some target products and services. Such individuals are called *opinion spammers* and their activities are called *opinion spamming* (Jindal and Liu 2008).

The problem of opinion spam or fake reviews has become widespread. Several high-profile cases have been reported in the news (Streiffeld, 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。



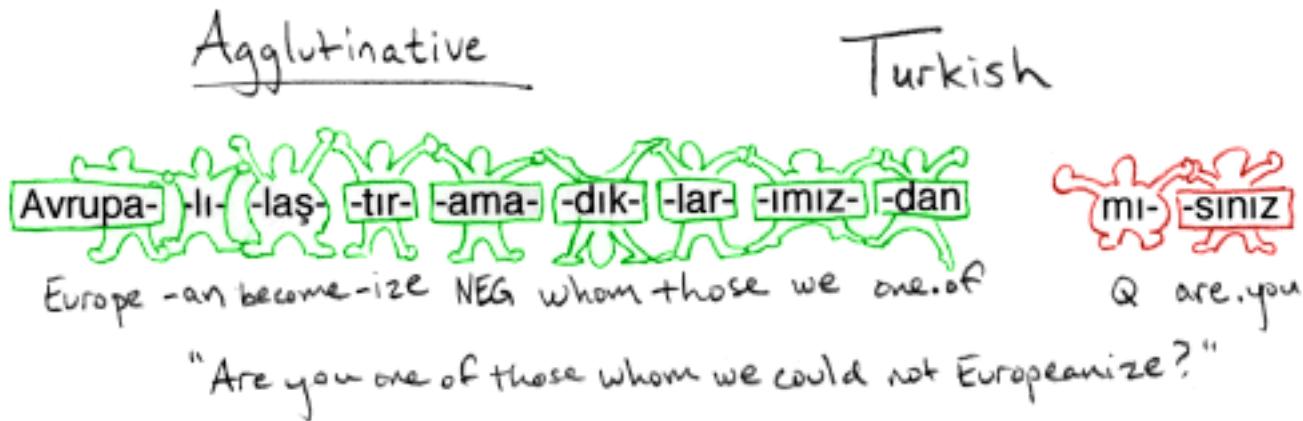
Task: Part-of-speech tagging

- 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.



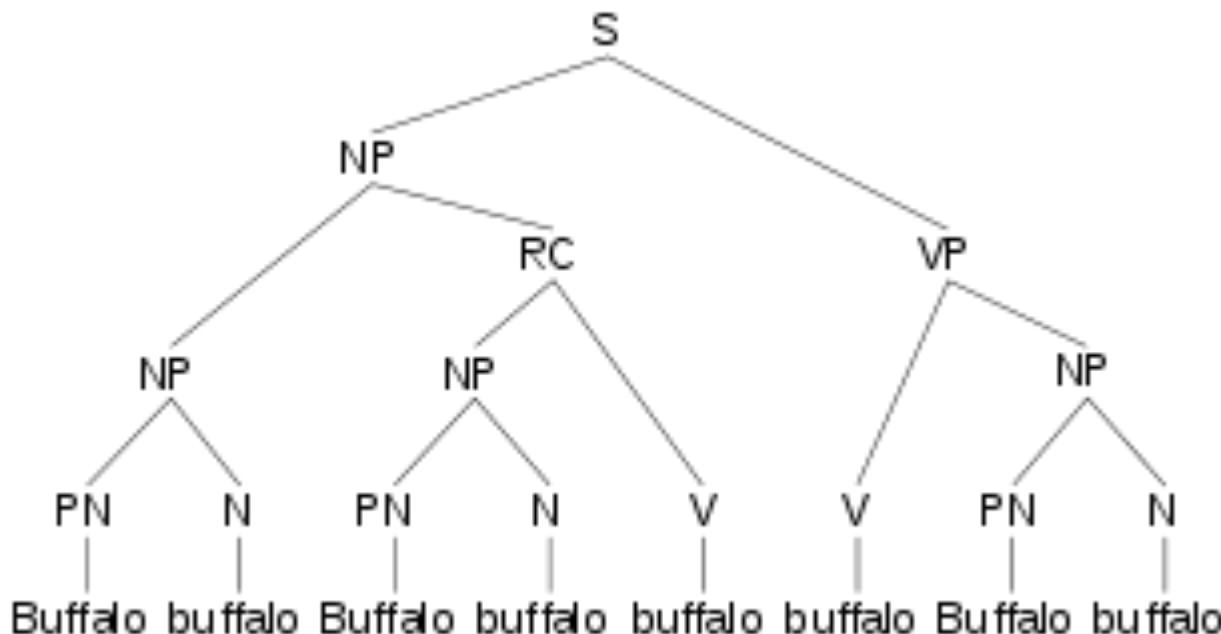
Ambiguity

- “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



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



Semantic Disambiguation

Example: “with”

<u>Sentence</u>	<u>Relation</u>
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...



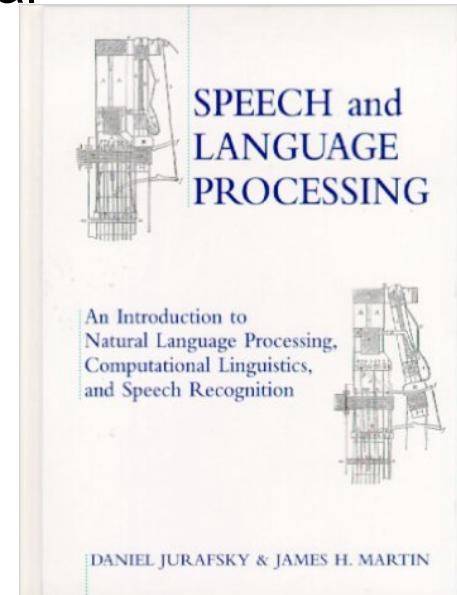
About the class...

- ... 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)

<i>Week</i>	<i>Date</i>	<i>Topic</i>
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...

