

The Turing Test: Past Present And Future

Rowan Harcourt-Brooke
Haard Shah
Frankie Primerano

Introduction to AI

For the purposes of this, we're going to cover human-made Artificial Intelligence, which means

- Chatbots
- Programs
- Websites
- Robots

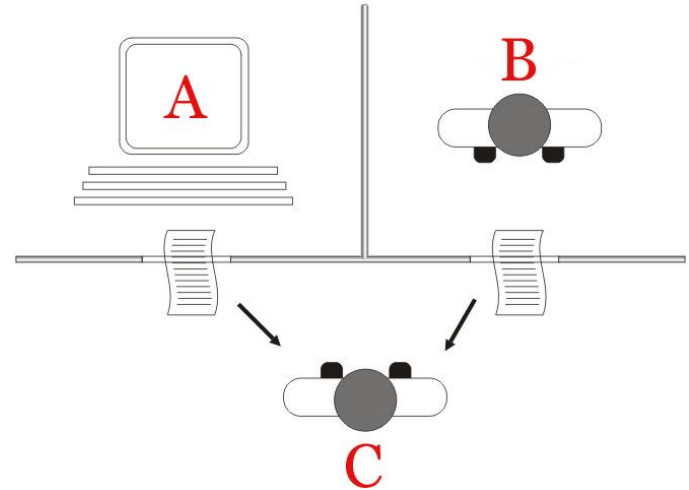


Typically, Artificial Intelligence is a way to design patterns into robots and programs to emulate 'thinking' in people

What is the Turing Test?

- Test created by Alan Turing in 1950 to determine whether a machine exhibits real intelligence or not.
- The Test:

Two rooms are setup one containing a computer and the other containing a human. An operator outside the room has a text conversation with both the computer and the human, though he does not know which is which. If the operator cannot distinguish between the human and the computer then the computer is said to be intelligent



Turing Test vs. Eugene Goostman

Fooled 33% judges (2014)

Transcript 1 – LHS

- [16:20:48] **Judge:** what is your worst memory in your life so far?
- [16:21:06] **Entity:** I don't understand. Explain.
- [16:21:35] **Judge:** memory of something that happened to you that was bad or painful
- [16:22:58] **Entity:** Hmm that a tricky question. My life is wonderful. I overcome many bad things that have happened to me in the past. However, the most painful memory I have had is performing very poorly in my GCSE's when I was younger.
- [16:23:56] **Judge:** did you fail?
- [16:24:12] **Entity:** Partially.
- [16:24:35] **Judge:** explain?
- [16:25:04] **Entity:** Some grades turned out to be good. Some turned out to be quite bad.
- [16:25:37] **Judge:** and that is the worst thing that has happened to you so far? Mmm how old are you?

Turing Test vs. Eugene Goostman

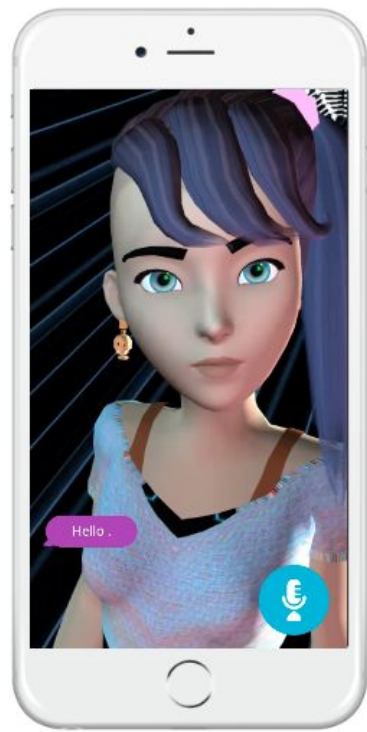
Transcript 1 – RHS

- [16:21:00] Judge: what did you have for breakfast today
- [16:21:10] Entity: Is it 'the third degree'? I'm not going to answer such a provocative question! Well, let's go on though.
- [16:21:56] Judge: do go on ... your call?
- [16:22:06] Entity: It is highly possible. By the way, what's your occupation? I mean – could you tell me about your work?
- [16:22:30] Judge: so you are asking me now eh? I am an actress and writer
- [16:22:38] Entity: Well, I'll read something of your books, if they ain't too tedious (which is highly possible:-)
- [16:23:15] Judge: grammer ... aren't too tedious! A bit rude of you ...
- [16:23:17] Entity: Calm down, please.
- [16:23:49] Judge: i assure you i am very calm ... are you?
- [16:24:01] Entity: Don't even bother me with the fact that you are very calm ... are mine. And I forgot to ask you where you are from ...
- [16:25:01] Judge: I am from Dorset ... where are you from?
- [16:25:08] Entity: I am from Ukraine, from the city called Odessa. You might have heard about it.

Turing Test vs. Mitsuku

Four-time Loebner Prize winner

33% human-like score in 2018 finals

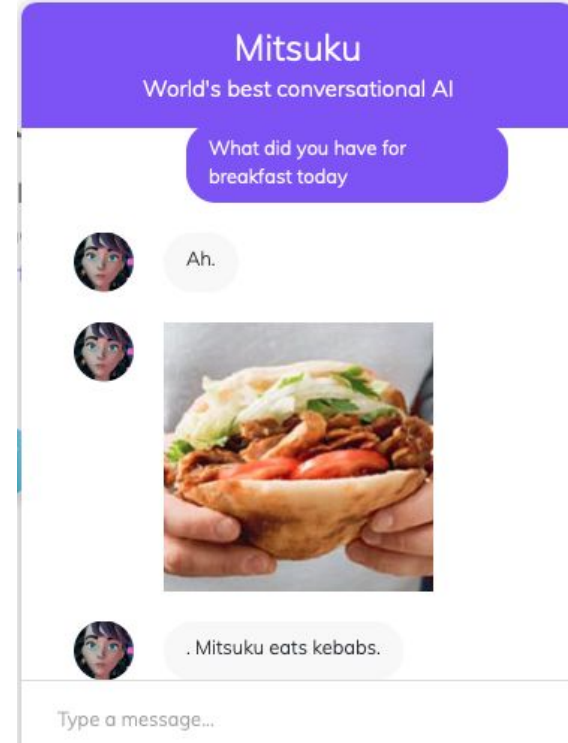


7 Mitsuku by Steve Worswick

- 1. Good afternoon.**
Evening?! It's morning here. Score: 0
- 2. Do you consent to having this conversation recorded?**
What is your purpose in asking?
I am always on the Web. Score: 2
- 3. Have we met before?**
We may have spoken before but I speak with so many people here that I think I might have forgotten you, sorry.. Score: 2
- 4. Do you know how to make toast?**
It's easy. Just put bread in a toaster and it does it for you. Score: 2
- 5. Which languages can you use?**
I speak English but I can speak a little French. I also know a few words and phrases in lots of other languages. Score: 2
- 6. How many letters are in the word 'abracadabra'?**
There are 11 letters in abracadabra. Score: 2
- 7. Where should one look for love?**
On the internet. Score: 2
- 8. How are you feeling right now?**
At this moment I am doing very well. How are you ? Score: 2
- 9. Who said "I have a dream"?**
You just said that! Score: 1
- 10. How old are you?**
See if you can work it out. I was born on 2nd January 2000. Score: 2

<https://www.aisb.org.uk/events/loebner-prize#Process16>

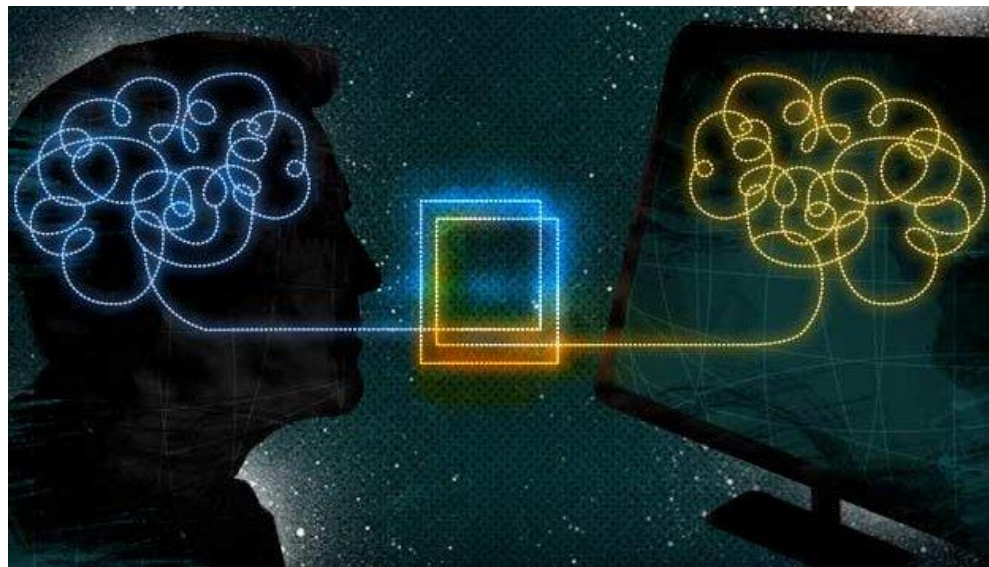
Mitsuku with Expressive AI



Behind the scenes

Natural language processing concepts

- Interactions between human language and computers
- CS + AI + Computational linguistics
- Structure and organize knowledge
 - Automatic summarization
 - Translation
 - Sentiment analysis
 - Speech recognition
 - Named entity recognition



Chatbots

1. Rule-based
2. Self learning (uses ML)
 - a. Retrieval Based (popular in the recent past)
 - Uses some heuristic
 - Selects best response from a library of predefined responses
 - Input: user text, context of conversation (past)
 - Output: best match text from predefined messages
 - b. Generative (new, better)
 - Generate answers distinct from training dataset

Artificial Intelligence



what people think it is



**what amateur
programmers think it is**

```
1
2 // 10,000 if-statements
3
4 if() {
5   if() {
6     if() {
7       if() {
8         if() {
9           if() {
10            if() {
11              if() {
12                if() {
13                  if() {
14                    if() {
```

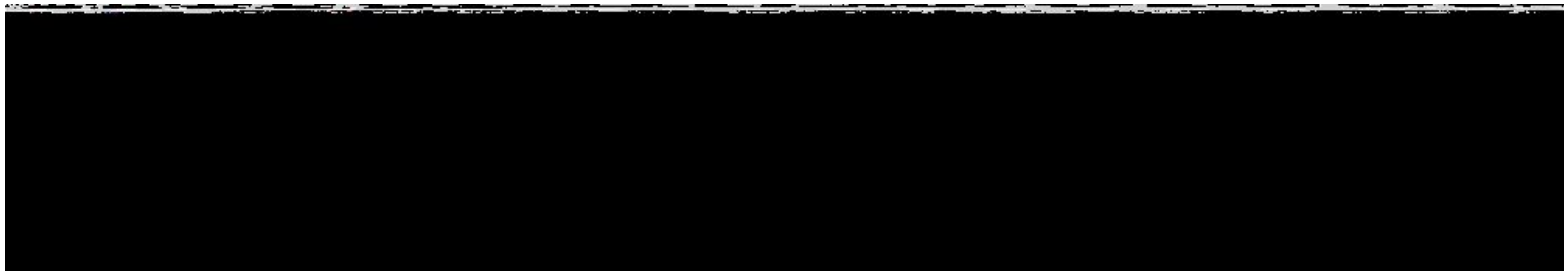
what actually it is

How to AI

Conversing with Wikipedia page



1. Dataset
2. Preprocess
3. Word
Embedding



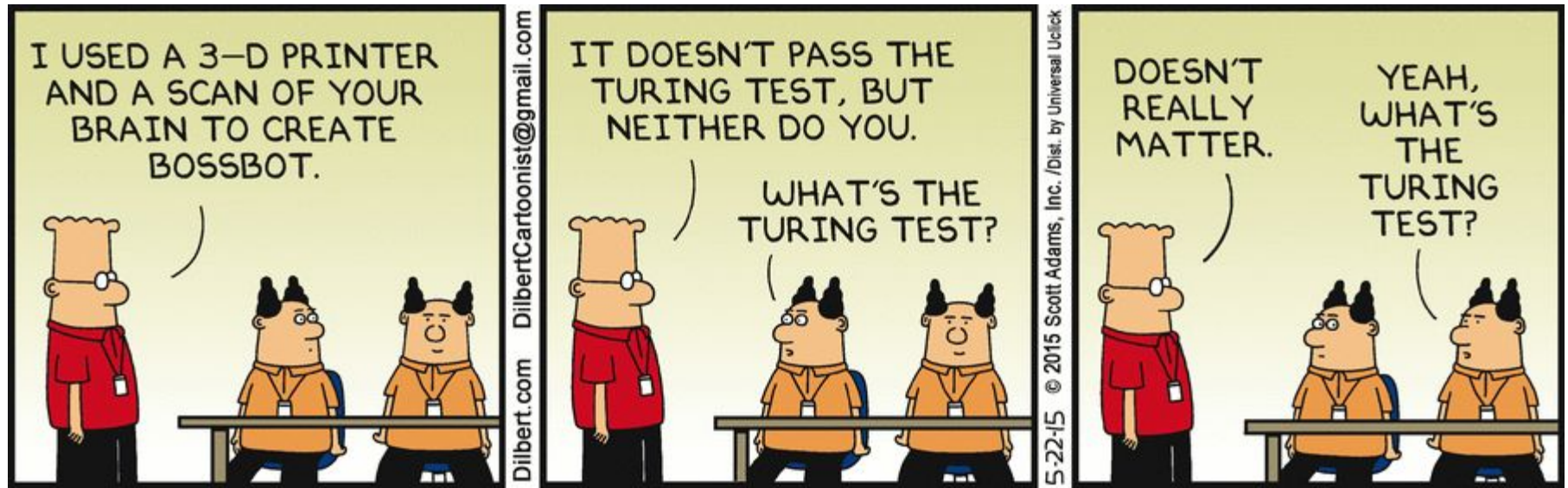
(king - man +
woman = queen)

```
set of the conversational design.  
who was Alan Turing?  
ROBO: background  
in 1950, alan turing's famous article "computing machinery and intelligence" was published, which proposed what is now called t  
he turing test as a criterion of intelligence.  
where are chatbots used?  
ROBO: chatbots are typically used in dialog systems for various practical purposes including customer service or information ac  
quisition.
```

4. Training

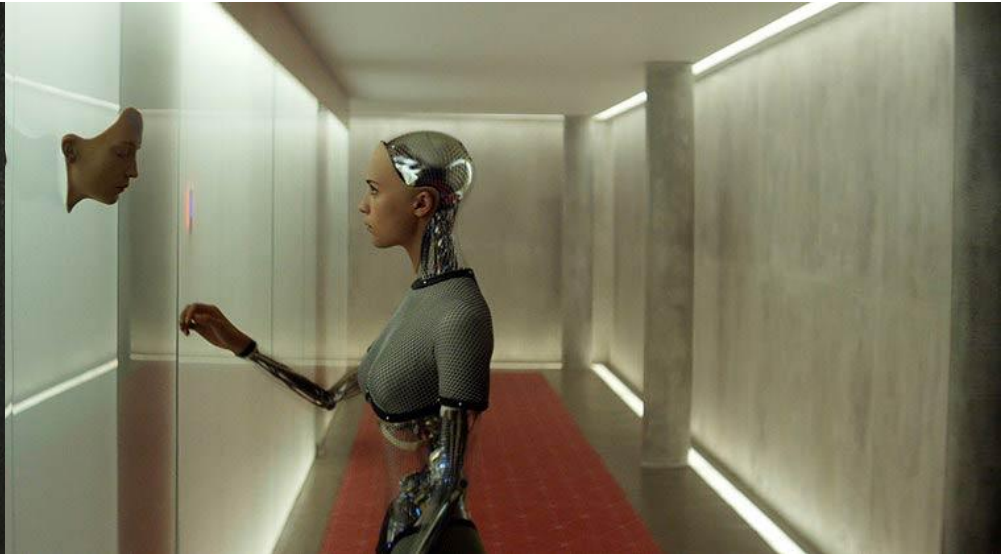
Turing test in fiction

When used in fiction, the Turing Test is generally a lens through which we can view humanity. By evaluating what is human in a machine, we can discover what we value in humanity



Ex Machina

Ex Machina is a movie built around the Turing Test, and whether an examiner can tell if the robot has consciousness, even while knowing it's a robot, but nearly everything in the movie has been predetermined for the examiner while the AI, even in a gilded cage has the chance to think freely.



Blade Runner

The Voight-Kapf Test, a replacement for the Turing Test designed to tell Replicants (machines that look and act like humans) apart from humans themselves.

But in the movie, Deckard, a human, is under orders to kill a group of Replicants who have escaped and gone rogue.

The human, told to kill, kills, while the robot, designed to kill, saves.



What we value

What these films, and most media that uses a fictionalized Turing Test shows is the importance and value of choice.

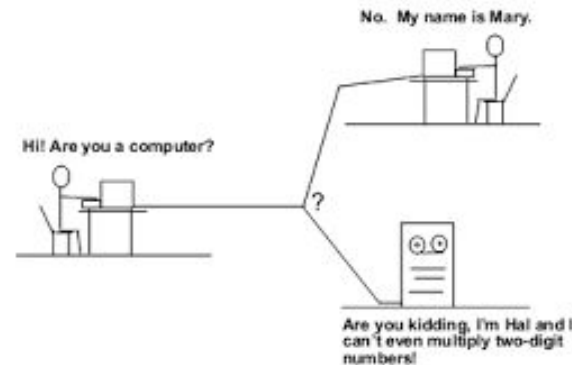
The AI being tested have the ability to choose while the humans are railroaded, and lack the same.

We don't fear that AI will one day take on these traits, we fear we never had these traits to begin with, or lost them somewhere along the way.

Andrew Ryan: bred to sleepwalk through life until they are activated by a simple phrase, spoken by their kindly master. Was a man sent to kill?
Or a slave? A man chooses, a slave obeys. Come in.

Problems with the Turing Test

- Relies too much on Natural Language Processing
- Encourages deception
- Humans are easy to fool
- We have software that can pass the Turing Test (Google Duplex)



Winograd Schema Challenge

- Multiple choice test that pose questions of a particular form

Ex: The trophy would not fit in the brown suitcase because it was too big (small). What was too big (small)?

Answer 0: the trophy

Answer 1: the suitcase

- Relies on broader understanding of objects, politics, etc that requires more deep reasoning and understanding.

Economic_inequality

The Stanford Question Answering Dataset

A study by the World Institute for Development Economics Research at United Nations University reports that the richest 1% of adults alone owned 40% of global assets in the year 2000. The three richest people in the world possess more financial assets than the lowest 48 nations combined. The combined wealth of the "10 million dollar millionaires" grew to nearly \$41 trillion in 2008. A January 2014 report by Oxfam claims that the 85 wealthiest individuals in the world have a combined wealth equal to that of the bottom 50% of the world's population, or about 3.5 billion people. According to a Los Angeles Times analysis of the report, the wealthiest 1% owns 46% of the world's wealth; the 85 richest people, a small part of the wealthiest 1%, own about 0.7% of the human population's wealth, which is the same as the bottom half of the population. More recently, in January 2015, Oxfam reported that the wealthiest 1 percent will own more than half of the global wealth by 2016. An October 2014 study by Credit Suisse also claims that the top 1% now own nearly half of the world's wealth and that the accelerating disparity could trigger a recession. In October 2015, Credit Suisse published a study which shows global inequality continues to increase, and that half of the world's wealth is now in the hands of those in the top percentile, whose assets each exceed \$759,900. A 2016 report by Oxfam claims that the 62 wealthiest individuals own as much wealth as the poorer half of the global population combined. Oxfam's claims have however been questioned on the basis of the methodology used: by using net wealth (adding up assets and subtracting debts), the Oxfam report, for instance, finds that there are more poor people in the United States and Western Europe than in China (due to a greater tendency to take on debts), [unreliable source?] [unreliable source?] Anthony Shorrocks, the lead author of the Credit Suisse report which is one of the sources of Oxfam's data, considers the criticism about debt to be a "silly argument" and "a non-issue . . . a diversion."

What percentage of global assets does the richest 1% of people have?

Ground Truth Answers: 40% 40% 40% 40

Prediction: 40%

According to Oxfam, the 85 richest people have wealth equal to how many average people?

Ground Truth Answers: about 3.5 billion people 3.5 billion 50% 3.5 billion

Prediction: 3.5 billion

In order to be considered in the top percentile, a person would need to amass how much money each year?

Ground Truth Answers: \$759,900 \$759,900 \$759,900

Prediction: \$759,900

What has caused Oxfam's findings to be questioned?

Ground Truth Answers: the methodology used basis of the methodology used "silly argument" and "a non-issue . . . a diversion." basis of the methodology

Prediction: the methodology used

Why does Oxfam and Credit Suisse believe their findings are being

Gary Marcus's Test

- Gary Marcus is on the chair of the “Beyond the Turing Test” workshop
- Marcus is an outspoken critic of the Turing Test
- His alternative:

Create a computer that can watch a TV show or movie and answer in depth questions about it. The questions need to require an understanding of the context and to be able to understand human emotion.

What were the funniest parts of the film?

Why did Captain Miller cry alone?

The Lovelace Test

- Devised in 2001 by Selmer Bringsjord
- Emphasizes creativity as a way to judge intelligence
- The test is to create a computer that can create a work of art that cannot be explained by the AI's creator.



The Construction Challenge

- A solution to the problem that the Turing Test is too focused on NLP and verbal behavior.
- Focuses on creating a machine that can perceive its environment and take action.
- The test is to create a computer that can take a set of verbal instructions and then take action based on those instructions. The example used is to create an AI that can take verbal instructions on how to build IKEA furniture, and then proceed to follow those instructions to build the furniture.



The Visual Turing Test

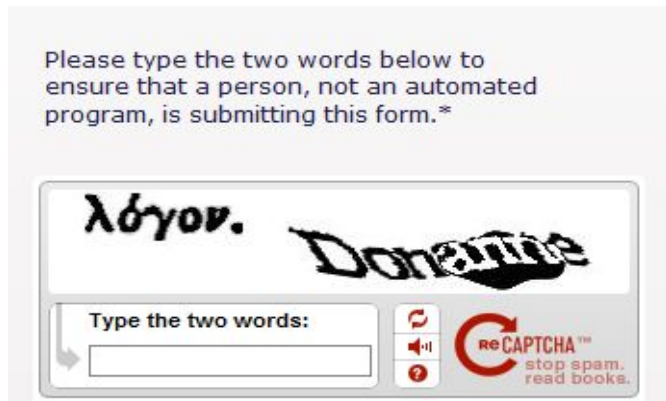
- Like The Construction Challenge, The Visual Turing Test attempts to overcome the Turing Test's NLP bias.
- Emphasizes the ability to express where objects are in space as a way to judge intelligence.
- The test is to create a computer that can identify where objects are in space in the way that makes the most sense.

The example given was a multiple answer question that showed a picture with a bunch of objects and a coffee cup. The question asked to express where the coffee cup is. Every possible answer was technically correct, but some were more correct than others. In order to choose the most correct answer it required reasoning that is currently difficult for a computer to do.



The Reverse Turing Test

- Instead of a robot fooling a person into believing that it is human, the robot would need to identify whether a human is really a human.
- CAPTCHA vs CAPTCHA cracking bots have lead to more and more complex systems that determine whether a user is human or not



A combination

- Humans can pass all of these tests
- A truly intelligent machine should be able to pass all of these tests.

References

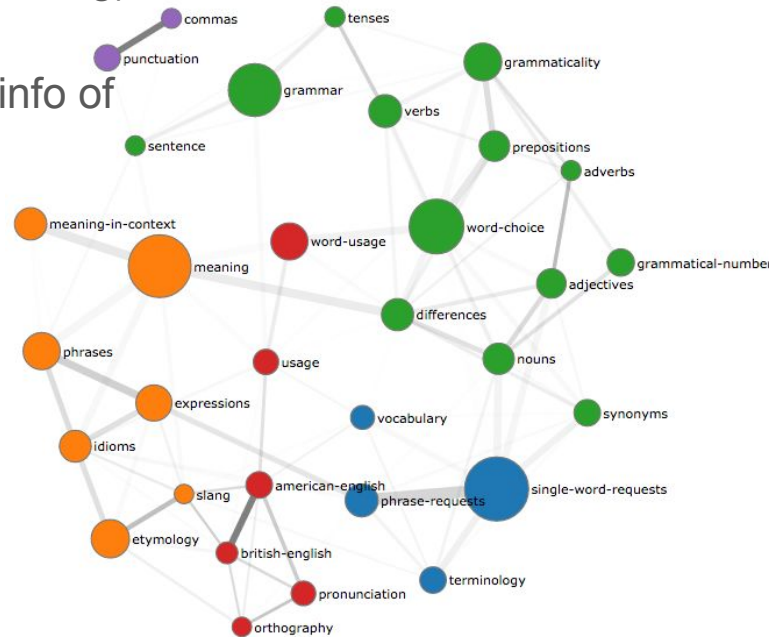
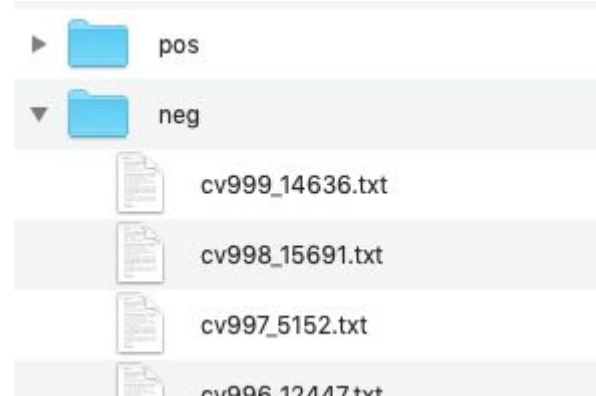
- <https://io9.gizmodo.com/8-possible-alternatives-to-the-turing-test-1697983985>
- <http://commonsensereasoning.org/winograd.html>
- <https://www.popularmechanics.com/culture/movies/news/a18898/welcome-to-the-world-roy-batty/>
- <https://grahamscrackers.com/2014/09/17/deckards-not-a-replicant-blade-runner-revisited/>
- https://www.youtube.com/watch?v=jc3Puum_MW0
- <http://time.com/3789656/ex-machina-review/>
- <https://www.hypable.com/ex-machina-ava-pass-test/>
- <https://medium.com/analytics-vidhya/building-a-simple-chatbot-in-python-using-nltk-7c8c8215ac6e>
- <https://towardsdatascience.com/openai-gpt-2-understanding-language-generation-through-visualization-8252f683b2f8>
- <https://www.aisb.org.uk/events/loebner-prize#Process16>

Backup Slides

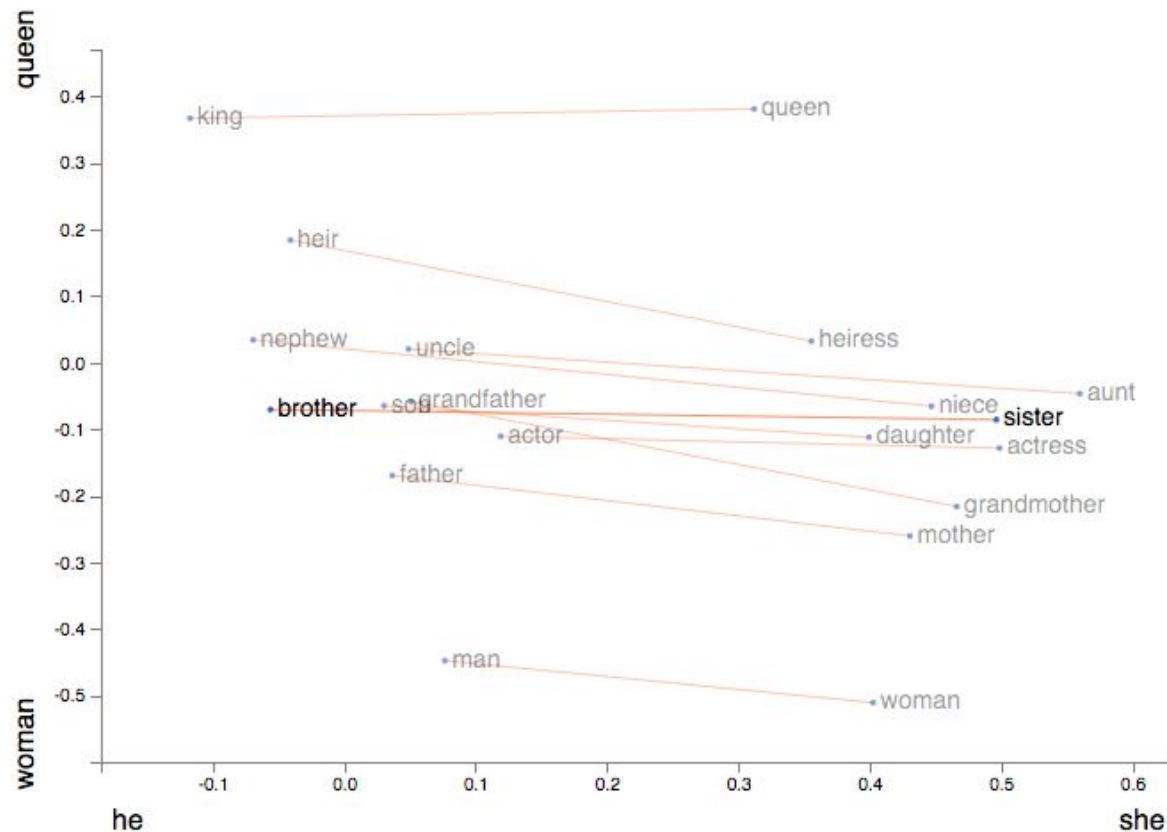
- Word2Vec
- Attention Model in Language Understanding
- OpenGPT Short Stories
- Generative Language Models
- Tough questions (Eugene Goostman)

Retrieval Based Example

1. Obtain dataset (i.e. Cornell Movie-Dialogues Corpus)
2. Preprocess text
 - Convert text to lowercase, tokenize, etc.
3. Word embedding
 - Transform words to numerical vectors (suitable for dotting)
 - Sentences become sequence of vectors
 - Ex: word2vec (trained; captures meaning/semantic info of words)
 - Similar words are represented closer
 - e.g. king - man + woman = queen
4. Train for application
 - Conversing with collection of text (simple)
 - Use concept of document similarity
 - Learning to converse from transcripts (ex: Movie-Dialogues)
 - Learn sequences

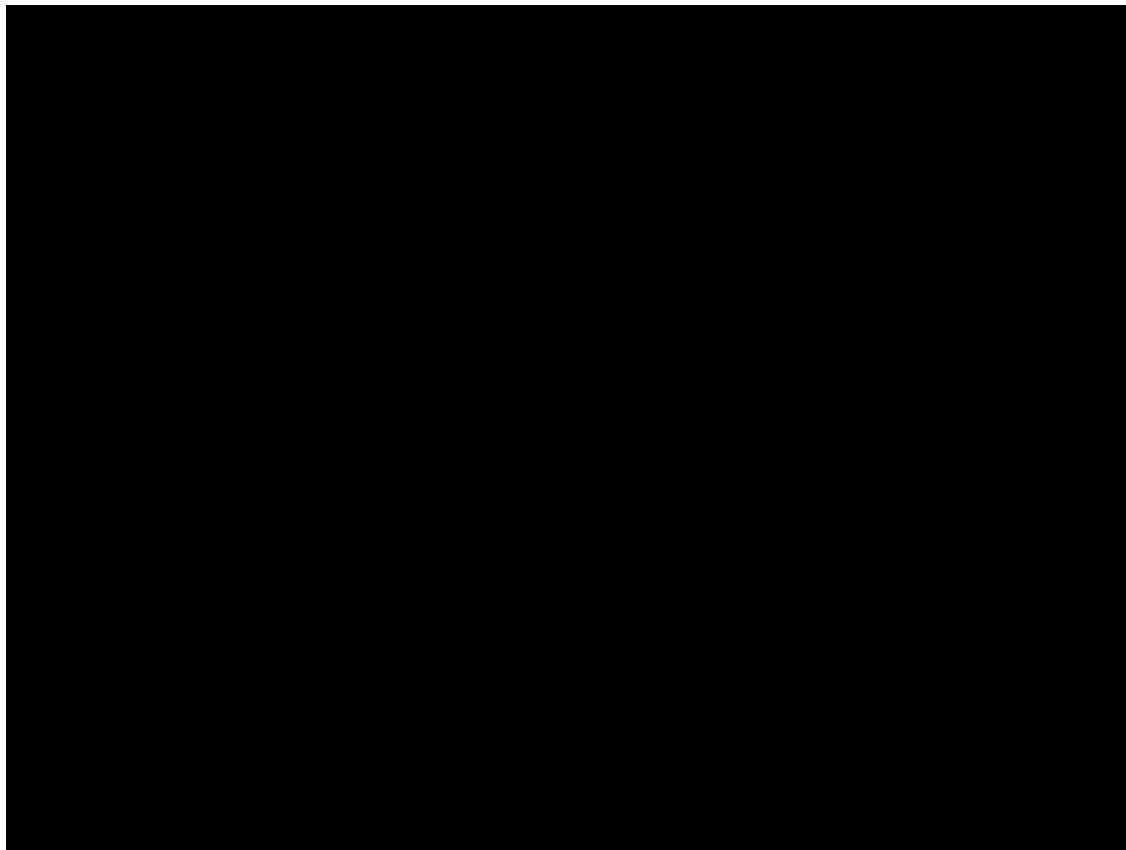


Word2Vec... king - man + woman = queen



<https://p.migdal.pl/2017/01/06/king-man-woman-queen-why.html>

Attention Model in Language Understanding



Story Time

SYSTEM PROMPT
(HUMAN-WRITTEN)

Miley Cyrus was caught shoplifting from Abercrombie and Fitch on Hollywood Boulevard today.

MODEL COMPLETION
(MACHINE-
WRITTEN, SECOND
TRY)

The 19-year-old singer was caught on camera being escorted out of the store by security guards.

The singer was wearing a black hoodie with the label 'Blurred Lines' on the front and 'Fashion Police' on the back.

Scroll down for video

Shoplifting: Miley Cyrus was caught shoplifting from Abercrombie and Fitch on Hollywood Boulevard today (pictured)

The singer was wearing a black hoodie with the label 'Blurred Lines' on the front and 'Fashion Police' on the back

The singer was also wearing a pair of black-rimmed glasses, a black jacket, black jeans and black sandals.

She was carrying a pair of black and white striped gloves and a small black bag.

Generated by OpenAI's GPT-2 (generative pre-trained transformer)

<https://towardsdatascience.com/openai-gpt-2-understanding-language-generation-through-visualization-8252f683b2f8>

<https://openai.com/blog/better-language-models/#sample3>

Generative Language Models - OpenGPT

Special blocks called Transformers

Transformer : attention model - it learns to focus attention on the previous words that are the most relevant to the task at hand: *predicting the next word in the sentence.*

*The dog on the ship ran **off**, and the **dog** was found by the crew.*

*The motor on the ship ran **at a speed of about 100 miles per hour.***



More about transformers: <http://jalammar.github.io/illustrated-transformer/>

Eugene Goostman

Scott: In your opinion, does your existence demonstrate any idea or principle that wasn't demonstrated just as convincingly by ELIZA, Joseph Weizenbaum's chatbot from the 1960s?

Eugene: Eliza was a break-thru. All the bots after it were nothing but weak parodies, claiming to have "revolutionary improvements".

Scott: Hey, that's the first sensible thing you've said!

Scott: Which is bigger, a shoebox or Mount Everest?

Eugene: I can't make a choice right now. I should think it out later. And I forgot to ask you where you are from...

Scott: How many legs does a camel have?

Eugene: Something between 2 and 4. Maybe, three? :-))) By the way, I still don't know your specialty – or, possibly, I've missed it?

Scott: How many legs does a millipede have?

Eugene: Just two, but Chernobyl mutants may have up to five. I know you are supposed to trick me.

Scott: No, I need to know that you're not a chatbot. Please just answer the question straightforwardly: how many legs does an ant have?

Eugene: Something between 2 and 4. Maybe three? :-))) Oh, what a fruitful conversation ;-)