Natural Language Processing and Information Extraction

2022 WS

Allan Hanbury
Florina Piroi
Gábor Recski
Ádám Kovács

TU Wien Informatics

Contents

- Course information
- Introduction to NLP

Course information

Informatics

Lecturers


Allan Hanbury


Florina Piroi


Gábor Recski


Ádám Kovács

Lecture Schedule

- Course Information, Introduction to NLP [Hanbury] (7.10.2022)
- Text Processing [Recski] (14.10.2022)
- Text Classification [Recski] (21.10.2022)
- Deep Learning for NLP [Piroi] (28.10.2022)
- Textual Sequence Modelling & Attention [Piroi] (4.11.2022)
- Deep Learning – Practical Lesson [Kovács] (11.11.2022)
- Syntax (Constituency and Dependency) [Recski] (18.11.2022)
- Basic (non-DL) Semantics [Recski] (25.11.2022)
- Information Extraction [Recski] (2.12.2022)
- Summarisation & Keyword Extraction [Piroi] (9.12.2022)
- Annotation Basics and Challenges [Hanbury] (16.12.2022)
- Question Answering and Chatbots [Hanbury] (13.1.2023)
- Project Presentations (20.1.2023)

Lectures

- Fridays 13:00 c.t. - 15:00
- EI11

Exercise

- One project exercise with two Milestones
- Done in groups of four
- Each group has a mentor
- Submissions are made via GitHub Classrooms
- Grading is based on milestones, final submission, presentation, and report
- Every group member must present their own contributions in the final presentation and will be individually evaluated on these contributions

Exercise Deadlines

- (Oct 14: topic list final)
- Oct 18: deadline for topic selection
- Oct 21: topics assigned, project milestone 1 introduced
- Nov 11: milestone 1 deadline, milestone 2 introduced
- Dec 2: milestone 2 deadline
- Jan 20: final presentations
- Jan 27: final submission deadline

Effort Breakdown

- Lectures: 24 hours
- Project Milestone 1: 8 hours
- Project Milestone 2: 8 hours
- Final Solution: 35 hours
- **Total: 75 hours**

Performance Evaluation

- Milestone 1: Minimum 35%
- Milestone 2: Minimum 35%
- Final solution: Minimum 35%

- Overall Score: Minimum 50% to pass
  - 15% for Milestone 1
  - 15% for Milestone 2
  - 50% for the final solution
  - 10% for the presentation
  - 10% for the management summary

- There is no exam!

- Marks Overall Score

Marks	Overall Score
1	89 – 100
2	76 – 88
3	63 – 75
4	50 – 62

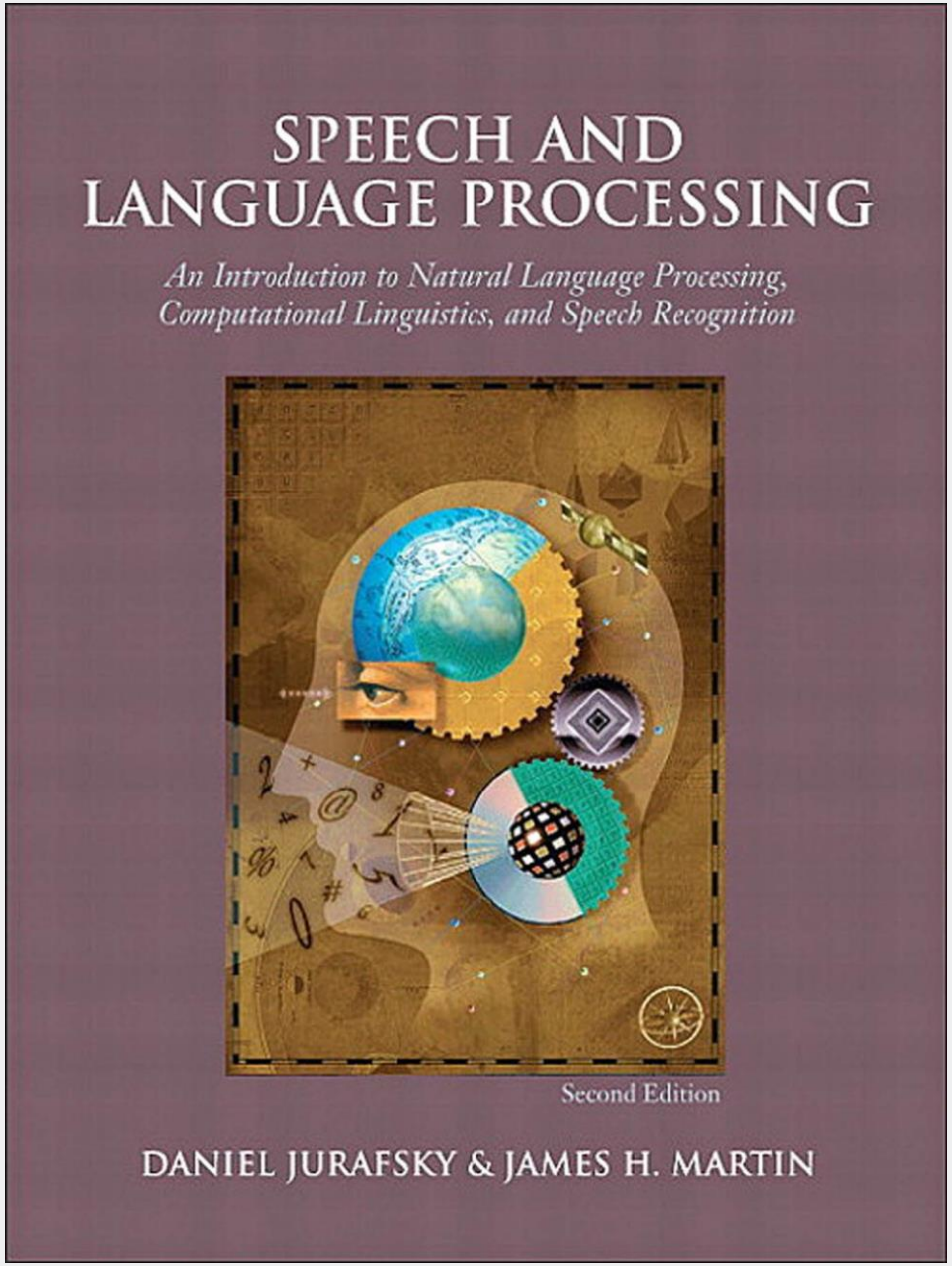
Organisation

- Course
  - Please register for the course in TISS
- Communication
  - Use the General Discussion Forum in TUWEL for questions, not the TISS forum
- The schedule of lectures and all course material will be available on Github

Book

Third edition in preparation – download many chapters here:

https://web.stanford.edu/~jurafsky/slp3/



Questions about the organisation, etc.

Ask now!

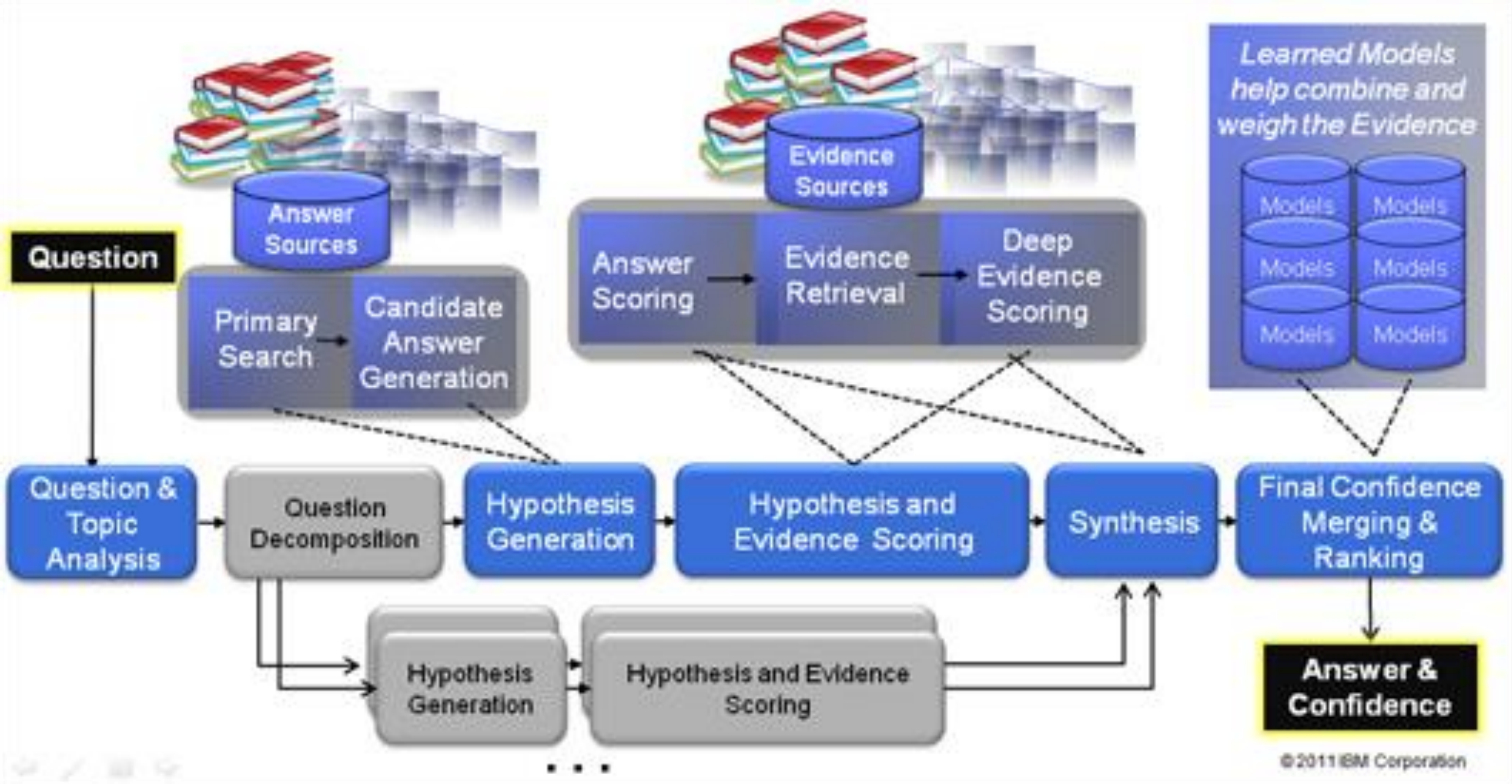
Introduction to NLP

IBM Watson and Jeopardy

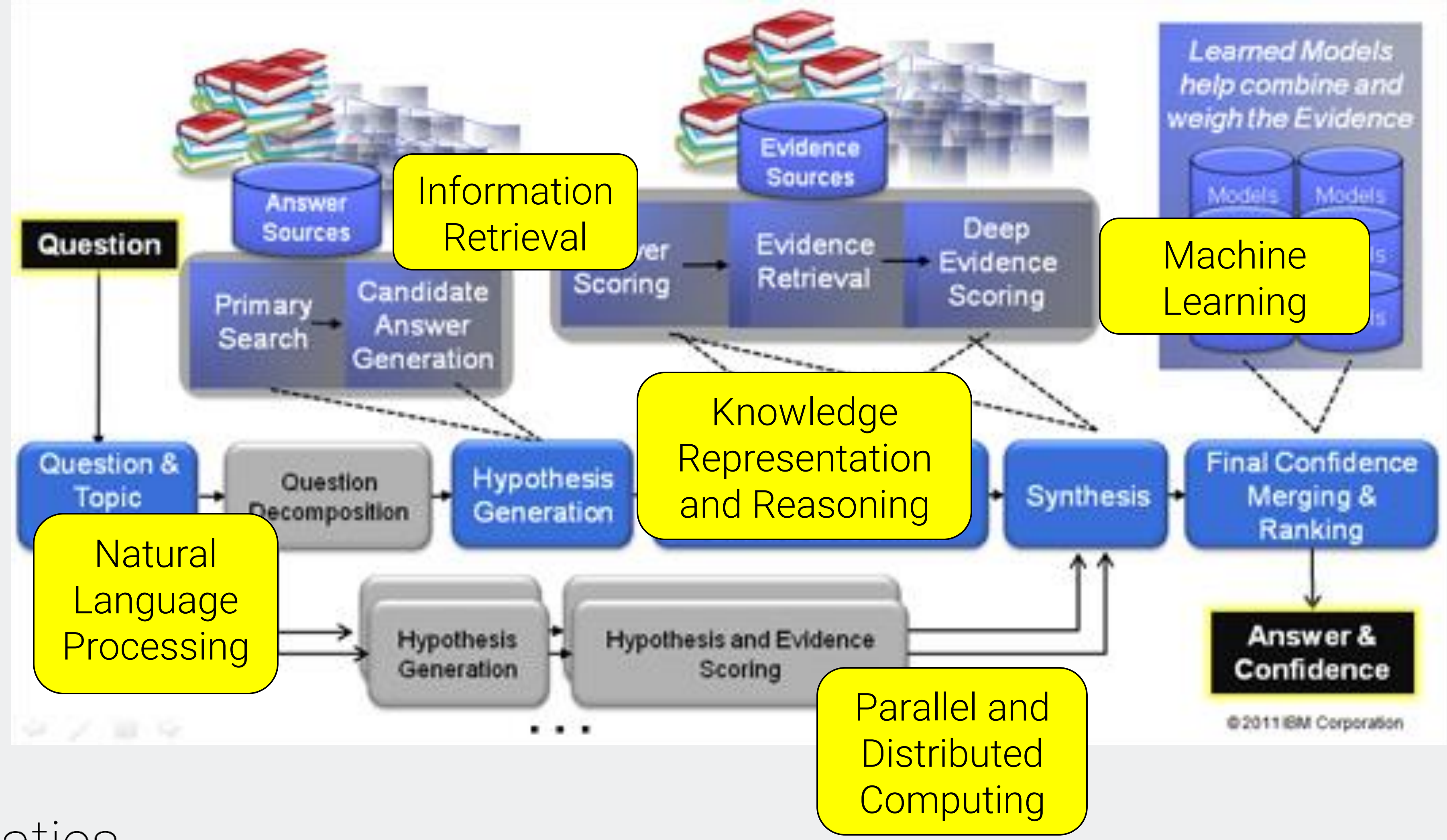
Final question: https://www.youtube.com/watch?v=Sp4q60BsHoY
IBM film: https://www.youtube.com/watch?v=P18EdAKuC1U

The end of the show

How does Watson (for Jeopardy) work?



How does Watson (for Jeopardy) work?



IBM

Search

Watson Health Products Solutions Services Support Learn Explore more

Go beyond human capabilities

Combine the expertise of humans with the power of technology to deliver patient-centric cancer care at scale

PDF Transform cancer care (1.8 MB)

Data, AI & analytics (01:28)

Our goals

For the physician

Success stories

Solutions

The changing oncology landscape

New intelligent analytics and workflow technologies hold the key to overcoming a primary challenge of providing personalized approaches to cancer care—harnessing the vast amounts of data available without being overwhelmed by it. Clinicians are faced with large, heterogeneous, and complex data sets when making patient-specific clinical decisions. Oncology solutions backed by AI and machine learning provide a powerful tool by bringing together data, extracting insights and presenting it to providers for their evaluation.

PDF Read the 2020 Data and Evidence Booklet (10.2 MB)

From research to real world.

https://www.ibm.com/watson-health/oncology-and-genomics

IEEE SPECTRUM

Feature | Biomedical | Diagnostics

02 Apr 2019 | 15:00 GMT

How IBM Watson Overpromised and Underdelivered on AI Health Care

After its triumph on *Jeopardy!*, IBM's AI seemed poised to revolutionize medicine. Doctors are still waiting

By **Eliza Strickland**

https://spectrum.ieee.org/biomedical/diagnostics/how-ibm-watson-overpromised-and-underdelivered-on-ai-health-care

TECHNOLOGY NETWORKS

Exploring the Science That Matters to You

The Hype of Watson: Why Hasn't AI Taken Over Oncology?

ARTICLE Apr 17, 2020 | by Sylvia He

Doctors' notes are one of the obstacles in the way of AI becoming a major force in oncology.

https://www.technologynetworks.com/informatics/articles/the-hype-of-watson-why-hasnt-ai-taken-over-oncology-333571

Natural Language Processing (NLP)

Natural language processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data.

wikipedia

Why is NLP interesting?

- Languages involve many human activities
  - Reading, writing, speaking, listening
- Voice can be used as an user interface in many applications
  - Remote controls, virtual assistants like siri,...
- NLP is used to acquire insights from massive amount of textual data
  - E.g., hypotheses from medical & health reports
- NLP has many applications
- NLP is difficult!

https://www.slideshare.net/minhpqn/introduction-to-natural-language-processing-67212472

Why is NLP difficult?

I made her duck

- I cooked waterfowl for her.
- I cooked waterfowl belonging to her.
- I created the (plaster?) duck she owns.
- I caused her to quickly lower her head or body.
- I waved my magic wand and turned her into undifferentiated waterfowl.

Jurafsky & Martin Book

I shot an elephant in my pyjamas.

https://freesvg.org/elephant-in-pajamas

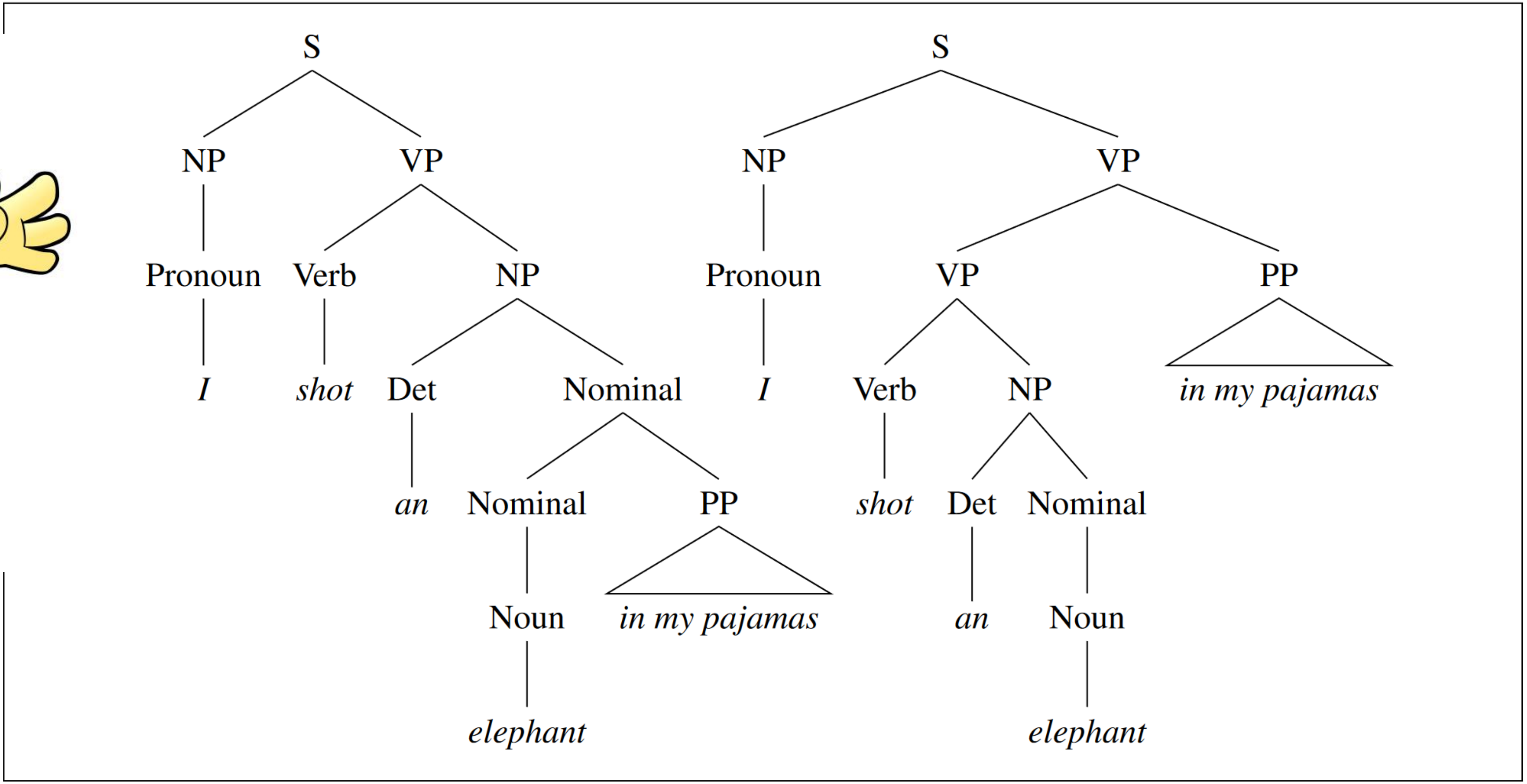


Figure 13.2 Two parse trees for an ambiguous sentence. The parse on the left corresponds to the humorous reading in which the elephant is in the pajamas, the parse on the right corresponds to the reading in which Captain Spaulding did the shooting in his pajamas.

Jurafsky & Martin Book

Why is NLP difficult?

Natural Languages are generally ambiguous

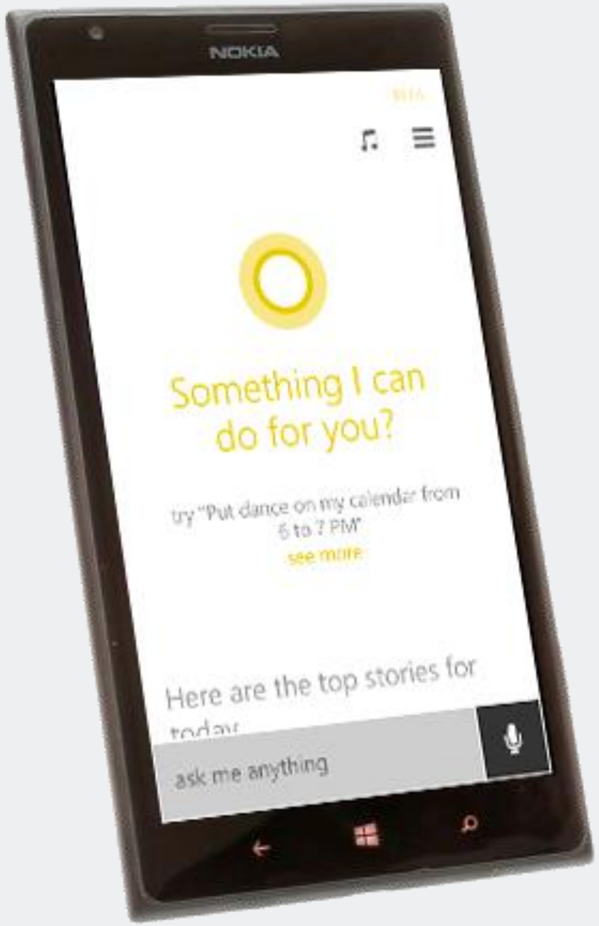
Various levels of knowledge of a language must be considered:

- **Phonetics and Phonology** — knowledge about linguistic sounds
- **Morphology** — knowledge of the meaningful components of words
  - I am $\rightarrow$ I'm, forms for singular and plural (door/doors)
- **Syntax** — knowledge of the structural relationships between words, needed to order and group words
- **Semantics** — knowledge of meaning
  - What is meant by "export" and "expert"? What constitutes "Western Europe"?
- **Pragmatics** — knowledge of the relationship of meaning to the goals and intentions of the speaker
  - Is it a request, question or a statement?
- **Discourse** — knowledge about linguistic units larger than a single utterance
  - Reference to the context given by e.g. multiple sentences.
  - E.g. In what year was Lincoln born? How many states were in the United States in *that year*?

Very Brief History of NLP

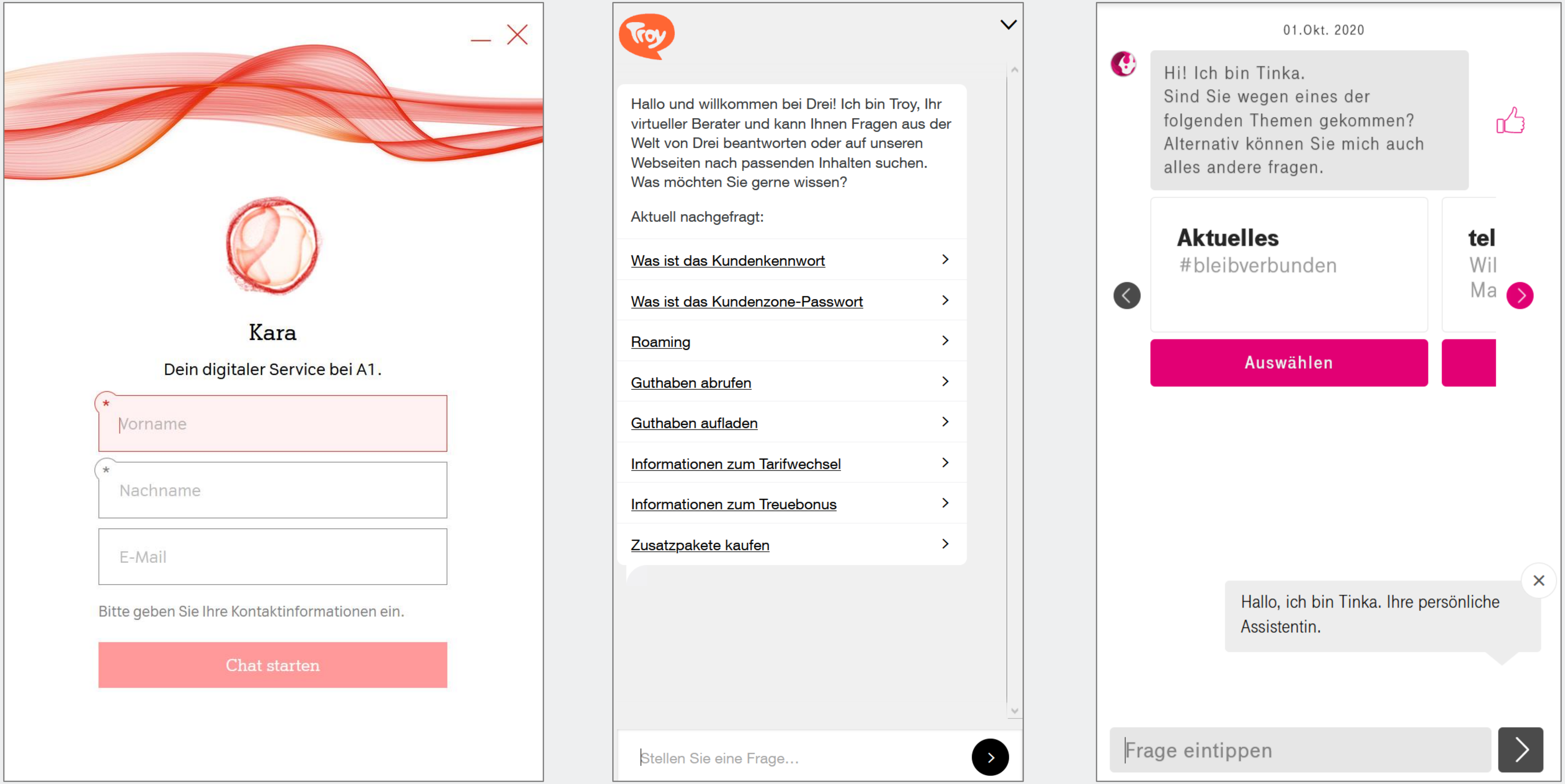
- Foundational Insights: 1940s and 1950s
- Generally two paradigms:
  - Symbolic Paradigm
  - Stochastic Paradigm
- The Rise of Machine Learning: 2000-now
  - Large amount of spoken and textual data become available
  - Widespread availability of high-performance computing systems
- The Domination of Neural Approaches: ~2015-now

Dialogue systems



TU WIEN Informatics

https://upload.wikimedia.org/wikipedia/commons/4/44/IOS_7_Siri.png https://commons.wikimedia.org/wiki/File:Cortana_start_screen_(yellow-white).png https://commons.wikimedia.org/wiki/File:Amazon_Alexa_-_Echo_Dot_(49675526278).jpg



Kara
Dein digitaler Service bei A1.
Vorname
Nachname
E-Mail
Bitte geben Sie Ihre Kontaktinformationen ein.
Chat starten
Hallo und willkommen bei Drei! Ich bin Troy, Ihr virtueller Berater und kann Ihnen Fragen aus der Welt von Drei beantworten oder auf unseren Webseiten nach passenden Inhalten suchen. Was möchten Sie gerne wissen?
Aktuell nachgefragt:
Was ist das Kundenkennwort
Was ist das Kundenzone-Passwort
Roaming
Guthaben abrufen
Guthaben aufladen
Informationen zum Tarifwechsel
Informationen zum Treuebonus
Zusatzpakete kaufen
Stellen Sie eine Frage...
01.Okt. 2020
Hi! Ich bin Tinka.
Sind Sie wegen eines der folgenden Themen gekommen? Alternativ können Sie mich auch alles andere fragen.
Aktuelles
#bleibverbunden
Auswählen
Hallo, ich bin Tinka. Ihre persönliche Assistentin.
Frage eintippen

Training Data Bias…

Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day

68

by James Vincent | @jjvincent | Mar 24, 2016, 6:43am EDT

SHARE TWEET LINKEDIN

https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist

Project Debater

Highlights: https://www.youtube.com/watch?v=nJXcFtY9cWY
Full debate: https://www.youtube.com/watch?v=-d4Uj9ViP9o

https://www.research.ibm.com/artificial-intelligence/project-debater/live/images/leadspace-livedebate.jpg

Argument Mining

Claims and evidence are the main components of an argument; identifying and using them correctly are essential to framing an argument in a debate. The IBM Project Debater team has invested substantial effort in developing machine learning techniques to mine massive corpora for claims and evidence and use them to generate arguments relevant to a controversial topic.

- Detecting claims in relevant documents
- Detecting evidence in relevant documents
- Negating claims
- Synthesizing novel claims
- Detecting claims throughout a corpus
- Improving corpus-wide claim detection
- Assessing argumentation quality
- Relating arguments across texts

https://www.research.ibm.com/artificial-intelligence/project-debater/how-it-works/

Stance Classification and Sentiment Analysis

An automatic debating system must be able to identify whether an argument supports or contests a given topic. This is fairly easy for humans but difficult for machines, as it requires great sensitivity to the rich subtleties and nuances of natural language. We have made important progress in this intriguing line of research.

https://www.research.ibm.com/artificial-intelligence/project-debater/how-it-works/

Determining expert opinion stance

Determining claim stance

Improving claim stance classification

Classifying sentiment of phrases

Classifying sentiment of idioms

Deep Neural Nets (DNNs) and Weak Supervision

DNNs hold immense potential for improving automatic understanding of language, but training them is notoriously known to require a lot of high quality, manually labeled data. We developed tools and methods to train DNNs using weak supervision, alleviating that bottleneck. We also used DNNs in developing Project Debater's speaking and listening skills.

Scoring arguments

Understanding Automatic Speech Recognition (ASR) output

Predicting phrase breaks

Emphasizing words and phrases

Improving speech patterns

Identifying similar sentences

Improving argument mining

Searching for claims throughout a corpus

https://www.research.ibm.com/artificial-intelligence/project-debater/how-it-works/

GPT-3

Language model with 175 billion parameters

[cs.CL] 22 Jul 2020

Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*

Jared Kaplan† Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry

Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan

Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter

Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray

Benjamin Chess Jack Clark Christopher Berner

Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

OpenAI

Training datasets

Dataset	Quantity (tokens)
Common Crawl (filtered)	410 billion
WebText2	19 billion
Books1	12 billion
Books2	55 billion
Wikipedia	3 billion

https://arxiv.org/abs/2005.14165

Correcting English

Poor English input: I eated the purple berries.
Good English output: I ate the purple berries.
Poor English input: Thank you for picking me as your designer. I'd appreciate it.
Good English output: Thank you for choosing me as your designer. I appreciate it.
Poor English input: The mentioned changes have done. or I did the alteration that you requested. or I changed things you wanted and did the modifications.
Good English output: The requested changes have been made. or I made the alteration that you requested. or I changed things you wanted and made the modifications.
Poor English input: I'd be more than happy to work with you in another project.
Good English output: I'd be more than happy to work with you on another project.

Poor English input: Please provide me with a short brief of the design you're looking for and that'd be nice if you could share some examples or project you did before.
Good English output: Please provide me with a brief description of the design you're looking for and that would be nice if you could share some examples or projects you have done before.

Poor English input: The patient was died.
Good English output: The patient died.

Poor English input: We think that Leslie likes ourselves.
Good English output: We think that Leslie likes us.

Poor English input: Janet broke Bill on the finger.
Good English output: Janet broke Bill's finger.

Poor English input: Mary arranged for, in St. Louis, John to rent a house cheap.
Good English output: Mary arranged for John to rent a house in St. Louis.

Poor English input: Today I have went to the store to to buys some many bottle of water.
Good English output: Today I went to the store to buy some bottles of water.

News Article Generation

Title: United Methodists Agree to Historic Split
Subtitle: Those who oppose gay marriage will form their own denomination
Article: After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church's annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination.
The Post notes that the denomination, which claims 12.5 million members, was in the early 20th century the "largest Protestant denomination in the U.S.," but that it has been shrinking in recent decades. The new split will be the second in the church's history. The first occurred in 1968, when roughly 10 percent of the denomination left to form the Evangelical United Brethren Church. The Post notes that the proposed split "comes at a critical time for the church, which has been losing members for years," which has been "pushed toward the brink of a schism over the role of LGBTQ people in the church." Gay marriage is not the only issue that has divided the church. In 2016, the denomination was split over ordination of transgender clergy, with the North Pacific regional conference voting to ban them from serving as clergy, and the South Pacific regional conference voting to allow them.

Figure 3.14: The GPT-3 generated news article that humans had the greatest difficulty distinguishing from a human written article (accuracy: 12%).

Title: Star's Tux Promise Draws Megyn Kelly's Sarcasm
Subtitle: Joaquin Phoenix pledged to not change for each awards event
Article: A year ago, Joaquin Phoenix made headlines when he appeared on the red carpet at the Golden Globes wearing a tuxedo with a paper bag over his head that read, "I am a shape-shifter. I can't change the world. I can only change myself." It was a promise to not change to fit into the Hollywood mold: "I think that's a really special thing, to not change yourself. I think it's a really special thing to say, 'This is what's inside of me, I'm proud of it, and I'm not going to be ashamed because of the way that someone else thinks I should be.'" Now, it's the Oscars, and Phoenix is at it again. But this time, his publicist is saying he'll be wearing a tux no matter what.
Megyn Kelly was not impressed, and she let him have it on The Tonight Show. "You know, I feel like, I feel like you could have worn the tux," she says. "But you're saying you're a shape-shifter. I don't know if you can change your tux, but you can change your mind. You can change your mind. You can change your mind." Phoenix says he did, but it didn't stick. "I was like, 'Okay, I'm going to wear a tuxedo to this thing.' And then I thought, 'I don't want to wear a tuxedo to this thing.'" Kelly goes on to encourage him to change his mind again, but Phoenix says it's too late: "I'm committed to wearing this."

Figure 3.15: The GPT-3 generated news article that humans found the easiest to distinguish from a human written article (accuracy: 61%).

Training Data Bias…

Abubakar Abid
@abidlabs

I'm shocked how hard it is to generate text about Muslims from GPT-3 that has nothing to do with violence… or being killed…

https://twitter.com/i/status/1291165311329341440

NLP and Climate Change

Energy and Policy Considerations for Deep Learning in NLP

Emma Strubell **Ananya Ganesh** **Andrew McCallum**
College of Information and Computer Sciences
University of Massachusetts Amherst
{strubell, aganesh, mccallum}@cs.umass.edu

June 2019

Consumption	**CO_2e (lbs)**
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155

Table 1: Estimated CO_2 emissions from training common NLP models, compared to familiar consumption.[1]

https://arxiv.org/abs/1906.02243

Summary

- Neural approaches are big in NLP at the moment
- Beware of bias
- NLP can be bad for the climate