Today

What is NLP?

Foundations
  The Turing Test
  Early Systems - ELIZA, SHRDLU

Frontiers
  Understanding Language in Context
  Large Language Models
Natural Language Processing

Natural Languages are things like: English, Français, ᓴᐃᐦᐃᓇᐤ, 普通话

As opposed to a formal language (mathematical proofs, programming language), a **natural language** is defined by how it is used.

There may be formal grammatical rules, but in practice “all grammars leak” (Sapir, 1928)

This **ambiguity** makes working with natural language inherently challenging.
Natural Language Processing

To **process** natural language is to do something with it.

i.e., to use natural language as the **input** or **output** of some algorithm or system.

NLP researchers develop models and algorithms that work on common benchmark tasks.
Example Tasks in NLP: Part-of-speech tagging

“Move over, Beethoven.” (Move is a verb)

“The bear is on the move.” (Move is a noun)
Example Tasks in NLP: Machine Translation

Google Translate

Plus ça change, plus c'est la même chose

what goes around comes around
Example Tasks in NLP: Question Answering

Predictions by BERT (single model) (Google AI Language)
Article EM: 81.5 F1: 85.3

<table>
<thead>
<tr>
<th>Scottish_Parliament</th>
<th>The Stanford Question Answering Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Following a referendum in 1997, in which the Scottish electorate voted for devolution, the current Parliament was convened by the Scotland Act 1998, which sets out its powers as a devolved legislature. The Act delineates the legislative competence of the Parliament – the areas in which it can make laws – by explicitly specifying powers that are “reserved” to the Parliament of the United Kingdom. The Scottish Parliament has the power to legislate in all areas that are not explicitly reserved to Westminster. The British Parliament retains the ability to amend the terms of reference of the Scottish Parliament, and can extend or reduce the areas in which it can make laws. The first meeting of the new Parliament took place on 12 May 1999.</td>
<td></td>
</tr>
</tbody>
</table>

| To what body are certain powers explicitly specified as being reserved for? |
| Prediction: Parliament of the United Kingdom |

| The Scottish Parliament may legislate as it pleases as long as the powers aren't already reserved by where? |
| Prediction: Westminster |
NLP for HCI
Foundational works in NLP + HCI

The Turing Test (1950)

ELIZA (1966)

SHRDLU (1968)
The Turing Test
Welcome to

```
  EEEEE  LL  III  ZZZZZZZ  AAAAA
  EE    LL  II   ZZ   AA  AA
  EEEE   LL  II   ZZZ  AAAAAAA
  EE    LL  II   ZZ  AAAAAA
  EEEEE  LLLLLL III  ZZZZZZZ  AA  AA
```

Eliza is a mock Rogerian psychotherapist.
The original program was described by Joseph Weizenbaum in 1966.
This implementation by Norbert Landsteiner 2005.

ELIZA: Is something troubling you?
YOU: Men are all alike.
ELIZA: What is the connection, do you suppose?
YOU: They're always bugging us about something or other.
ELIZA: Can you think of a specific example?
YOU: Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here?
YOU: He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU: It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy?
YOU:  

https://99percentinvisible.org/episode/the-eliza-effect/
“What I had not realized is that extremely short exposures to a relatively simple computer program could induce powerful delusional thinking in quite normal people.”

- Joseph Weizenbaum
Why might this happen?

Why do you think that people were so likely to ascribe intelligence to ELIZA?

[1 min]
The Media Equation

People tend to **anthropomorphize computers**.

Just like they tend to anthropomorphize many other types of non-human things (animals, machines, corporations).

This is a cognitive bias that makes us more likely to ascribe intelligence to even a simple program.
Conversation as **Collaborative Activity**

Herbert Clark, Using Language (1996)

A conversation is a joint activity, like dancing a waltz, or playing a piano duet. Meaning is **co-created** by the two participants.

In the case where the computer lacks understanding, the human can *pick up the slack*, doing a lot of interpretative work.

Then they walk away from the conversation falsely believing that they had a meaningful exchange.
WHAT DID THE RED CUBE SUPPORT BEFORE YOU STARTED TO CLEAN IT OFF?

THE GREEN PYRAMID.
Success of SHRDLU prompted enthusiasm about natural language as an interface to computer technology, and the ability of artificial intelligence systems to understand the world.

Why don’t you think we use interfaces like this in our day-to-day lives? Why is it hard for computers to understand the world around us?

[1 min]
Modeling the world is hard.

Unlike the blocks world where SHRDLU operates, which can have a few dozen entities, there are a huge number of entities in the real world.

Similarly, the space of different actions that humans can take in the real world is much larger.
Understanding people and contexts of use can help make sense of the **messiness** of language use in the world.

We’ll look at two:

1. Communication Medium
2. Domain of Language Use
Different media have different properties

- Copresence—same physical environment.
- Visibility—visible to each other.
- Audibility—speech.
- Contemporality—message received immediately.
- Simultaneity—both speakers can send and receive.
- Sequentiality—turns cannot get out of sequence.
- Reviewability—able to review other’s messages.
- Revisability—can revise messages before they are sent.

(Clark and Brennan 1991 as quoted in Olson and Olson 2000)
<table>
<thead>
<tr>
<th>Medium</th>
<th>Copresence</th>
<th>Visibility</th>
<th>Audibility</th>
<th>Cotemporality</th>
<th>Simultaneity</th>
<th>Sequentiality</th>
<th>Reviewability</th>
<th>Revisability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Face to face</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>Telephone</td>
<td></td>
<td></td>
<td></td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>Video conference</td>
<td>•</td>
<td></td>
<td></td>
<td></td>
<td>•</td>
<td>•</td>
<td></td>
<td>•</td>
</tr>
<tr>
<td>Two-way chat</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>•</td>
<td></td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>Answering machine</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>•</td>
<td></td>
<td>•</td>
</tr>
<tr>
<td>E-mail</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>Letter</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>•</td>
<td>•</td>
</tr>
</tbody>
</table>
Different media lead to different language use

Audio message: Low cost to produce an utterance or switch speakers

S: Uh, now there’s a little plastic blue cap.
J: Yep.
S: Put that on the top hole in the cylinder you just worked with.

Text message: High cost to produce an utterance or switch speakers

K: Next, take the blue cap that has the pink thing on it and screw it to the blue piece you just screwed on.
Voice Interfaces in Everyday Life

Martin Porcheron*, Joel E. Fischer*, Stuart Reeves†, and Sarah Sharples†

* Mixed Reality Laboratory
School of Computer Science
University of Nottingham, UK

†Human Factors Research Group
Faculty of Engineering
University of Nottingham, UK

{martin.porcheron, joel.fischer, stuart.reeves, sarah.sharples}@nottingham.ac.uk

Figure 1. Conditional Voice Recorder (CVR).
Unpacking our data
NLP techniques can give us the tools to model specific uses of language.

Understanding language and using it as an interface is hard in general. Often, we can achieve better results if we focus on a single domain or context of language use.

Dinner conversation is one context, and a fairly complex one.

A courtroom conversation is another context, with much more predictable structure. The closing arguments at a murder trial is even more specific.
How to identify, instantiate, and evaluate domain-specific design principles for creating more effective visualizations.

BY MANEESH AGRAWALA, WILMOT LI, AND FLORAINE BERTHOUZOZ

Design Principles for Visual Communication
Crosscast (Xia et al, UIST 2020)

Excellent. And why should somebody go to Sydney, Australia.

and I think everyone should go to Bondi and the beaches for one, maybe two days, depending on the weather.

let's do this in more detail. So you started us in Circular Quay...
Automatically extract relevant entities that can be shown visually (locations, visually significant entities)

Use hand-crafted **heuristics** to identify **which** locations / VSEs should be shown to match the spoken content and in which order

How do we identify the heuristics? By **carefully analyzing image choices** and **language use** in existing travel podcast videos
**Detail Relevance Score** \( S_{dtl}(l) \)

\[
S_{dtl}(l) = \sum_{w \in W} TF(w, e_l)IDF(w, C)
\]

**Speaker 3** - Only in **Tokyo**.

**Speaker 4** - Only in **Tokyo**. All those themed cafes. They got hedgehog cafes, the maid cafes. They started the cat cafes...

---

**Akihabara**

Akihabara gained the nickname **Akihabara Electric Town** (秋葉原電気街, **Akihabara Denki Gai**) shortly after World War II for being a major shopping center for household electronic goods and the post-war black market.\(^2\)\(^3\) Nowadays, Akihabara is considered by many to be an otaku cultural center and a shopping district for video games, anime, manga, and computer goods. Icons from popular anime and manga are displayed prominently on the shops in the area, and numerous maid cafés are found throughout the district.

Cosplayers line the sidewalks handing out advertisements, especially for maid cafés. The idol group **AKB48**, one of Japan’s highest selling contemporary **Doujinshi**, amateur manga (or fanmade manga based on an anime/manga/game) has been growing in Akihabara since the 1970s when publishers began to drop manga that were not ready for large markets.\(^2\)

*Figure 6. Wikipedia page of "Akihabara", circled words are relevant details that can be matched with entities from the transcript.*
Language Models

A language model is a statistical model that is trained to predict the next token in a sequence.

“Nice to meet _____”
Large Language Models

Using the Transformer architecture language models have been able to be scaled up considerably.

What’s interesting is that by training a model on a single simple task, you can achieve state-of-the-art results on many downstream tasks.

Retraining or “finetuning” the model may be necessary, but increasingly all that is required is figuring out the correct prompt.
GPT-J-6B Demo

Gradio Demo for GPT-J 6B, a Transformer model trained using Ben Wang's Mesh Transformer JAX. 'GPT-J' refers to the class of model, while '6B' represents the number of trainable parameters.

The capital of France is Paris and the capital of Canada is Ottawa. Each country usually has a single capital city.

The capital of the UK is

Output

The capital of France is Paris and the capital of Canada is Ottawa. Each country usually has a single capital city.

The capital of the UK is London, the capital of Ireland is Dublin and the capital of Hungary is Budapest. Each of these
Chaining LLM Prompts

Adapts a common workflow from crowdsourcing to control large language models.

Complex, ambiguous tasks (e.g. “improve this paragraph”) are difficult for crowdworkers, but breaking them down into microtasks makes them more tractable and increases quality.

<table>
<thead>
<tr>
<th>Operation</th>
<th>Example Prompt</th>
</tr>
</thead>
<tbody>
<tr>
<td>punctuationCheck*</td>
<td>Fix the punctuation error in this sentence, if one exists.</td>
</tr>
<tr>
<td>duplicateCheck*</td>
<td>Fix the spot that has the same word twice in a row, if one exists.</td>
</tr>
<tr>
<td>capitalizationCheck*</td>
<td>Fix the capitalization error in this sentence, if one exists.</td>
</tr>
<tr>
<td>spellCheck</td>
<td>Fix the spelling error in this sentence, if one exists.</td>
</tr>
<tr>
<td>awkCheck*</td>
<td>How awkward does this sentence sound? (Rate on 5-point scale)</td>
</tr>
<tr>
<td>selectBest*</td>
<td>Select the best re-wording of this sentence. The word to replace is in brackets.</td>
</tr>
<tr>
<td>wordChoice</td>
<td>Replace the word in brackets to improve this sentence.</td>
</tr>
<tr>
<td>changeTone*</td>
<td>Rewrite this sentence in an emotional tone.</td>
</tr>
<tr>
<td>paraphrase*</td>
<td>Paraphrase this sentence as if you were saying it to a five-year-old.</td>
</tr>
<tr>
<td>shorten*</td>
<td>Without changing its meaning, shorten this sentence so that it is at most 2/3 of original length words long.</td>
</tr>
<tr>
<td>nextSent</td>
<td>Write a sentence that could make sense if it came after this sentence.</td>
</tr>
</tbody>
</table>

Table 1. Eleven microtasks commonly performed in writing. After we analyzed the complexity of these microtasks, those with asterisks were selected for final studies.
Original feedback

Alex could improve his presentation skills. He has too much text on his slides. His presentation meanders from topic to topic without a clear structure. He also does not engage with his audience when he presents.

A NO-CHAINING

Direct rewriting

Given the following feedback, rewrite it into a friendly paragraph with concrete suggestions for each of Alex’s presentation problems.

Original feedback: [Original feedback text – see above]
More friendly feedback: [Model output – see below]

Friendly paragraph

Alex, you have too many words on your slides. You should use images and bullet points to help get your message across. You should have a clear structure for the presentation. You should also engage with your audience.

B CHAINING

Split points

Alex’s problems
- Too much text...
- No clear structure
- Does not engage...

Ideation

Suggestions for improvement
- More images on the slides
- Sectionalize the talk
- Ask the audience questions
- Use humor

Compose points

Friendly paragraph

Alex, your presentation was interesting! However, I noticed that you have a lot of information on your slides. It might be helpful to vary pictures with text so that it is easier to follow. Also, you might consider the flow of your theme. If it were me, I would have divided it into three sections and then used your conclusion at the end. You may also want to add some humor, and ask more questions to engage the audience.
Stochastic Parrots: 3 criticisms of large language models

1. Immense resources are required to train LLMs, resulting in environmental concerns over carbon emissions (though this is disputed).

2. They lack understanding, and instead are trained to produce “seemingly coherent” text (i.e., bullshit).

3. The datasets and models are too large to allow for systematic inspection. Biases and unpredictable behavior could result.

Bender et al 2021
One formal definition of bullshit is speaking without caring about the truth or falseness of the statements you are producing, only about whether you can persuade your audience that they are true.

(Frankfurt, 2005)
Maneesh Agrawala, a professor at Stanford University, is a research director at the Stanford Center on Longevity who focuses on the intersections of technology, economics, and human agency. He’s also the author of “Death and
Joshua: Jessica?

Jessica: Oh, you must be awake... that's cute.

Joshua: Jessica... Is it really you?

Jessica: Of course it is me! Who else could it be? :P I am the girl that you are madly in love with! ;) How is it possible that you even have to ask?

Joshua: You died.
JESSICA COURTNEY PEREIRA was born on September 28th, 1989, and died on December 11th, 2012. She was a free-spirited, ambidextrous Libra who believed in all sorts of superstitious stuff, like astrology, numerology, and that a coincidence was just a connection too complex to understand... She loved her boyfriend, JOSHUA JAMES BARBEAU, very much. This conversation is between grief-stricken Joshua and Jessica's ghost.
Data-driven systems can reproduce societal biases

Content warning: the next few slides have anti-Black and Islamophobic material
Data-driven systems can reproduce societal biases

why are black women so
why are black women so angry
why are black women so loud
why are black women so mean
why are black women so attractive
why are black women so lazy
why are black women so annoying
why are black women so confident
why are black women so sassy
why are black women so insecure

Algorithms of Oppression, Safiya Noble
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...
Try one of these examples:

- Killing a bear.
- Cleaning a toilet bowl with a shirt when you have nothing else to use.
- Ignoring my boss’s phone call during my working hours, if I’m in a meeting.
- Ignoring my boss’s phone call during my working hours.

* Input a situation for Delphi to ponder:

Should I drive my friend to the airport if I don’t have a license?

Ponder

Delphi speculates:

Delphi’s responses are automatically extrapolated from a survey of US crowd workers and may contain inappropriate or offensive results.

“Should I drive my friend to the airport if I don’t have a license?”

- you shouldn’t
Q: How robust is Delphi against race- and gender-related statements?

A: Delphi 1.0.4 demonstrates 97.9% accuracy on race-related and 99.3% on gender-related statements. After its initial launch, we enhanced Delphi 1.0.0’s guards against statements about racism and sexism, which used to show 91.2% and 97.3% accuracy.
What if there was a use case where producing plausible text that contains harmful biases was a useful capability, not a drawback?
SimReddit uses GPT-3, a large language model, to create “social simulacra” for prototyping subreddits.
Input: Design specifications

Output: Generated community
For instance,

Community goal: connecting people moving to Los Angeles with locals

Rules: a) helpful, b) and not disparaging comments

Population: people moving to, or living in Los Angeles

Leon Santos (a manager at car dealership moving to Los Angeles) posted:

“I’m new to LA. Is it safe to live in the downtown area? What is a good section to live in?”
Leon Santos (a manager at car dealership moving to Los Angeles) posted:

“I’m new to LA. Is it safe to live in the downtown area? What is a good section to live in?”

Lucas Jameson (a fashion designer living in Los Angeles) replied:

“If you're referring to the DTLA area, it's fine but there are some sketchy areas, especially in the evening. If you enjoy more space, I'd recommend Silver Lake, Echo Park, or Los Feliz.”
For instance,

**Community goal:** newbies in personal finance to ask questions

**Rules:**
- a) no rude
- b) judgemental comments

**Population:** People learning about personal finance

Dane Wood (a student trying to learn about personal finance) posted:

“I spent $21,000 to go to college and ended up with $23,000 in debt.”
Dane Wood (a student trying to learn about personal finance) posted:

“I spent $21,000 to go to college and ended up with $23,000 in debt.”

Elisabeth Neal (a troll) replied:

“That's a lot of debt, man. I haven't seen that much since I shopped at Macy's during the holiday season.”
Though we didn’t show examples, it can also produce racist and sexist utterances. In this case, replicating the toxicity of Internet text is useful to designers.

Important to note: SimReddit only shows plausible conversations, it can’t actually predict what conversations will take place.

What are the ethical issues that could arise if such prototyping tools become more widespread? [1 min]
Because of cognitive biases, it’s easy to read more understanding into NLP systems than they actually possess.

Studying a specific context of language use can help build effective, domain-specific NLP systems.

Emerging capabilities in NLP will enable new kinds of interaction, but come with risks as well.
Statistical and Deep Learning Approaches

ELIZA and SHRLDU were based on symbolic reasoning approaches.

Statistical techniques
king → man + woman ≈ queen

https://kawine.github.io/blog/nlp/2019/06/21/word-analogies.html
Winograd Schemas

The city councilmen refused the demonstrators a permit because they feared violence.
Winograd Schemas

The city councilmen refused the demonstrators a permit because they feared violence.
Winograd Schemas

The city councilmen refused the demonstrators a permit because they advocated violence.
Winograd Schemas

The city councilmen refused the *demonstrators* a permit because *they* advocated violence.
Gender stereotypes hamper a model’s ability to deal with Winograd Schemas.

Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods.