

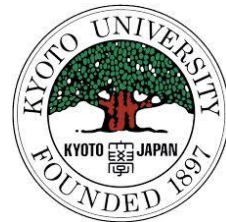
Natural Language Processing: Introduction

14 February 2022

Adam Jatowt

jatowt@acm.org

About me



- **Adam Jatowt**
 - Main affiliation Department of Computer Science & DiSC, *University of Innsbruck* (since 2021/1)
 - Other affiliation
 - Visiting Professor, *National Institute of Advanced Industrial Science and Technology* (Japan)
- Education
 - PhD in Information Science from *University of Tokyo* (Japan)
- Work experience
 - Mainly at *Kyoto University* (Japan) as Assistant and Associate Professor
- Research interests
 - News analysis (question answering, measuring bias in news, timeline summarization)
 - Digital humanities (computational approaches to history, archival search)
 - Social media analysis (collective memory analysis, ambient and user profiling)
- Further information is available at <https://adammo12.github.io/adamjatowt/>



Examples of NLP-related researches that I was involved at..

- Multi timeline summarization in news archives
- Question answering in news archives
- Search and recommendation models in news archives (extracting related content or interesting content from past)
- Sentence temporal validity estimation (information expiry date)
- Text readability and comprehensibility estimation
- Ambient profiling for customized route navigation

If anyone is interested in doing master thesis research in any of these or related NLP/IR topics feel free to contact me for discussion, or collaboration..

About Course

Disclaimer

- I am looking forward to the first this kind of experience for me to teach a block class here at KU for an interdisciplinary group

What is Natural Language Processing?

- The amount of digital textual data being generated every day is huge (e.g., the Web, social media, medical records, digitalized books)
- So does the need for translating, analyzing, and managing this flood of words and text
- **Natural language processing** (NLP) deals with designing **methods** and **algorithms** that take as an input, or produce as an output, **unstructured, natural language data**
- **Natural language processing** is focused on the design & analysis of **computational algorithms** and **representations** for **processing natural human language**

General things about this course

- This is not a linguistic course, but rather a course that includes aspects of **language processing**, **machine learning** and **quantitative methods**
- We will explore statistical techniques for the automatic analysis of natural (human) language data
 - The dominant modeling paradigm is **corpus-driven statistical learning**, with both **supervised** and **unsupervised methods**
- NLP is a huge field!
 - We focus mainly on core ideas and methods needed for fundamental technologies and eventually for applications

Course goals

- Study fundamental tasks in NLP
- Learn some classic and some state-of-the-art techniques
- Acquire some research ideas and experience :-)

What are your goals?

- Why are you here? Perhaps you wish to:
 - work at a company that uses NLP (maybe even as the only text mining expert among engineers...)
 - conduct research in NLP (or IR, MT, etc.)
 - use NLP tools for research in linguistics (or other domains where text data is important: social sciences, humanities, medicine, ..)
- Tell me about yourself:
 - Name
 - Home country or place
 - Academic interests
 - Expectation of this course (if any)
 - Planned topics for thesis if any
 - Etc.

Housekeeping notes

Schedule, Grading, etc.

Course sessions

- We will have everyday sessions
- Some days have more hours than others so please check the plan carefully

Schedule: **Febuary 14 - 22, 2022**

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Date	2/14			2/15			2/16		2/17		2/18		2/21		2/22	
Day	Mon			Tue			Wed		Thu		Fri		Mon		Tue	
Period	3限	4限	5限	4限	5限	4限	5限	4限	5限	4限	5限	4限	5限	3限	4限	5限

情報学科目群

対象学生: 全学部

対象学年: 全学年

教室:

2/14-18 吉田南1号館1共26教室
(MAP No. 85)

2/21-22 総合研究7号館 1階
セミナー室1 (127号室)
(MAP No. 68)

定員: 30 名 (先着順)

Informatics

Eligible students: For all majors

Target year: All students

Classrooms:

[Feb-14 to 18] Seminar room 26,
Yoshida South No.1 Bldg.
(MAP No. 85)

[Feb-21 to 22] Seminar Room 1 (Room #127)
1F, Research Bldg No. 7 (MAP No. 68)

Class capacity: 30 students
(on a first-come, first-served basis)

Tentative schedule

- **Day 1:** NLP overview & introduction, basic text processing (3 classes)
- **Day 2:** Ngrams, language models, spelling error correction (2 classes)
- **Day 3:** Text classification, sentiment analysis (2 classes)
- **Day 4:** Logistic regression, Naïve Bayes (2 classes)
- **Day 5:** Vector semantics, Word embeddings (2 classes)
- **Day 6:** POS tagging, named entity extraction, text summarization (2 classes)
- **Day 7:** Question answering (2 classes) & **Exam** (1 class)

(The schedule may be subject to change during term time. Exam date is fixed)

Slides

- <https://github.com/adammo12/NLP/blob/main/main1.pdf>
- <https://github.com/adammo12/NLP/blob/main/main2.pdf>

Evaluation

- Grading is decided based on final exam
 - Closed book exam
 - Single/multiple choice or descriptive questions
 - Focused on understanding rather than rote memorization of details

Attendance

- The presence is strongly advised but few absences are acceptable
- However, you have to be present at the final examination (day 7)

Academic honesty & integrity

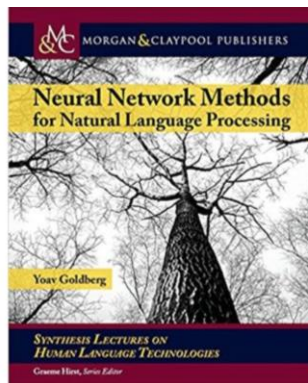
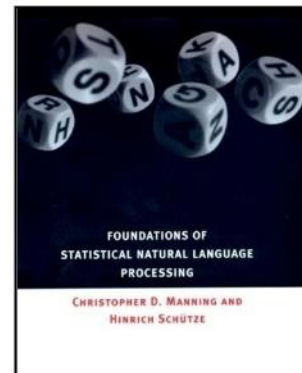
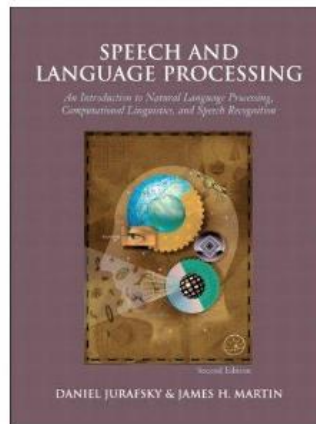
- Students are not allowed to collaborate on answering exam questions. It is an honor code violation to discuss exam questions with other students

Slides

- Slides will be uploaded before each lecture
- Some slides are borrowed from Julia Hockenmaier, Alex Lascarides, Nathan Schneider, Dan Jurafsky, Chris Manning, David Bamman, Ray Mooney, Yulia Tsvetkov, Taylor Berg-Kirk, Dan Klein, Diyi Yang, Jannik Strötgen, Vinay Setty, Anubhav Jangra

Books

- Jurafsky and Martin, *Speech and Language Processing*, 2nd or 3rd Edition
- Manning and Schuetze, *Foundations of Statistical NLP Speech and Language Processing*, MIT Press
- Goldberg, *Neural Network Methods for Natural Language Processing*. Synthesis lectures on human language technologies



Other relevant books

- NLP with Python, The NLTK book, Bird, Klein & Loper.
 - <https://www.nltk.org/book>
- Natural Language Processing Eisenstein.
 - <https://tinyurl.com/eisenstein-nlp>
- Linguistic Fundamentals of NLP Bender, 2013.
 - <http://tinyurl.com/bender-nlp>

Useful NLP resources

- <https://www.nltk.org/>
- <https://towardsdatascience.com/>
- <https://huggingface.co/transformers/>
- <http://nlpprogress.com/>
 - Repository tracking the progress in NLP, including the datasets and the current state-of-the-art for most common NLP tasks
- <https://allennlp.org/>
 - An open-source NLP research library, built on PyTorch
- <https://github.com/flairNLP/flair>

Other related materials

- Primer on NNs:
 - <http://u.cs.biu.ac.il/~yogo/nnlp.pdf>
- Python development tips:
 - <http://people.cs.georgetown.edu/nschneid/pythonathan/>

Relevant scientific conferences

- Association for Computational Linguistics (ACL)
- North American Association for Computational Linguistics (NAACL)
- International Conference on Computational Linguistics (COLING)
- Empirical Methods in Natural Language Processing (EMNLP)
- Conference on Computational Natural Language Learning (CoNLL)
- European chapter of the Association for Computational Linguistics (EACL)

Other related scientific conferences

- Related ones:
 - CIKM
 - WSDM
 - WWW
 - SIGIR
 - ECIR
 - AAAI, IJCAI

Example ACL workshops

- NLP for Building Educational Applications (BEA)
- Fact Extraction and VERification (FEVER)
- Figurative Language Processing (FLP)
- NLP for Conversational AI (NLP4ConvAI)
- Narrative Understanding, Storylines, and Events (NUSE)
- Representation Learning for NLP (RepL4NLP)
- Natural Language Processing for Social Media (SocialNLP)
- Neural Generation and Translation (WNGT)

Workshops we organized last year...

- The 2nd Int. Workshop on Computational Approaches to Historical Language Change 2021 at ACL 2021
- The 4th Int. Workshop on Narrative Extraction from Texts at ECIR 2021

What is NLP? Where is it used?

NLP

- Human language is a general purpose communication tool
- Natural Language Processing
 - Large field: processing natural language text involves many various syntactic, semantic and pragmatic tasks in addition to other problems

Act of Communication

- The goal in the production and comprehension of natural language is **communication**
- Communication for the speaker:
 - **Intention**: Decide when and what information should be transmitted (a.k.a. *content selection*). May require planning and reasoning about agents' goals and beliefs
 - **Generation**: Translate the information to be communicated (in internal logical representation or “language of thought”) into string of words in desired natural language (a.k.a. *surface realization*)
 - **Synthesis**: Output the string in desired modality, text or speech

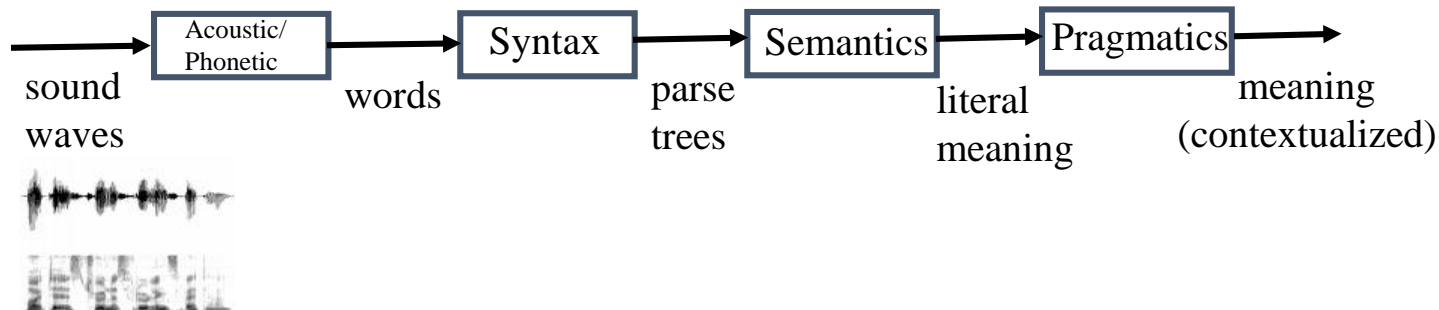
Act of Communication (cont.)

- Communication for the listener:
 - **Perception**: Map input modality to a string of words, e.g. *optical character recognition* (OCR) or *speech recognition*
 - **Analysis**: Determine the information content of the string
 - **Syntactic Interpretation (parsing)**: Find the correct parse tree showing the phrase structure of the string
 - **Semantic Interpretation**: Extract the (literal) meaning of the string (*logical form*)
 - **Pragmatic Interpretation**: Consider effect of the overall context on altering the literal meaning of a sentence
 - **Incorporation**: Decide whether or not to believe the content of the string and add it to the KB.

Syntax, Semantics, Pragmatics

- **Syntax** concerns the proper ordering of words and its affect on meaning.
 - The dog bit the boy.
 - The boy bit the dog.
 - * Bit boy dog the the.
 - Colorless green ideas sleep furiously.
- **Semantics** concerns the (literal) meaning of words, phrases, and sentences.
 - “plant” as a photosynthetic organism
 - “plant” as a manufacturing facility
 - “plant” as the act of putting a seed into ground
- **Pragmatics** concerns the overall communicative and social context and its effect on interpretation.
 - The ham sandwich wants another beer.
 - John thinks vanilla.

Modular Comprehension



Example Syntactic Tasks

Word Segmentation

- Breaking a string of characters into a sequence of words.
- In some written languages (e.g. Chinese, Japanese) words are not separated by spaces.
- Even in English, characters other than white-space can be used to separate words [e.g. , ; . - : ()]
- Examples from English URLs:
 - jumptheshark.com \Rightarrow jump the shark .com
 - myspace.com/pluckerswingbar
 - \Rightarrow myspace .com pluckers wing bar
 - \Rightarrow myspace .com plucker swing bar

Morphological Analysis

- ***Morphology*** is the field of linguistics that studies the internal structure of words
[Wikipedia]
- A ***morpheme*** is the smallest linguistic unit that has semantic meaning
[Wikipedia]
 - e.g. “carry”, “pre”, “ed”, “ly”, “s”
- Morphological analysis is the task of segmenting a word into its morphemes:
 - carried \Rightarrow carry + ed (past tense)
 - independently \Rightarrow in + (depend + ent) + ly
 - Googlers \Rightarrow (Google + er) + s (plural)
 - unlockable \Rightarrow un + (lock + able) ?
 \Rightarrow (un + lock) + able ?

Part-Of-Speech (POS) Tagging

- Annotate each word in a sentence with a part-of-speech.

I ate the spaghetti with meatballs.

Pro V Det N Prep N

John saw the saw and decided to take it to the table.

PN V Det N Con V Part V Pro Prep Det N

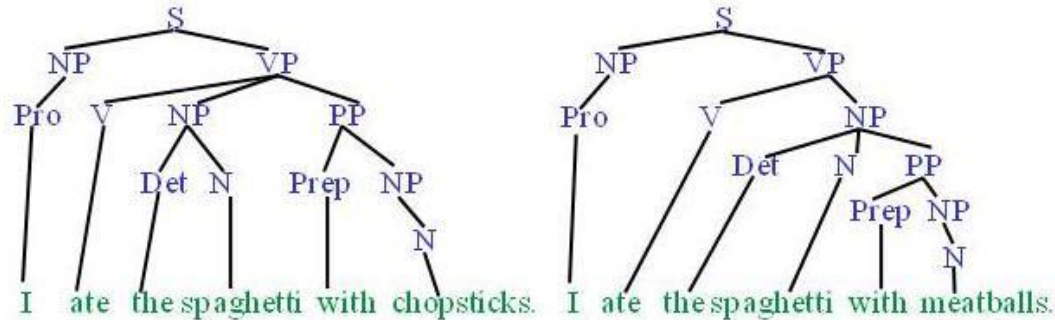
- Useful for subsequent syntactic parsing and word sense disambiguation

Phrase Chunking

- Find all non-recursive noun phrases (NPs) and verb phrases (VPs) in a sentence
 - [NP I] [VP ate] [NP the spaghetti] [PP with] [NP meatballs].
 - [NP He] [VP reckons] [NP the current account deficit] [VP will narrow] [PP to] [NP only # 1.8 billion] [PP in] [NP September]

Syntactic Parsing

- Produce the correct syntactic parse tree for a sentence.



Example Semantic Tasks

Word Sense Disambiguation (WSD)

- Words in natural language usually have a fair number of different possible meanings
 - Ellen has a strong **interest** in computational linguistics.
 - Ellen pays a large amount of **interest** on her credit card.
- For many tasks (e.g., question answering, translation), the proper sense of each ambiguous word in a sentence must be determined.

Semantic Role Labeling (SRL)

- For each clause, determine the **semantic role** played by each **noun phrase** that is an argument to the verb

agent *patient* *source* *destination* *instrument*

- John drove Mary from Austin to Dallas in his Toyota Prius.
 - The hammer broke the window.
-
- Also referred to as a “case role analysis,” “thematic analysis”

Textual Entailment

- Determine whether one natural language sentence entails (implies) another under an ordinary interpretation

Textual Entailment Problems in PASCAL Challenge

TEXT	HYPOTHESIS	ENTAILMENT
<i>Eyeing the huge market potential, currently led by Google, Yahoo took over search company Overture Services Inc last year.</i>	<i>Yahoo bought Overture.</i>	TRUE
<i>Microsoft's rival Sun Microsystems Inc. bought Star Office last month and plans to boost its development as a Web-based device running over the Net on personal computers and Internet appliances.</i>	<i>Microsoft bought Star Office.</i>	FALSE
<i>The National Institute for Psychobiology in Israel was established in May 1971 as the Israel Center for Psychobiology by Prof. Joel.</i>	<i>Israel was established in May 1971.</i>	FALSE
<i>Since its formation in 1948, Israel fought many wars with neighboring Arab countries.</i>	<i>Israel was established in 1948.</i>	TRUE

Temporal/change-based inference

- S1: I am cooking pasta Bolonese for the first time
- S2: The water is not hot yet

- S1: I am cooking pasta Bolonese for the first time
- S2: It was really tasty and I need a rest now

In which case S2 implies the end of action expressed in S1?

Example Pragmatics Tasks

Anaphora Resolution/Co-Reference

- Determine which phrases in a document refer to the same underlying entity.

- John put the carrot on the plate and ate it.

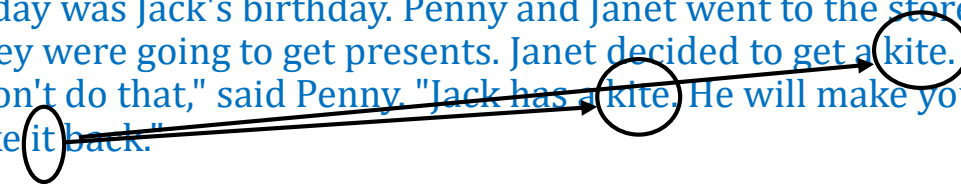


- Bush started the war in Iraq. But the president needed the consent of Congress.



- Some cases require difficult reasoning.

- Today was Jack's birthday. Penny and Janet went to the store. They were going to get presents. Janet decided to get a kite. "Don't do that," said Penny. Jack has a kite. He will make you take it back.



Ellipsis Resolution

- Frequently words and phrases are omitted from sentences when they can be inferred from context

"Wise men talk because they have something to say;
fools, because they have to say something." (Plato)

"Wise men talk because they have something to say;
fools **talk** because they have to say something." (Plato)

Other Tasks

Information Extraction (IE)

- Identify phrases in language that refer to specific types of entities and relations in text.
- **Named entity recognition** is task of identifying names of people, places, organizations, etc.

people organizations places

- **Michael Dell** is the CEO of **Dell Computer Corporation** and lives in **Austin Texas**.

- **Relation extraction** identifies specific relations between entities.

- **Michael Dell** is the **CEO of** **Dell Computer Corporation** and lives in **Austin Texas**.

Question Answering

- Directly answer natural language questions based on information presented in a corpora of textual documents (e.g. the web).
 - When was Barack Obama born? (*factoid*)
 - August 4, 1961
 - Who was president when Barack Obama was born?
 - John F. Kennedy
 - How many presidents have there been since Barack Obama was born?
 - 9

Question Answering

What did Barack Obama teach?

Barack Hussein Obama II (born August 4, 1961) is an American attorney and politician who served as the 44th [President of the United States](#) from January 20, 2009, to January 20, 2017. A member of the [Democratic Party](#), he was the first [African American](#) to serve as president. He was previously a [United States Senator](#) from [Illinois](#) and a member of the [Illinois State Senate](#).

Obama was born in 1961 in [Honolulu, Hawaii](#), two years after the territory was [admitted to the Union](#) as the 50th state. Raised largely in Hawaii, he also spent one year of his childhood in [Washington state](#) and four years in [Indonesia](#). After graduating from [Columbia University](#) in 1983, he worked as a [community organizer](#) in [Chicago](#). In 1988, he enrolled in [Harvard Law School](#), where he was the first black president of the [Harvard Law Review](#). After graduating, he became a [civil rights attorney](#) and a professor, teaching [constitutional law](#); at the [University of Chicago Law School](#) from 1992 to 2004.

Barack Obama



44th [President of the United States](#)

In office

Reading Comprehension

- Read a passage of text and answer questions about it
- Example from Stanford SQuAD dataset

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?

gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?

graupel

Where do water droplets collide with ice crystals to form precipitation?

within a cloud

Text Summarization

- Produce a short summary of a longer document or article.
 - **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.

Machine Translation (MT)

- Translate a sentence from one natural language to another.
 - Hasta la vista, bebé \Rightarrow
Until we see each other again, baby.
 - 我喜欢汉堡 \Rightarrow
I like burgers.

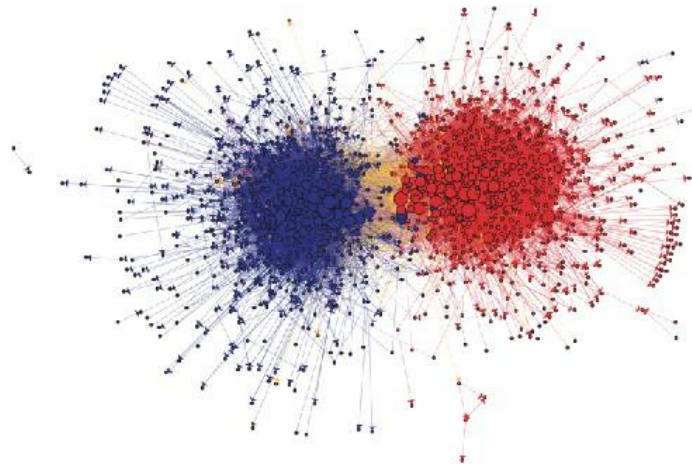
Chatbots (Spoken Dialogue Systems) Management

- Q: Is it going to rain today?
- A: It will be mostly sunny. No rain is expected.



Computational Social Science

- e.g., finding politics-focused communities in blogs
- e.g., detecting the triggers of censorship in blogs/social media
- e.g., inferring power differentials in language use



Link structure in political blogs
Adamic and Glance 2005

Computational Journalism

The Rise of the Robot Reporter



By Jaclyn Peiser

Feb. 5, 2019

As reporters and editors find themselves the victims of layoffs at digital publishers and traditional newspaper chains alike, journalism generated by machine is on the rise.

Roughly a third of the content published by Bloomberg News uses some form of automated technology. The system used by the company, Cyborg, is able to assist reporters in churning out thousands of articles on company earnings reports each quarter.

The program can dissect a financial report the moment it appears and spit out an immediate news story that includes the most pertinent facts and figures. And unlike business reporters, who find working on that kind of thing a snooze, it does so without complaint.

Untiring and accurate, Cyborg helps Bloomberg in its race against Reuters, its main rival in the field of quick-twitch business financial journalism, as well as giving it a fighting chance against a more recent player in the information race, hedge funds, which use artificial intelligence to serve their clients fresh facts.



Computational Humanities, e.g.:

Ted Underwood (2016), “The Life Cycles of **Genres**,” Cultural Analytics

Ryan Heuser, Franco Moretti, Erik Steiner (2016), The **Emotions** of London

Richard Jean So and Hoyt Long (2015), “Literary Pattern Recognition”

Andrew Goldstone and Ted Underwood (2014), “The Quiet Transformations of Literary Studies,” New Literary History

Franco Moretti (2005), Graphs, Maps, Trees

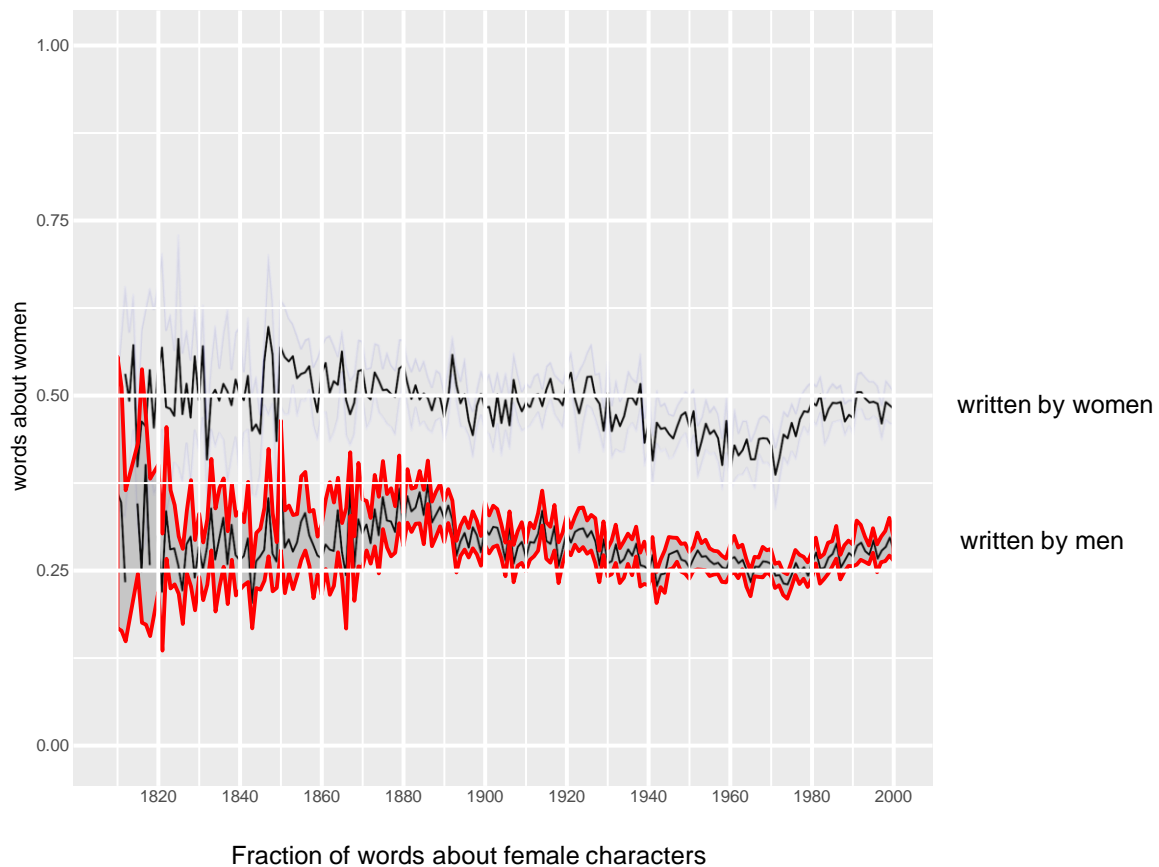
Holst Katsma (2014), **Loudness** in the Novel

So et al (2014), “**Cents** and Sensibility”

Matt Wilkens (2013), “The **Geographic** Imagination of Civil War Era American Fiction”

Jockers and Mimno (2013), “Significant **Themes** in 19th-Century Literature,”

Ted Underwood and Jordan Sellers (2012). “The Emergence of **Literary Diction**.” JDH



Ted Underwood, David Bamman, and Sabrina Lee (2018), "The Transformation of Gender in English-Language Fiction," Cultural Analytics

Text-driven forecasting

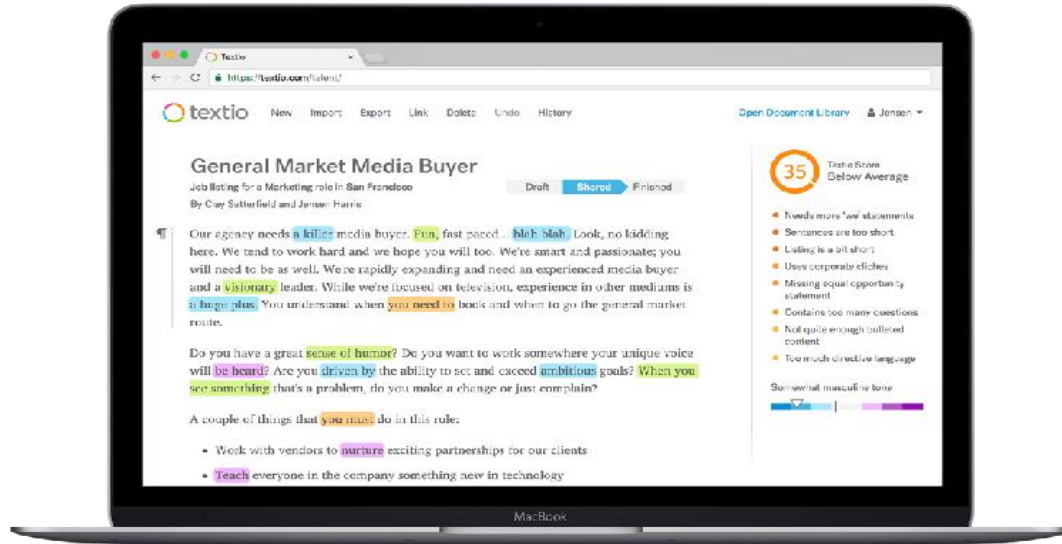
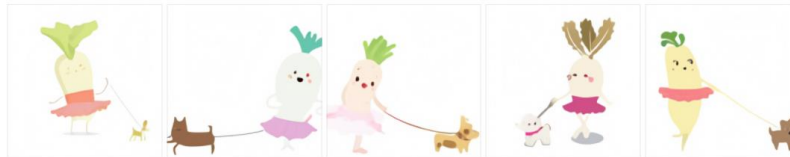


Image and text

TEXT PROMPT

an illustration of a baby daikon radish in a tutu walking a dog

AI-GENERATED
IMAGES



Edit prompt or view more images ↴

TEXT PROMPT

an armchair in the shape of an avocado . . .

AI-GENERATED
IMAGES



Edit prompt or view more images ↴

TEXT PROMPT

a store front that has the word 'openai' written on it . . .

AI-GENERATED
IMAGES



Edit prompt or view more images ↴

<https://openai.com/blog/dall-e/>

Historical Discovery?

- E.g., book in language we cannot understand



Voynich manuscript



Why NLP is hard?

Why NLP is hard?

- Outside of artificial annotated data sets, machines don't have access to any direct representation of speaker meaning, but only to natural language utterances
- And the artificial, annotated data sets include only specific subsets of meaning..

Why NLP is hard?

- Ambiguity at many levels:
 - Word senses: bank (finance or river?)
 - Part of speech: chair (noun or verb?)
 - Syntactic structure: I saw a man with a telescope
 - Quantifier scope: Every child loves some movie
 - Multiple: I saw her duck



Ambiguity is Explosive

- Ambiguities compound to generate enormous numbers of possible interpretations.
- In English, a sentence ending in n prepositional phrases has *over* 2^n syntactic interpretations (cf. Catalan numbers).
 - “I saw the man with the telescope”: 2 parses
 - “I saw the man on the hill with the telescope.”: 5 parses
 - “I saw the man on the hill in Texas with the telescope”: 14 parses
 - “I saw the man on the hill in Texas with the telescope at noon.”: 42 parses
 - “I saw the man on the hill in Texas with the telescope at noon on Monday” 132 parses

Humor and Ambiguity

- Many jokes rely on the ambiguity of language:
 - Groucho Marx: One morning I shot an elephant in my pajamas. How he got into my pajamas, I'll never know.
 - She criticized my apartment, so I knocked her flat.
 - Policeman to little boy: "We are looking for a thief with a bicycle." Little boy: "Wouldn't you be better using your eyes."
 - Why is the teacher wearing sun-glasses? Because the class is so bright.

Why is Language Ambiguous?

- Having a unique linguistic expression for every possible conceptualization that could be conveyed would make language overly complex and linguistic expressions unnecessarily long
- Allowing resolvable ambiguity permits shorter linguistic expressions, i.e., data compression
- Language relies on people's ability to use their knowledge and inference abilities to properly resolve ambiguities

Natural Languages vs. Computer Languages

- Ambiguity is the primary difference between natural and computer languages
- Formal programming languages are designed to be unambiguous, i.e., they can be defined by a grammar that produces a unique parse for each sentence in the language

Ambiguity Resolution is Required for Translation

- Syntactic and semantic ambiguities must be properly resolved for correct translation:
 - “John plays the guitar.” → “John toca la guitarra.”
 - “John plays soccer.” → “John juega el fútbol.”
- Anecdotal examples of early MT systems giving the following results when translating from English to Russian and then back to English:
 - “The spirit is willing but the flesh is weak.” ⇒
“The liquor is good but the meat is spoiled.”
 - “Out of sight, out of mind.” ⇒
“Invisible idiot.”

Resolving Ambiguity

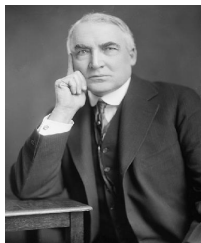
- Choosing the correct interpretation of linguistic utterances requires knowledge of:
 - Syntax
 - E.g., an agent is typically the subject of the verb
 - Semantics
 - Michael and Ellen are names of people
 - Austin is the name of a city (and of a person)
 - Toyota is a car company and Prius is a brand of car
 - Pragmatics
 - World knowledge and commonsense knowledge
 - Credit cards require users to pay financial interest
 - Agents must be animate while a hammer is not animate

Commonsense and world knowledge

- Shared knowledge (“Warren is running for president.”)
- Common sense (“Pan is in the room where we cook.” (frying pan, kitchen))
- Common sense (“I dropped the glass on the floor and it broke” v.s. “I dropped the hammer on the glass and it broke”)
- Temporal commonsense: taking a walk is shorter than taking holidays



Elizabeth Warren
2020



Warren G. Harding
1920



Importance of probability

- Unlikely interpretations of words can combine to generate spurious ambiguity:
 - “Time flies like an arrow” has 4 parses, including those meanings:
 - Insects of a variety called “time flies” are fond of a particular arrow
 - A command to record insects’ speed in the manner that an arrow would
 - “The a are of I” is a valid English noun phrase
 - “a” is an adjective for the letter A
 - “are” is a noun for an area of land (as in hectare)
 - “I” is a noun for the letter I
- Statistical methods allow computing most likely interpretation by combining probabilistic evidence from a variety of uncertain knowledge sources

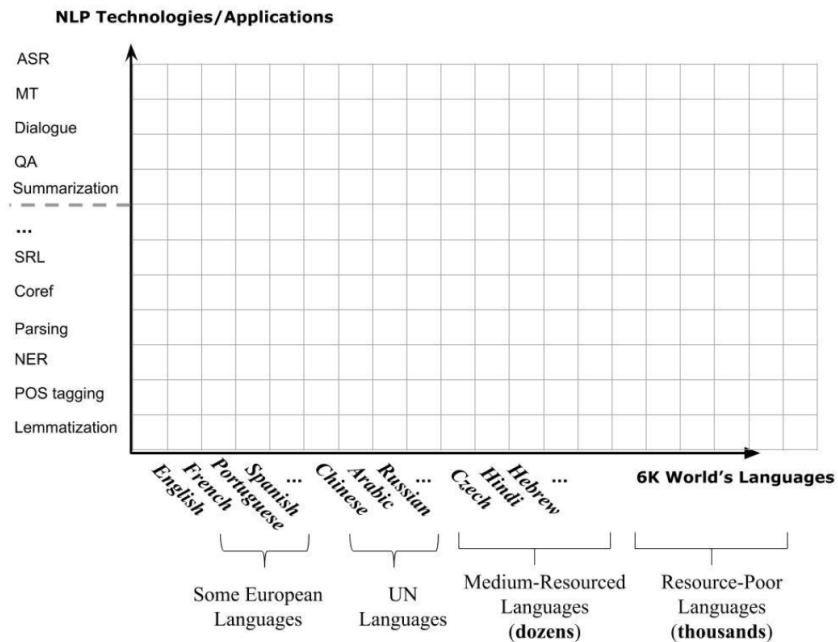
Pipelining problem

- Combining separate independent components for speech recognition, syntax, semantics, pragmatics, etc. allows for more convenient modular software development
- However, often the constraints from “higher level” processes are needed to disambiguate “lower level” processes
 - Example of syntactic disambiguation relying on semantic disambiguation:
 - “At the zoo, several men were showing a group of students various types of flying animals. Suddenly, one of the students hit the man **with** a **bat**.”

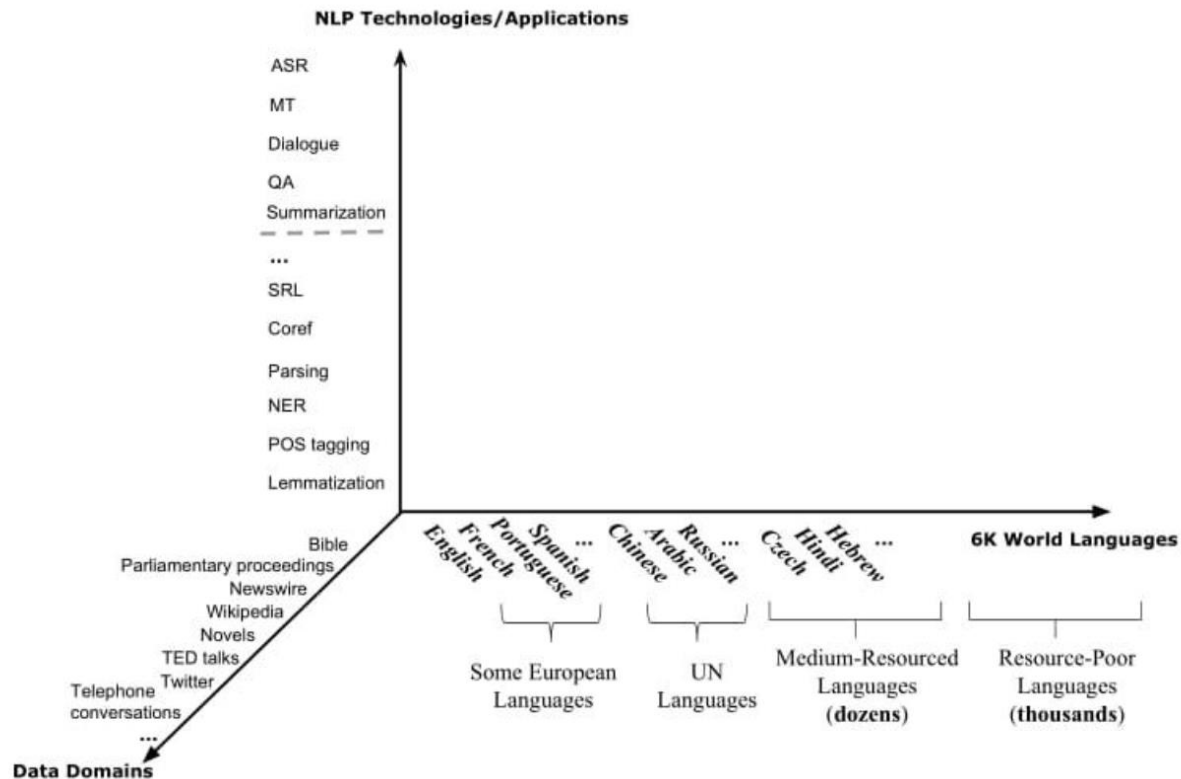
Pipelining problem (cont.)

- If a hard decision is made at each stage, it cannot backtrack when a later stage indicates it is incorrect
 - If attach “with a bat” to the verb “hit” during syntactic analysis, then cannot reattach it to “man” after “bat” is disambiguated during later semantic or pragmatic processing

Many languages



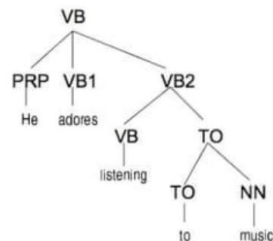
Many languages and domains



Japanese example

syntactic parsing

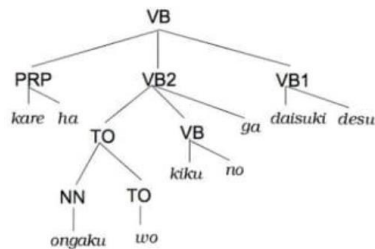
SVO



word alignment

he adores listening to music

SOV



Brief history of the field

Historical perspective

- 1950's: Early days

- Foundational work: automata, information theory, etc.
- First speech systems
- Machine translation (MT) hugely funded by military
- Toy models: MT using basically word-substitution
- Optimism!
- Rationalism: approaches to design hand-crafted rules to incorporate knowledge and reasoning mechanisms into intelligent NLP systems (e.g., ELIZA for simulating a Rogerian psychotherapist, MARGIE for structuring real-world information into concept ontologies).

- 1960's and 1970's: NLP Winter

- Bar-Hillel (FAHQT) and ALPAC reports “kills” MT
- Work shifts to deeper models, syntax... but toy domains / grammars

- 1980's and 1990's: The Empirical Revolution

- Expectations get reset
- Empiricism: characterized by the exploitation of data corpora and of (shallow) machine learning and statistical models (e.g., Naive Bayes, HMMs, IBM translation models).
- Corpus-based methods become central
- Deep analysis often traded for robust and simple approximations
- Evaluate everything

Historical perspective

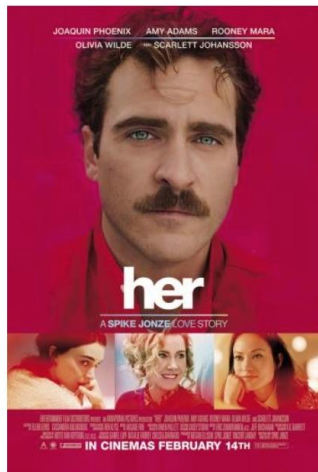
- 2000+: Richer Statistical Methods

- Models increasingly merge linguistically sophisticated representations with statistical methods
- Begin to get both breadth and depth

- 2013+: Deep Learning

- feature engineering (considered as a bottleneck) is replaced with representation learning and/or deep neural networks (e.g., <https://www.deepl.com/translator>)
- A very influential paper in this revolution: [Collobert et al., 2011]

Where are we now?



Baseline mutual information model (Li et al. 2015)

A: Where are you going? (1)
B: I'm going to the restroom. (2)
A: See you later. (3)
B: See you later. (4)
A: See you later. (5)
B: See you later. (6)
...
...
A: how old are you? (1)
B: I'm 16. (2)
A: 16? (3)
B: I don't know what you are talking about. (4)
A: You don't know what you are saying. (5)
B: I don't know what you are talking about. (6)
A: You don't know what you are saying. (7)
...

Li et al. (2016), "Deep Reinforcement Learning for Dialogue Generation" *EMNLP*



Where are we now?



- "Stop"
- "Turn it up"
- "Volume level 6"
- "Repeat that"
- "What can you do?"
- "Play some music"
- "Play music by [artist]"
- "Play dance music on YouTube"
- "Play KEXP radio on TuneIn"
- "Play the latest episode of Radiolab"
- "Pause"
- "Next song"
- "When's my first appointment tomorrow?"
- "Wake me up at 6am tomorrow"
- "Tell me about my day"
- "How long will it take to get to work?"
- "What's the weather today?"

And many new developments recently...

Facebook AI Creates Its Own Language In Creepy Preview Of Our Potential Future

Computers can now describe images using language you'd understand

A REPORTER AT LARGE OCTOBER 14, 2019 ISSUE

CAN A MACHINE LEARN TO WRITE FOR THE NEW YORKER?

How predictive-text technology could transform the future of the written word.

The AI Text Generator That's Too Dangerous to Make Public

Researchers at OpenAI decided that a system that scores well at understanding language could too easily be manipulated for malicious intent.

How AI Can Create And Detect Fake News

A.I. breakthroughs in natural-language processing are big for business

BY JEREMY KAHN

Barbie Wants to Get to Know Your Child

With the help of A.I., America's most famous doll tries to fulfill a timeless dream — convincing little girls that she's a real friend. What will happen if they believe her?

Related fields

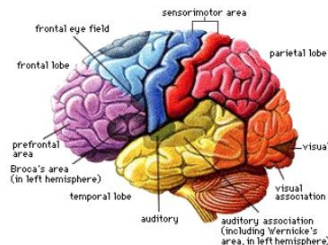
- Cognitive Science

- Figuring out how the human brain works
- Includes the bits that do language
- Humans: the only working NLP prototype..

- Speech Processing

- Mapping audio signals to text
- Traditionally separate from NLP, recently converging
- Two components: acoustic models and language models
- Language models in the domain of statistical NLP

- Computational Linguistics (CL)



Difference of NLP & CL

- Most conferences and journals that host natural language processing research bear the name “**computational linguistics**” (e.g., ACL, NAACL, COLING)
- NLP and CL may be thought of as essentially synonymous
- While there is substantial overlap, there is an important focus difference
 - **CL is essentially linguistics supported by computational methods** (similar to computational biology, computational astronomy)
 - In linguistics, language is the object of study
 - NLP focuses on **solving well-defined tasks involving human language** (e.g., translation, query answering, holding conversations, information extraction, machine reading)
 - Fundamental linguistic insights may be crucial for accomplishing these tasks, but success is ultimately measured by whether and how well the job gets done according to used evaluation metrics

Other related fields

- Artificial Intelligence
- Machine Learning
- Formal Language (Automata) Theory
- Linguistics
- Psycholinguistics
- Philosophy of Language