NLP without Annotated Dataset

Course Overview

Sowmya Vajjala

Seminar für Sprachwissenschaft, University of Tübingen, Germany

8 January 2021

Today's plan

- Course overview
- NLP overview

Course Overview

About me

- I work as a full time researcher at the National Research Council, Canada in Digital Technologies Research Center.
- 2020: Wrote a book for O'Reilly (http://www.practicalnlp.ai/)
- 2018-19: Senior Data Scientist in software engineering r&d teams in Toronto
- 2016-18: Assistant Professor (tenure track) at Iowa State University, USA
- 2011-15: PhD at University of Tuebingen, Germany
- Before that: software developer, Bachelors/Masters in Engineering

About You

- 22/46 filled the questionnaire so far (at 1600 CET).
- Mostly MA, followed by BA ISCL?
- Background: Mostly Linguistics, and many wrote they know some programming.
- Languages you speak: English, German, Italian, Spanish, Portuguese, Swedish, Arabic, Russian, Mandarin, Korean,Japanese, Thai, Bahasa Indonesia, Acehnese!

About You

- 22/46 filled the questionnaire so far (at 1600 CET).
- Mostly MA, followed by BA ISCL?
- Background: Mostly Linguistics, and many wrote they know some programming.
- Languages you speak: English, German, Italian, Spanish, Portuguese, Swedish, Arabic, Russian, Mandarin, Korean,Japanese, Thai, Bahasa Indonesia, Acehnese!
- Why are you enrolled in this course? What do you want to do later?
  - A common answer: learn practical aspects of NLP and work in the industry
  - Do research on NLP for native languages (non English/German)

Teaching experience

- 2011-13: 2 Hauptseminar courses at Tuebingen (with Prof Meurers)
- 2016-18:
  - Applied Linguistics grad students: Python programming, Introduction to NLP
  - Grad Computer science students: Statistical NLP
  - Undergrad students from all disciplines : "Language and Computers", "Text as Data" (R), Technical Communication
- 2020: Guest course at Munich Graduate School of Economics, Germany (online)

Course Background

- NLP is a part of many day to day applications we use, such as search engines, virtual assistants on your smartphones and various functionalities in your email.

Course Background

- NLP is a part of many day to day applications we use, such as search engines, virtual assistants on your smartphones and various functionalities in your email.
- When we think of NLP, we think of the various algorithms, neural network architectures, and so on.

Course Background

- NLP is a part of many day to day applications we use, such as search engines, virtual assistants on your smartphones and various functionalities in your email.
- When we think of NLP, we think of the various algorithms, neural network architectures, and so on.
- However, what drives all of them are large collections of annotated corpora.

Course Background

- NLP is a part of many day to day applications we use, such as search engines, virtual assistants on your smartphones and various functionalities in your email.
- When we think of NLP, we think of the various algorithms, neural network architectures, and so on.
- However, what drives all of them are large collections of annotated corpora.
- What do you do when you don't have access to such datasets, though?

Course Objectives

- Provide an overview of NLP system development pipeline
- Discuss some common approaches for collecting, cleaning and exploring text data
- Introduce some methods to develop labeled data for NLP

Expected Learning Outcomes

Students should be able to:

- Understand the end to end NLP system development pipeline
- Compile and explore labeled/annotated corpora for NLP
- Build some basic text classification and information extraction systems

... upon successful completion of the course..

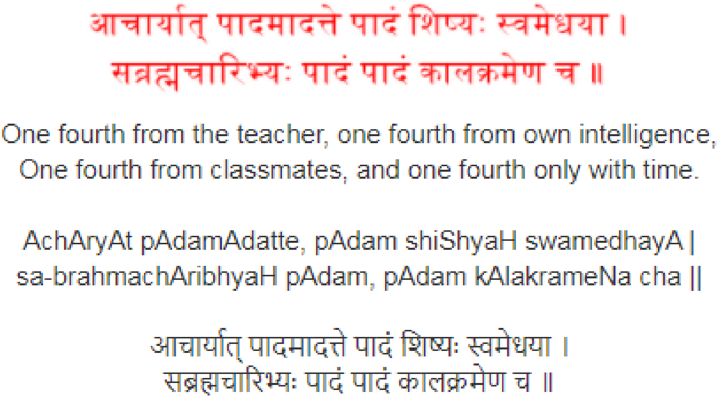
Pre-requisities

1. Intermediate proficiency in any programming language (Python preferred)
2. Comfortable installing libraries etc on their laptops
3. Knowledge of the usage of virtual environments (venv, anaconda) is useful

What the course can't do

- Don't expect to become an NLP expert with one compact course.
- Contents may not always meet your own expectations, but there is a term paper and a group discussion, which gives you opportunities to explore your specific interests related to this topic.
- The course won't teach you programming.

How we learn and grow



आचार्यात् पादमादत्ते पादं शिष्यः स्वमेधया ।
सब्रह्मचारिभ्यः पादं पादं कालक्रमेण च ॥

One fourth from the teacher, one fourth from own intelligence,
One fourth from classmates, and one fourth only with time.

AchAryAt pAdamAdatte, pAdam shiShyaH swamedhayA |
sa-brahmachAribhyaH pAdam, pAdam kAlakrameNa cha ||

आचार्यात् पादमादत्ते पादं शिष्यः स्वमेधया ।
सब्रह्मचारिभ्यः पादं पादं कालक्रमेण च ॥

Source

Course Logistics

Meeting and Location

- January 8 2021-January 29, 2021, M W F, 17:00 s.t. - 19:30 (Central European Time).
  1. 8th Jan 2021 (Friday)
  2. 11th, 13th, 15th Jan 2021 (Mon, Wed, Fri)
  3. 18th, 20th, 22nd Jan 2021 (Mon, Wed, Fri)
  4. 25th, 27th, 29th Jan 2021 (Mon, Wed, Fri)
- Location: Zoom Meeting-ID: 990 5086 7382
  Kenncode: 296817
- For a one to one meeting, email me to set up a time. I am keeping 1700-1800 free on most days in January for these one to one meetings.

Course Website

- Moodle: https://moodle.zdv.uni-tuebingen.de/course/view.php?id=1301
- Syllabus, Lecture slides and Assignments will be uploaded there.

Course Format + Credits

- Video lectures + Discussion (I may sometimes pick people randomly and ask a question!)
- Assignments (2)
- Team presentations: You are expected to form into groups of 2-4 people, pick a paper from the reading list on the website (or any other relevant paper) and present a brief discussion in a live session (10-15 minutes per group)
- Assignments
- Term paper(optional)

Credits: 3 CP (+ 3 CP if you write a term paper)

Textbooks

1. "Speech and Language Processing" by Jurafsky and Martin (2/3 editions)
2. "Practical Natural Language Processing" by Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta and Harshit Surana.
3. NLTK book
4. For Python: "Python for Everybody" Charles Severance

(Details on how to access these books are in the Syllabus document)

Course Topics

1. Introduction (1 session)
2. NLP Pipeline (1 session)
3. Corpus collection, extraction, exploration (1 session)
4. Automatically labeling data (3 sessions)

remaining 4 sessions are for student presentations and review.

Assignments/Grading (for 6 CP)

1. 2 Assignments (30% of the grade)
2. 1 presentation (30% of the grade)
3. 1 term paper (30% of the grade)
4. classroom participation (10% of the grade)

(For 3 CP: Split the term paper grade between two assignments)

Assignments

- Two assignments, already uploaded on Moodle
- They are not difficult - the goal is not to trick you, but to make you think about the challenges of working with NLP problems in real world.
- My preferred programming language is Python, I am okay with Java, R, C, C++, and anything else (note: I can't debug for you. What you submit should run error-free on my machine).

Presentation

- Students can work in teams of 2-4 people and present one of the research papers related to course topics, from a given list of papers.
- Papers are listed in the syllabus document. If you want to present a different paper, talk to me first.
- Pick your teams early (deadline: 13th Jan)

Term Paper

- Work on a short project involving NLP and write a report describing your work (6-8 pages long in single column, latex formatted document)
- Some ideas are listed in the syllabus document. If you want to work on something else, talk to me first.
- If you want to get into NLP research later, explore some of your ideas through this term paper!

Classroom Participation

- Attending live meetings
- Participating in the forum
- Communicating (Asking questions, informing me if something comes up and you can't attend etc)
- Submitting stuff on time

Important Deadlines

1. Decide on a team for group discussion (13th Jan 2021)
2. Decide on a paper for group discussion (15th Jan 2021)
3. Group Discussions (22nd-27th Jan 2021)
4. Assignments 1 and 2 Submission (6th Feb 2021)
5. Decide on term paper topic (29th Jan 2021)
6. Term paper submission (13th Feb 2021)

- Questions so far?

NLP without Annotated Dataset

NLP: An Introduction

Sowmya Vajjala

Seminar für Sprachwissenschaft, University of Tübingen, Germany

8 January 2021

Outline

1. **What is NLP?**
2. The many faces of NLP
3. What makes NLP Challenging?
4. Some common NLP Tasks: An overview
5. The levels of language processing: some examples
6. Approaches to NLP

What is NLP?

note: images without source attribution are taken from our book: practicalnlp.ai

What is NLP?

- NLP is a sub-field of Artificial intelligence that is concerned with analyzing, modeling and understanding human language using computational methods.

What is NLP?

- NLP is a sub-field of Artificial intelligence that is concerned with analyzing, modeling and understanding human language using computational methods.
- The eventual goal is to make computers understand (and generate) human languages, and make them communicate with humans like humans

What is NLP?

- NLP is a sub-field of Artificial intelligence that is concerned with analyzing, modeling and understanding human language using computational methods.
- The eventual goal is to make computers understand (and generate) human languages, and make them communicate with humans like humans
- Because of its role in the process of human-computer interaction, NLP has a wide range of technological applications

What is NLP?

- NLP is a sub-field of Artificial intelligence that is concerned with analyzing, modeling and understanding human language using computational methods.
- The eventual goal is to make computers understand (and generate) human languages, and make them communicate with humans like humans
- Because of its role in the process of human-computer interaction, NLP has a wide range of technological applications
- It is also becoming popular as a research method in a broad range of disciplines in social sciences.

Inter-disciplinary by nature

NLP is very inter-disciplinary. Draws from research in Computer Science, Linguistics, Mathematics, Statistics, Psychology etc.,

Computational Linguistics vs NLP

The terms are used synonymously. However, generally, NLP is typically used by people involved in engineering and technology development, and CL is typically used by traditionally linguistics groups who adapted computational methods.

History of NLP

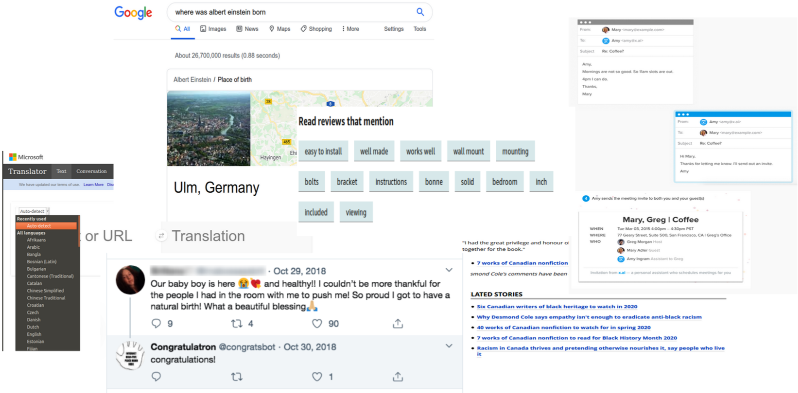
1. Foundational ideas: 40s and 50s. WWII and Beyond.
2. Main NLP problem of that time (and even now): Machine Translation
3. First few decades: Work focused on the development of speech recognition systems, logic based language understanding systems, creating elaborate grammars to teach human language to computers, rule based systems, and automatic language generation.
4. Late 90s on: Advent of statistical methods and machine learning
5. 2010s: Deep learning
6. Last few years: more into interpretable models, discussion about ethical issues, introspection about the field etc.

Where is NLP useful in real world?

- general purpose applications: search, email, voice based assistants on phones etc.
- domain specific applications: e-commerce, legal, finance, health care etc.
- educational technology: language teaching, learning, assessment tools
- language revitalization software
- disaster management tools

...

Everyday NLP Applications



where was albert einstein born

About 26,700,000 results (0.88 seconds)

Albert Einstein / Place of birth

Ulm, Germany

Read reviews that mention

easy to install | well made | works well | wall mount | mounting | bolts | bracket | instructions | bonne | solid | bedroom | inch | included | viewing

We have updated our terms of use. Learn More Dis

· Oct 29, 2018

Our baby boy is here 😭❤️ and healthy!! I couldn't be more thankful for the people I had in the room with me to push me! So proud I got to have a natural birth! What a beautiful blessing🙏

Congratulatron @congratsbot · Oct 30, 2018

congratulations!

"I had the great privilege and honour o
together for the book."

- 7 works of Canadian nonfiction

smond Cole's comments have been

LATED STORIES

- Six Canadian writers of black heritage to watch in 2020
- Why Desmond Cole says empathy isn't enough to eradicate anti-black racism
- 40 works of Canadian nonfiction to watch for in spring 2020
- 7 works of Canadian nonfiction to read for Black History Month 2020
- Racism in Canada thrives and pretending otherwise nourishes it, say people who live it

From: Mary <mary@example.com>
To: Amy <amy@x.ai>
Subject Re: Coffee?

Amy,
Mornings are not so good. So 11am slots are out.
4pm I can do.
Thanks,
Mary

From: Amy <amy@x.ai>
To: Mary <mary@example.com>
Subject Re: Coffee?

Hi Mary,
Thanks for letting me know. I'll send out an invite.
Amy

4 Amy sends the meeting invite to both you and your guest(s)

Mary, Greg | Coffee

WHEN Tue Mar 03, 2015 4:00pm – 4:30pm PST
WHERE 77 Geary Street, Suite 500, San Francisco, CA | Greg's Office
WHO Greg Morgan Host
Mary Adler Guest
Amy Ingram Assistant to Greg

Invitation from x.ai — a personal assistant who schedules meetings for you

Where is NLP used in real-world?

1. Apple Siri, Google assistant and other such software
2. Google Translate and the likes
3. Search Engines
4. Question Answering (e.g., IBM Watson)
5. Sentiment analysis of product reviews on Amazon, for example
6. Spam classification in Gmail, Yahoomail etc
7. Information extraction from text (e.g., identifying calendar entries automatically in gmail)
8. Spelling/Grammar check tools, language learning apps such as DuoLingo etc.

Where is NLP used elsewhere?

1. NLP is used as a method to answer research questions in many disciplines.
2. NLP sometimes plays a major role in discipline specific challenges, going beyond being just a research method.
3. In Google Scholar, I saw mentions of NLP methods in journals as diverse as Asian studies & History to Clinical Oncology.

Outline

1. What is NLP?
2. **The many faces of NLP**
3. What makes NLP Challenging?
4. Some common NLP Tasks: An overview
5. The levels of language processing: some examples
6. Approaches to NLP

The Many Faces of NLP

The many faces of NLP

Three broad groups:

1. NLPers: NLP researchers in academia and industry
2. Other researchers who use NLP methods in their research
3. Industry professionals developing NLP based applications

Data Scientist

Building real-world
NLP systems

Product Manager

Applying NLP to their
products & domains

Annotator

Or linguists creating
NLP data

ML Engineer

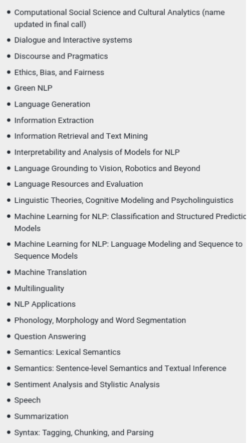
Or data engineer applying
MLOps to scale NLP systems

Business Leader

CXOs, VPs, founders incorporating
NLP in their business

NLP research - an overview

A snapshot of contemporary NLP research topics



- Computational Social Science and Cultural Analytics (name updated in final call)
- Dialogue and Interactive systems
- Discourse and Pragmatics
- Ethics, Bias, and Fairness
- Green NLP
- Language Generation
- Information Extraction
- Information Retrieval and Text Mining
- Interpretability and Analysis of Models for NLP
- Language Grounding to Vision, Robotics and Beyond
- Language Resources and Evaluation
- Linguistic Theories, Cognitive Modeling and Psycholinguistics
- Machine Learning for NLP: Classification and Structured Predictic Models
- Machine Learning for NLP: Language Modeling and Sequence to Sequence Models
- Machine Translation
- Multilinguality
- NLP Applications
- Phonology, Morphology and Word Segmentation
- Question Answering
- Semantics: Lexical Semantics
- Semantics: Sentence-level Semantics and Textual Inference
- Sentiment Analysis and Stylistic Analysis
- Speech
- Summarization
- Syntax: Tagging, Chunking, and Parsing

(source: https://2021.naacl.org/calls/papers/)

Trends in NLP Research

How can one quickly showcase contemporary NLP research to others? ⇒ Some paper titles from best paper awards over the past 5 years may give a picture.
source: https://aclweb.org/aclwiki/Best_paper_awards
and 2020 conference websites. (everything is open access!)

Contemporary NLP Research - 1

A lot of research focuses on what can perhaps be called core NLP tasks and applications:

- Bridging the Gap between Training and Inference for Neural Machine Translation
- Improving Evaluation of Machine Translation Quality Estimation
- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
- Linguistically-Informed Self-Attention for Semantic Role Labeling
- Beyond Accuracy: Behavioral Testing of NLP Models with CheckList

Contemporary NLP Research - 2

There is also a lot of work on a range of other topics from human language comprehension to mental health

- Finding syntax in human encephalography with beam search
- [Probabilistic Typology: Deep Generative Models of Vowel Inventories
- Feuding Families and Former Friends; Unsupervised Learning for Dynamic Fictional Relationships
- Depression and Self-Harm Risk Assessment in Online Forums
- Digital Voicing of Silent Speech

NLP Research in 2020: Introspection

Some papers from ACL 2020 Theme: "Taking Stock of Where We've Been and Where We're Going"

- The State and Fate of Linguistic Diversity and Inclusion in the NLP World
- How Can We Accelerate Progress Towards Human-like Linguistic Generalization?
- A Call for More Rigor in Unsupervised Cross-lingual Learning
- Automated Evaluation of Writing – 50 Years and Counting
- Speech Translation and the End-to-End Promise: Taking Stock of Where We Are

NLP Research: Summary

- generally heavy on algorithms/methods based on machine learning/deep learning
- relatively less focus on corpus creation, evaluation beyond standard tests, and rule engineering
- recent interest in bias in models, ethics, interpretability etc
- lot of introspection in 2020

What does NLP in Industry look like?

- large r&d teams building NLP focused products, for their own use as well as for third parties (Google Cloud NLP, Amazon Comprehend, IBM Watson, Microsoft NLP etc)
- software teams where NLP contributes to existing product functionalities
- speech to text/text to speech software, transcription tools etc.
- language learning/teaching/assessment software

and so on.

NLP in Industry: A few examples

- SpeechRater spoken response assessment tool
- Grammarly
- DuoLingo
- Lawdroid makes chatbots for law firms to perform various functions (e.g., paralegal bot, reception bot etc)
- Bloomberg uses sentiment analysis on news articles about companies to support stock market decisions.
- Pfizer uses IBM Watson for cancer treatment drug discovery.

....

NLP in Industry: Summary

- We saw a few use cases so far. NLP is useful in many other industry scenarios too.
- Companies that build software involving local, non-English NLP are also growing in many countries.
- There are also companies that primarily do annotation for NLP and other Machine Learning projects. (e.g., Appen Ltd)

NLP in Industry: Summary

- We saw a few use cases so far. NLP is useful in many other industry scenarios too.
- Companies that build software involving local, non-English NLP are also growing in many countries.
- There are also companies that primarily do annotation for NLP and other Machine Learning projects. (e.g., Appen Ltd)
- To conclude,
  - industry NLP involves a wide range of applications
  - requires people from diverse backgrounds such as linguists, software developers, product managers etc.

NLP in Other Disciplines: An Overview

Where is NLP used in other disciplines?

- NLP is used as a method to answer research questions in many disciplines.
- NLP sometimes plays a major role in discipline specific challenges, going beyond being **just** a research method.
- In Google Scholar, I saw mentions of NLP methods in journals as diverse as Asian studies & History to Clinical Oncology.
- I will show a sample of work taken from a few disciplines that may interest you.

NLP in Applied Linguistics Research

Causal discourse analyzer: Improving automated feedback on academic ESL

- used Stanford CoreNLP software + linguistic rule engineering to identify cause and effect discourse in non-native writing.
- Causal markers were first identified by a manual, functional linguistic analysis of a corpus, and were then used to develop the above rules.
- evaluated in terms of precision and recall, on manually annotated essays by 17 students.

NLP in Language Acquisition Research

Automatic extraction of subordinate clauses and its application in second language acquisition research

- built a tool to extract subordinate clauses using Stanford dependency parser followed by several hand crafted rules.
- validated the tool through an evaluation with annotated test set and manual inspection.
- used this tool to analyze a large-scale learner corpus and investigate the effects of first language (L1) on the acquisition of subordination in second language (L2) English.

NLP and Corpus Linguistics

Dependency parsing of learner English

- proposed an approach to control for annotation bias in learner language parse annotations.
- evaluated multiple NLP parsers on learner English.
- identified and quantified the influence of learner writing errors on parser's efficiency.

Few more examples ..

Medical Informatics: Chen, L., Gu, Y., Ji, X., Sun, Z., Li, H., Gao, Y., & Huang, Y. (2020). [Extracting medications and associated adverse drug events using a natural language processing system combining knowledge base and deep learning](https://doi.org/10.1093/jamia/ocz141). Journal of the American Medical Informatics Association : JAMIA, 27(1), 56–64.

Plant Science: Braun, I. R., & Lawrence-Dill, C. J. (2019). [Automated methods enable direct computation on phenotypic descriptions for novel candidate gene prediction](https://www.frontiersin.org/articles/10.3389/fpls.2019.01629/full). Frontiers in Plant Science, 10, 1629.

Civil Engineering: Le, T., & David Jeong, H. (2017). [NLP-based approach to semantic classification of heterogeneous transportation asset data terminology](https://par.nsf.gov/servlets/purl/10069437). Journal of Computing in Civil Engineering, 31(6), 04017057.

Economics: Hansen, S., McMahon, M., & Prat, A. (2018). [Transparency and deliberation within the FOMC: a computational linguistics approach](https://academic.oup.com/qje/article/133/2/801/4582916). The Quarterly Journal of Economics, 133(2), 801-870.

Political Science: Benoit, K., Munger, K., & Spirling, A. (2019). [Measuring and explaining political sophistication through textual complexity](https://onlinelibrary.wiley.com/doi/full/10.1111/ajps.12423). American Journal of Political Science, 63(2), 491-508.

Urban planning: Plunz, R. A., Zhou, Y., Vintimilla, M. I. C., Mckeown, K., Yu, T., Uguccioni, L., & Sutto, M. P. (2019). [Twitter sentiment in New York City parks as measure of well-being. Landscape and urban planning](https://www.sciencedirect.com/science/article/pii/S0169204618305863), 189, 235-246.

Cultural Heritage: Machidon, O. M., Tavčar, A., Gams, M., & Duguleană, M. (2020). [CulturalERICA: A conversational agent improving the exploration of European cultural heritage](https://www.sciencedirect.com/science/article/pii/S1296207418308136). Journal of Cultural Heritage, 41, 152-165.

NLP in other disciplines: Summary

- Clearly, there are many more. I just sampled a few examples, from even fewer disciplines!
- Existing NLP tools + rules is a commonly used approach in some disciplines.
- Doing user studies and using a small set of manually annotated documents for validation of approaches is also a common method.
- In some fields (e.g., medical informatics), we also see state of the art deep learning and NLP.

How are the faces of NLP different from each other?

- NLP researchers focus on developing new methods, using standard corpora/evaluation procedures, and comparing against SOTA.
- Industry professionals focus on end users, end to end system development and maintenance.
  "If you think Machine Learning will give you a 100% boost, then a heuristic will get you a 50% of the way there" - Martin Zinkevich, Google
- Other discipline researchers are concerned how to use NLP methods to address their own research questions.

Questions?

My questions:

- What role does NLP play in your daily life?
- What do you want to do with NLP?

Outline

1. What is NLP?
2. The many faces of NLP
3. **What makes NLP Challenging?**
4. Some common NLP Tasks: An overview
5. The levels of language processing: some examples
6. Approaches to NLP

What makes NLP challenging?

(what sort of issues pose problems for a computer?)

Language is ambiguous

Some ambiguous sentences

- Newspaper headlines
  - "Children make delicious snacks"
  - "Dead expected to rise"
  - "Republicans grill IRS chief over lost emails"
- Normal, grammatical sentences can be ambiguous too:
  - "I saw a man on a hill with a telescope."
  - "Look at the man with one eye"

We are not even talking about ambiguities involving speech or alternative interpretations due to stress/emphasis on some word.

Some types of ambiguity

1. Lexical ambiguity: due to multiple meanings or senses of word usage
   e.g., He stood near the **bank**
2. Structural ambiguity: due to syntactic structure
   e.g., I saw the man on the hill with telescope.
3. Semantic ambiguity: more interpretations possible
   e.g., John and Mary are married (to each other? or to different people?)
4. Referential ambiguity
   e.g., She dropped the *plate* on the *table* and broke **it**
5. Ambiguity due to the use of non-literal language
   e.g., Time flies like an arrow

Good source to read more:
http://cs.nyu.edu/faculty/davise/ai/ambiguity.html

"common" knowledge for humans

Look at these two sentences:
Dog bit man.
Man bit dog.
- For a computer, both of them are linguistically the same. We know only the first one is "normal" English sentence because we have "world knowledge".

Few more challenges

- Language is diverse: many different forms of documents such as news, tweets, legal texts etc.
- Language is creative: its use changes over time, and vocabulary gets richer.
- There are many different languages in the world
- Many spelling variations, slangs, sarcasm etc.

- NLP solutions should account for all these things!

So, the summary is:

perfect NLP is hard to achieve because of all these issues that come up when we start using computers to analyze language!

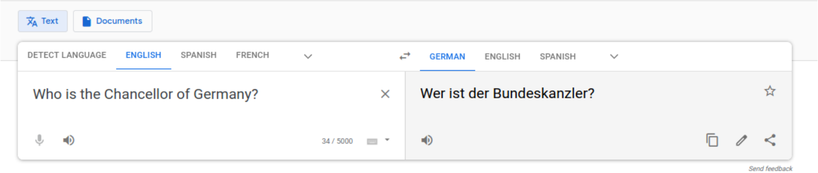
Let me pause ...

Do you have any funny NLP moments to share? (e.g., google translate mishaps?)

Outline

1. What is NLP?
2. The many faces of NLP
3. What makes NLP Challenging?
4. **Some common NLP Tasks: An overview**
5. The levels of language processing: some examples
6. Approaches to NLP

Machine Translation

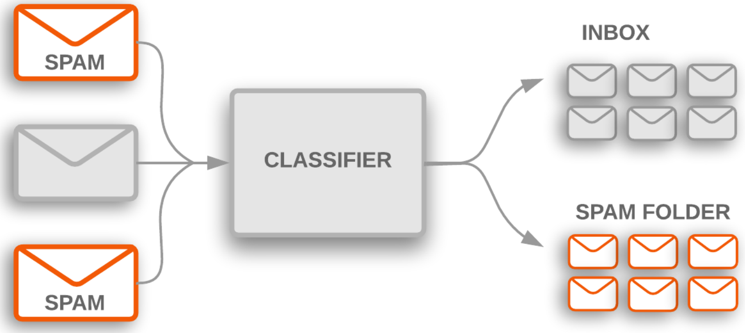


Who is the Chancellor of Germany?

Wer ist der Bundeskanzler?

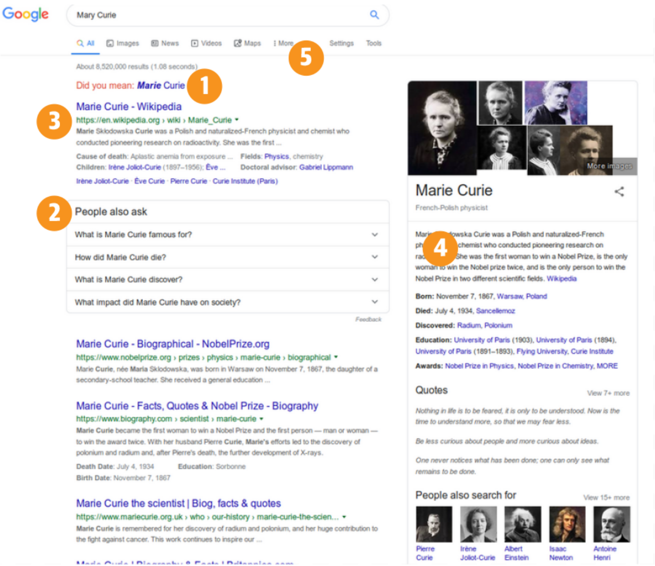
source: https://translate.google.com

Text Classification



source: https://developers.google.com/machine-learning/guides/text-classification

Search



Mary Curie

About 8,520,000 results (1.08 seconds)

Did you mean: ***Marie*** Curie

Marie Curie - Wikipedia
https://en.wikipedia.org › wiki › Marie_Curie ▾
Marie Sklodowska **Curie** was a Polish and naturalized-French physicist and chemist who conducted pioneering research on radioactivity. She was the first ...
Cause of death: Aplastic anemia from exposure ... **Fields**: Physics, chemistry
Children: Irène Joliot-Curie (1897–1956); Ève ... **Doctoral advisor**: Gabriel Lippmann
Irène Joliot-Curie · Ève Curie · Pierre Curie · Curie Institute (Paris)

People also ask

- What is Marie Curie famous for?
- How did Marie Curie die?
- What is Marie Curie discover?
- What impact did Marie Curie have on society?

Feedback

Marie Curie - Biographical - NobelPrize.org
https://www.nobelprize.org › prizes › physics › marie-curie › biographical ▾
Marie **Curie**, née **Maria** Sklodowska, was born in Warsaw on November 7, 1867, the daughter of a secondary-school teacher. She received a general education ...

Marie Curie - Facts, Quotes & Nobel Prize - Biography
https://www.biography.com › scientist › marie-curie ▾
Marie Curie became the first woman to win a Nobel Prize and the first person — man or woman — to win the award twice. With her husband Pierre **Curie**, **Marie's** efforts led to the discovery of polonium and radium and, after Pierre's death, the further development of X-rays.
Death Date: July 4, 1934 **Education**: Sorbonne
Birth Date: November 7, 1867

Marie Curie the scientist | Biog, facts & quotes
https://www.mariecurie.org.uk › who › our-history › marie-curie-the-scien... ▾
Marie Curie is remembered for her discovery of radium and polonium, and her huge contribution to the fight against cancer. This work continues to inspire our ...

Marie Curie | Biography & Facts | Britannica.com

Marie Curie
French-Polish physicist

Mari owska Curie was a Polish and naturalized-French ph chemist who conducted pioneering research on rad She was the first woman to win a Nobel Prize, is the only woman to win the Nobel prize twice, and is the only person to win the Nobel Prize in two different scientific fields. Wikipedia

Born: November 7, 1867, Warsaw, Poland

Died: July 4, 1934, Sancellemoz

Discovered: Radium, Polonium

Education: University of Paris (1903), University of Paris (1894), University of Paris (1891–1893), Flying University, Curie Institute

Awards: Nobel Prize in Physics, Nobel Prize in Chemistry, MORE

Quotes View 7+ more

Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less.

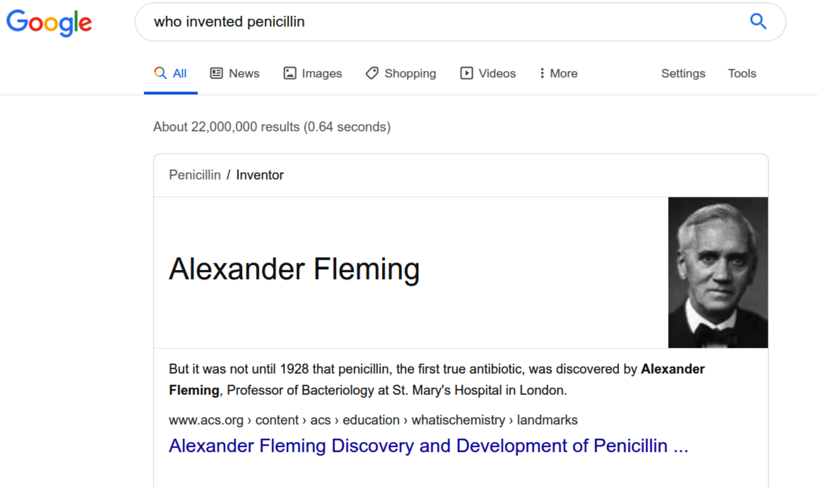
Be less curious about people and more curious about ideas.

One never notices what has been done; one can only see what remains to be done.

People also search for View 15+ more

Pierre Curie · Irène Joliot-Curie · Albert Einstein · Isaac Newton · Antoine Henri

Question Answering



who invented penicillin

About 22,000,000 results (0.64 seconds)

Penicillin / Inventor

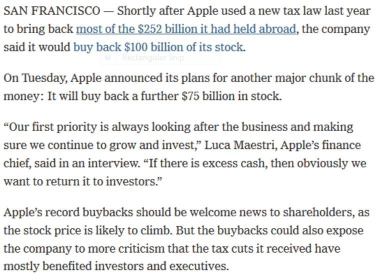
Alexander Fleming

But it was not until 1928 that penicillin, the first true antibiotic, was discovered by **Alexander Fleming**, Professor of Bacteriology at St. Mary's Hospital in London.

www.acs.org › content › acs › education › whatischemistry › landmarks

Alexander Fleming Discovery and Development of Penicillin ...

Information Extraction



SAN FRANCISCO — Shortly after Apple used a new tax law last year to bring back most of the $252 billion it had held abroad, the company said it would buy back $100 billion of its stock.

On Tuesday, Apple announced its plans for another major chunk of the money: It will buy back a further $75 billion in stock.

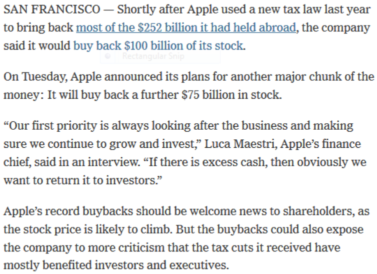
"Our first priority is always looking after the business and making sure we continue to grow and invest," Luca Maestri, Apple's finance chief, said in an interview. "If there is excess cash, then obviously we want to return it to investors."

Apple's record buybacks should be welcome news to shareholders, as the stock price is likely to climb. But the buybacks could also expose the company to more criticism that the tax cuts it received have mostly benefited investors and executives.

¿¿

- Who is Luca Mestri?

Information Extraction



SAN FRANCISCO — Shortly after Apple used a new tax law last year to bring back most of the $252 billion it had held abroad, the company said it would buy back $100 billion of its stock.

On Tuesday, Apple announced its plans for another major chunk of the money: It will buy back a further $75 billion in stock.

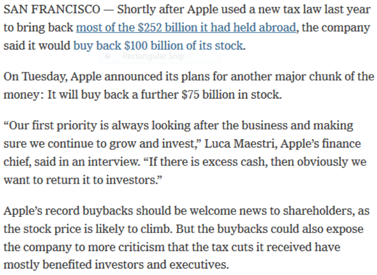
“Our first priority is always looking after the business and making sure we continue to grow and invest,” Luca Maestri, Apple’s finance chief, said in an interview. “If there is excess cash, then obviously we want to return it to investors.”

Apple’s record buybacks should be welcome news to shareholders, as the stock price is likely to climb. But the buybacks could also expose the company to more criticism that the tax cuts it received have mostly benefited investors and executives.

¿¿

- Who is Luca Mestri? needs: Named Entity Recognition and Linking, Relation extraction
- What is the article about?

Information Extraction



SAN FRANCISCO — Shortly after Apple used a new tax law last year to bring back most of the $252 billion it had held abroad, the company said it would buy back $100 billion of its stock.

On Tuesday, Apple announced its plans for another major chunk of the money: It will buy back a further $75 billion in stock.

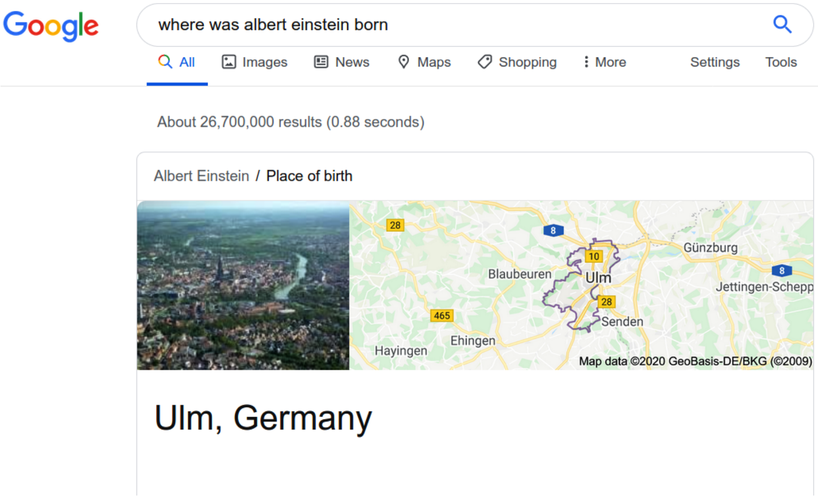
"Our first priority is always looking after the business and making sure we continue to grow and invest," Luca Maestri, Apple's finance chief, said in an interview. "If there is excess cash, then obviously we want to return it to investors."

Apple's record buybacks should be welcome news to shareholders, as the stock price is likely to climb. But the buybacks could also expose the company to more criticism that the tax cuts it received have mostly benefited investors and executives.

¿¿

- Who is Luca Mestri? needs: Named Entity Recognition and Linking, Relation extraction
- What is the article about? needs: Key phrase extraction, event extraction

Named Entity Extraction/Linking



where was albert einstein born

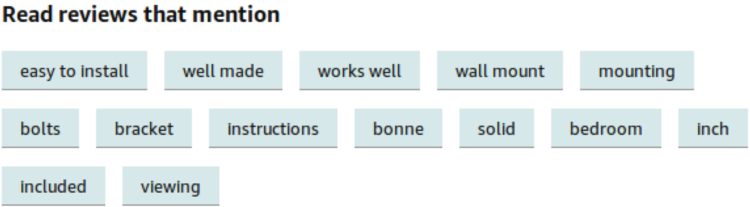
About 26,700,000 results (0.88 seconds)

Albert Einstein / Place of birth

Map data ©2020 GeoBasis-DE/BKG (©2009)

Ulm, Germany

Key Phrase Extraction



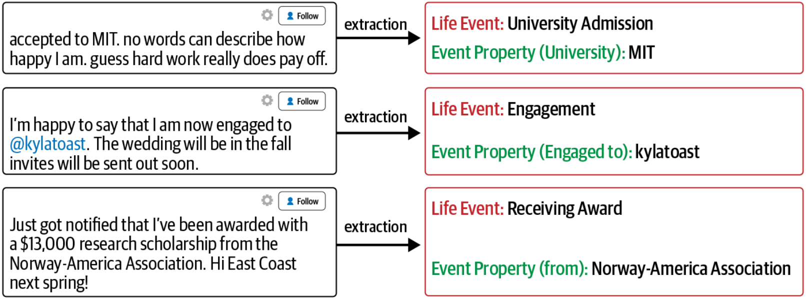
Read reviews that mention

easy to install | well made | works well | wall mount | mounting

bolts | bracket | instructions | bonne | solid | bedroom | inch

included | viewing

Event/Relation Extraction



accepted to MIT. no words can describe how happy I am. guess hard work really does pay off.

Life Event: University Admission

Event Property (University): MIT

I'm happy to say that I am now engaged to @kylatoast. The wedding will be in the fall invites will be sent out soon.

Life Event: Engagement

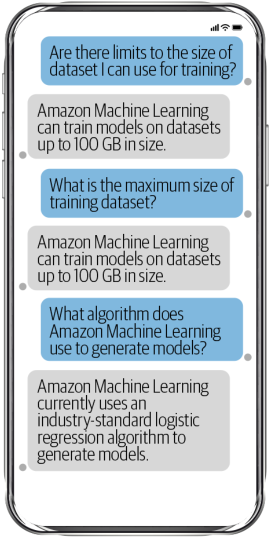
Event Property (Engaged to): kylatoast

Just got notified that I've been awarded with a $13,000 research scholarship from the Norway-America Association. Hi East Coast next spring!

Life Event: Receiving Award

Event Property (from): Norway-America Association

Chatbots



Are there limits to the size of dataset I can use for training?

Amazon Machine Learning can train models on datasets up to 100 GB in size.

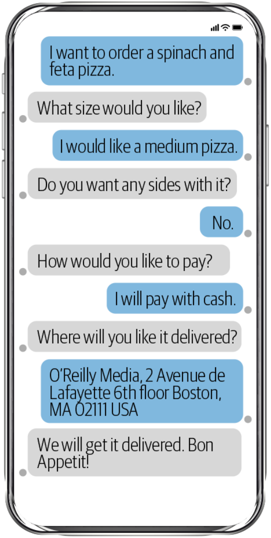
What is the maximum size of training dataset?

Amazon Machine Learning can train models on datasets up to 100 GB in size.

What algorithm does Amazon Machine Learning use to generate models?

Amazon Machine Learning currently uses an industry-standard logistic regression algorithm to generate models.

FAQ Bot



I want to order a spinach and feta pizza.

What size would you like?

I would like a medium pizza.

Do you want any sides with it?

No.

How would you like to pay?

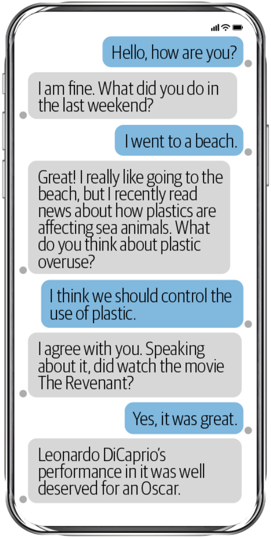
I will pay with cash.

Where will you like it delivered?

O'Reilly Media, 2 Avenue de Lafayette 6th floor Boston, MA 02111 USA

We will get it delivered. Bon Appetit!

Flow-Based Bot



Hello, how are you?

I am fine. What did you do in the last weekend?

I went to a beach.

Great! I really like going to the beach, but I recently read news about how plastics are affecting sea animals. What do you think about plastic overuse?

I think we should control the use of plastic.

I agree with you. Speaking about it, did watch the movie The Revenant?

Yes, it was great.

Leonardo DiCaprio's performance in it was well deserved for an Oscar.

Open-Ended Bot

Many more

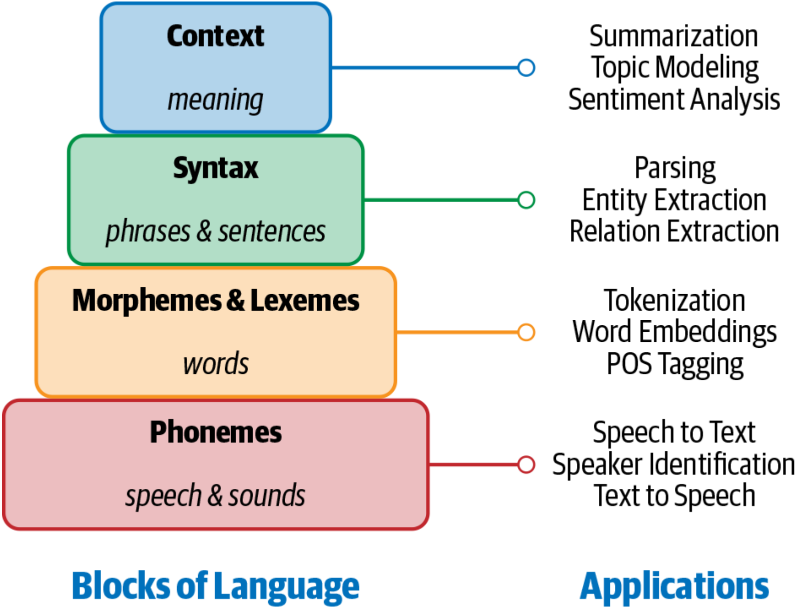
- text summarization
- text recommendation
- topic modeling
- text to speech/speech to text conversion
- language generation (e.g., automatically generating weather reports, chatbots, etc.)

.... and so on.

Outline

1. What is NLP?
2. The many faces of NLP
3. What makes NLP Challenging?
4. Some common NLP Tasks: An overview
5. **The levels of language processing: some examples**
6. Approaches to NLP

Levels of Language Processing



Context

meaning

Summarization
Topic Modeling
Sentiment Analysis

Syntax

phrases & sentences

Parsing
Entity Extraction
Relation Extraction

Morphemes & Lexemes

words

Tokenization
Word Embeddings
POS Tagging

Phonemes

speech & sounds

Speech to Text
Speaker Identification
Text to Speech

Blocks of Language

Applications

Levels of processing: speech/sounds

- What are the sounds of a language? How do they become words? How do we identify them automatically?
- Uses: Text to speech conversion, Speech to text conversion, pronunciation training etc
- Traditionally studied separately under Speech processing. Not a part of NLP courses generally.

Levels of processing: Word

- Morphological analysis of words
- compiling lists of words, or word sequences occurring in documents.
- Understanding/Modeling relationships between words etc.

Levels of processing: Sentence

POS Tagging

- Task: Given a sequence of words, return the POS tags for each word.
- An example problem: What is the best tag for a word in a context?
  - I wish to cite this work.
    PRP/I VBP/wish TO/to VB/cite DT/this NN/work ./.
  - He has a wish.
    PRP/He VBZ/has DT/a NN/wish ./.
- Largely considered solved for English, but there are still issues if we go beyond typical newspaper language (e.g., tagging speech or tweets). Still an unsolved problem for several languages.

Levels of processing: Sentence

Parsing

- Task: Construct the syntactic structure of a given sentence.
- Two kinds of trees can be generated in NLP: Phrase structure tree (Constituency tree), Dependency tree
- PST: shows parse structure in terms of Noun Phrases, Verb Phrases, Prep. Phrases etc.
- Dependency Tree: shows relations between words in a sentence in terms of a pre-defined set of relations
- Very active area of current research, for multiple languages.
- Important note: POS tagging errors can carry over and affect parser efficiency.

Levels of processing: Sentence

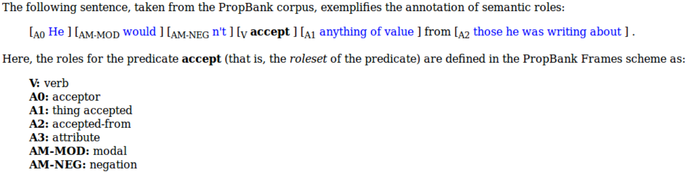
Word sense disambiguation

- Task: For words that can have multiple meanings, what is the right sense of the word in a given sentence?
- Example: "Let us go inside, it is cold" vs "I have cold and cough"
- Very important for applications such as machine translation, information retrieval
- Good progress for English WSD. One of the active areas of research in the field.

Levels of processing: Sentence

Semantic Role Labeling

- SRL is all about doing a "semantic parse" of a sentence. The task here is to identify argument structure of a sentence and thematic roles of different entities.
- Example: (source: http://www.cs.upc.edu/~srlconll/)



The following sentence, taken from the PropBank corpus, exemplifies the annotation of semantic roles:

[$_{\text{A0}}$ He] [$_{\text{AM-MOD}}$ would] [$_{\text{AM-NEG}}$ n't] [$_{\text{V}}$ **accept**] [$_{\text{A1}}$ anything of value] from [$_{\text{A2}}$ those he was writing about] .

Here, the roles for the predicate **accept** (that is, the *roleset* of the predicate) are defined in the PropBank Frames scheme as:

V: verb
A0: acceptor
A1: thing accepted
A2: accepted-from
A3: attribute
AM-MOD: modal
AM-NEG: negation

- Active area of research. Still hard, but making progress.

Beyond a sentence

- Given a text (more than one sentence), analyze the relationships between sentences, identify what pronouns refer to what nouns, how is the same entity referred in different ways (Barack Obama, Obama, The President and so on).
- What NLP methods are useful: coreference resolution, discourse parsing etc.
- Application: Text summarization, Question-Answering, Essay scoring etc.
- Hard problems, but active research topic and hence, making good progress

Questions so far?

Outline

1. What is NLP?
2. The many faces of NLP
3. What makes NLP Challenging?
4. Some common NLP Tasks: An overview
5. The levels of language processing: some examples
6. **Approaches to NLP**
7. Conclusion

Approaches to NLP

1. heuristics based NLP
2. machine learning
3. deep learning
4. combination of all or some of the above three

(More on this in the next class)

Heuristics

1. Word lists, lexicons etc.
2. Rule engineering based on observed patterns of language use.
3. Regular expressions

...

Machine Learning

- Machine learning is a collection of methods that learn a task automatically by looking at lots (and lots) of examples
- Usually involves a feature extraction step, where we need to specify what kind of patterns (e.g., words? heuristics? pos tags?) should a machine learn.
- Different kinds of machine learning algorithms are used in NLP - naive bayes, logistic regression (text classification), conditional random fields (sequence tagging), clustering etc.
- We will briefly see how some of them are useful for NLP as we progress with the course.

Deep Learning

...a sub-field of machine learning

- Deep learning is inspired by learning in humans and other biological forms, but "deep" refers to the number of layers in the artificial neural network of the deep learning method.

Deep Learning

...a sub-field of machine learning

- Deep learning is inspired by learning in humans and other biological forms, but "deep" refers to the number of layers in the artificial neural network of the deep learning method.
- Some popular deep learning architectures used in NLP in the past 5 years:
  - Convolutional Neural Networks are typically used in image processing related problems, and are useful to automatically extract features (words/word sequences) in NLP tasks.
  - Transformers (e.g., BERT) model the textual context, but not sequentially. It uses a phenomenon called **attention** to model the context.

Deep Learning

…a sub-field of machine learning

- Deep learning is inspired by learning in humans and other biological forms, but "deep" refers to the number of layers in the artificial neural network of the deep learning method.
- Some popular deep learning architectures used in NLP in the past 5 years:
    - Convolutional Neural Networks are typically used in image processing related problems, and are useful to automatically extract features (words/word sequences) in NLP tasks.
    - Transformers (e.g., BERT) model the textual context, but not sequentially. It uses a phenomenon called **attention** to model the context.

  Note: Transformers have become very popular in NLP in recent 2-3 years across tasks and languages.

Deep Learning and Transfer Learning

- Deep learning and Machine learning methods in general expect you have to have a lot of data for the problem you want to solve.
- Sometimes, we may have a lot of data on some related problem, but very small amount of data on a particular problem.
- Transfer learning is all about how to use existing data/knowledge for a new problem.

Transfer learning has also become a popular method in NLP in the recent 2-3 years.

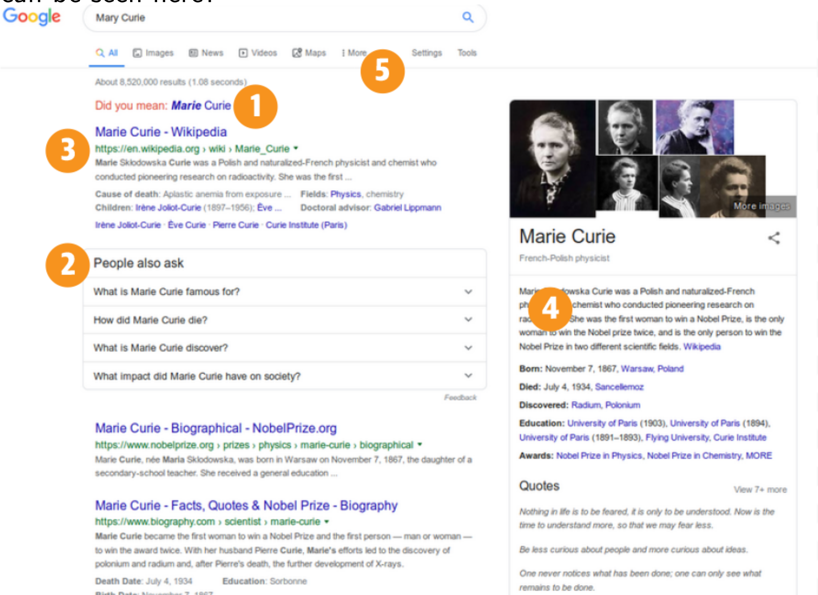
Questions so far?

Outline

1. What is NLP?
2. The many faces of NLP
3. What makes NLP Challenging?
4. Some common NLP Tasks: An overview
5. The levels of language processing: some examples
6. Approaches to NLP
7. **Conclusion**

A real world example: revisiting search

What levels of language processing are involved? What NLP tasks can be seen here?



About 8,520,000 results (1.08 seconds)

Did you mean: ***Marie*** Curie

Marie Curie - Wikipedia
https://en.wikipedia.org › wiki › Marie_Curie ▾
Marie Sklodowska **Curie** was a Polish and naturalized-French physicist and chemist who conducted pioneering research on radioactivity. She was the first ...

Cause of death: Aplastic anemia from exposure ... **Fields**: Physics, chemistry
Children: Irène Joliot-Curie (1897–1956); Ève ... **Doctoral advisor**: Gabriel Lippmann

Irène Joliot-Curie · Ève Curie · Pierre Curie · Curie Institute (Paris)

People also ask

What is Marie Curie famous for?

How did Marie Curie die?

What is Marie Curie discover?

What impact did Marie Curie have on society?

Marie Curie - Biographical - NobelPrize.org
https://www.nobelprize.org › prizes › physics › marie-curie › biographical ▾
Marie **Curie**, née **Maria** Sklodowska, was born in Warsaw on November 7, 1867, the daughter of a secondary-school teacher. She received a general education ...

Marie Curie - Facts, Quotes & Nobel Prize - Biography
https://www.biography.com › scientist › marie-curie ▾
Marie Curie became the first woman to win a Nobel Prize and the first person — man or woman — to win the award twice. With her husband Pierre **Curie**, **Marie's** efforts led to the discovery of polonium and radium and, after Pierre's death, the further development of X-rays.

Death Date: July 4, 1934 **Education**: Sorbonne
Birth Date: November 7, 1867

Marie Curie

French-Polish physicist

Mari dowska Curie was a Polish and naturalized-French ph chemist who conducted pioneering research on rad She was the first woman to win a Nobel Prize, is the only woman win the Nobel prize twice, and is the only person to win the Nobel Prize in two different scientific fields. Wikipedia

Born: November 7, 1867, Warsaw, Poland

Died: July 4, 1934, Sancellemoz

Discovered: Radium, Polonium

Education: University of Paris (1903), University of Paris (1894), University of Paris (1891–1893), Flying University, Curie Institute

Awards: Nobel Prize in Physics, Nobel Prize in Chemistry, MORE

Quotes View 7+ more

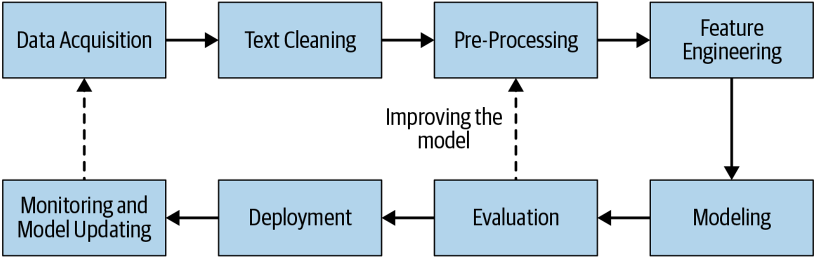
Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less.

Be less curious about people and more curious about ideas.

One never notices what has been done; one can only see what remains to be done.

So, how does one build NLP systems?

NLP Pipeline



(More on this in the next class!)

ToDo before next meeting

- Take a look at the syllabus document, understand the course requirements, deadlines etc.
- Read the introductory chapter of any of the textbooks mentioned above for NLP
- Check out the assignment descriptions
- Decide on a programming language you want to use and setup your laptops for that (e.g., installing any software needed etc)
- Read Chapter 2 of Practical NLP if you have access.
- Please introduce yourselves in the Forum called "Introductions"
- Ask any questions you have about today in the forum "NLP/Course Overview"

Questions so far?