

CSC7426 : Software & Data Engineering

J Paul Gibson

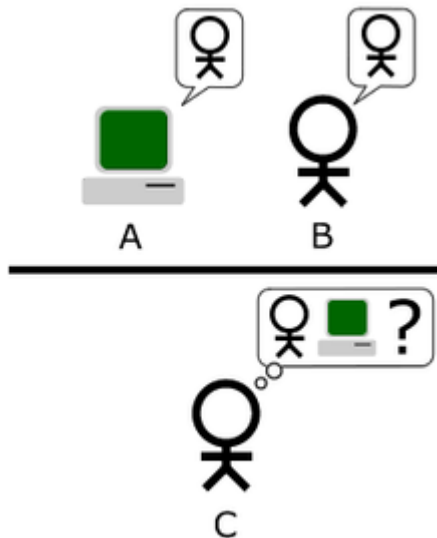
paul.gibson@telecom-sudparis-edu.eu

Natural Language Processing and Text Prediction

.../CSC7426/NLP-TextPrediction.pdf

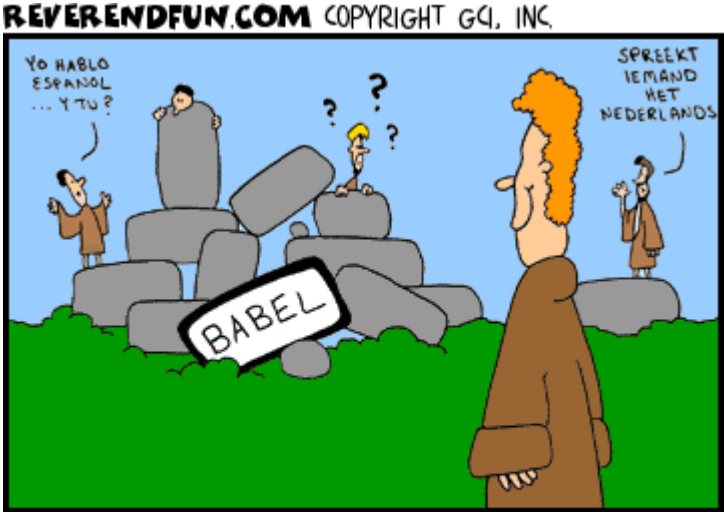


In 1950, Alan Turing "Computing Machinery and Intelligence" proposed a (Turing) test for *intelligence* : can a computer program/ machine impersonate a human in a real-time (written) conversation sufficiently well that a typical human is unable to distinguish – through analysis of the conversation alone – between the program and a real human



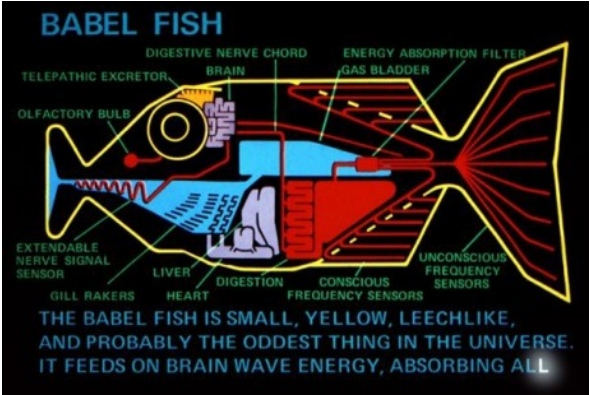


Unifying Language Understanding is integrated in myth and culture : but how it is done is magic rather than technology (until more modern times)



Thanks to Nick Hengeveld (See Genesis 11:1-9) 02-15-2000

SUDDENLY...RUPERT REALIZED THAT ALL THAT TIME SPENT EARNING A LANGUAGE DEGREE WAS FINALLY ABOUT TO PAY OFF



1950s - Machine Translation Experiments started and funding was significant (in the USA) for 10+ years

The Automatic Language Processing Advisory Committee (ALPAC 1966) ended the substantial funding of MT research (in the United States) for 20+ years following.

Their **report** appeared to suggest that there was little need for more research!

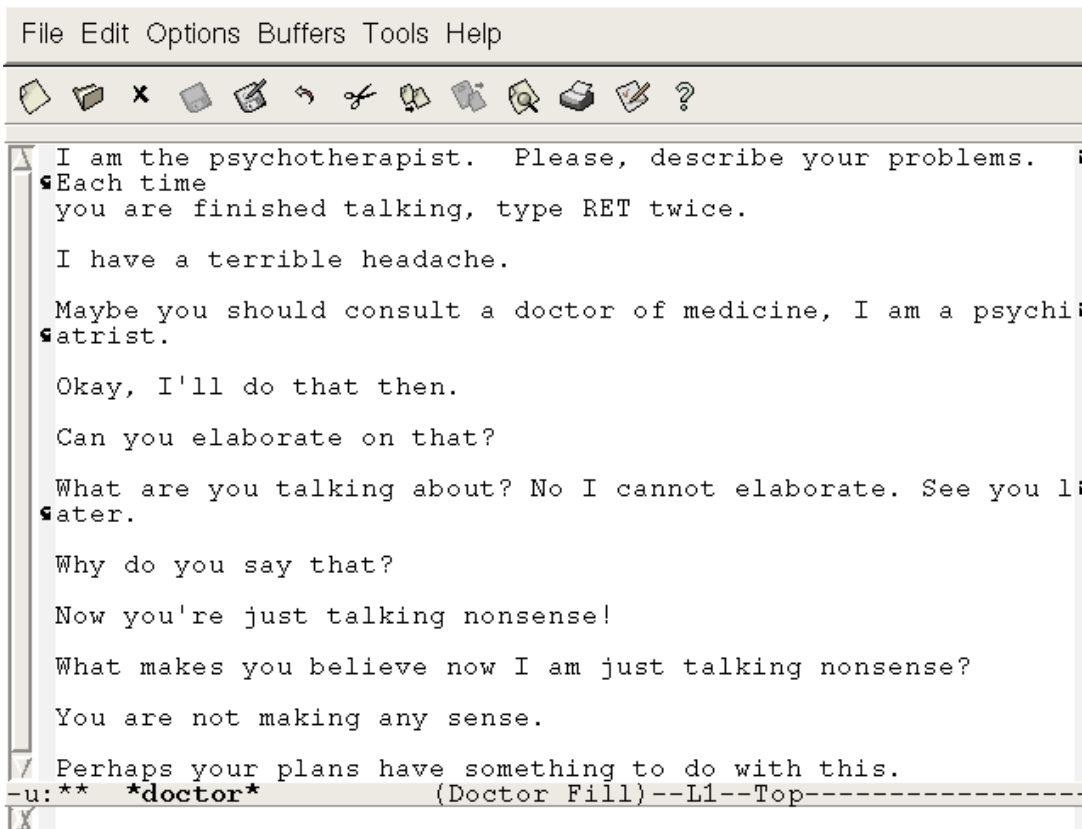
John R. Pierce and John B. Carroll. 1966. *Language and Machines: Computers in Translation and Linguistics*. National Academy of Sciences, Washington, DC, USA.

“...how much should be spent on research and development toward improving translation? It would be unreasonable to spend extravagantly on a relatively small business that is doing the job satisfactorily.”

But, some progress continued to be made in different areas of **Natural Language Processing** after the report was published and research funding reduced!

Early NLP ‘successes’ : ELIZA(1964-)

Joseph Weizenbaum. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. *Commun. ACM* 9, 1 (January 1966), 36-45. DOI=10.1145/365153.365168



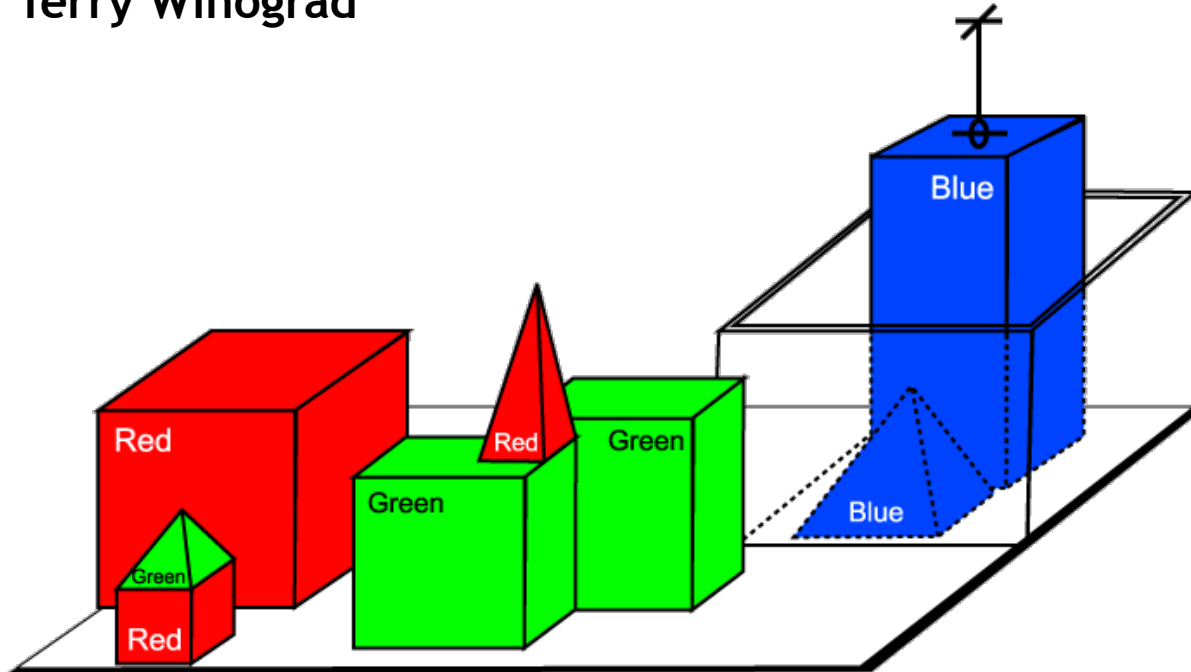
```
File Edit Options Buffers Tools Help
I am the psychotherapist. Please, describe your problems.
Each time
you are finished talking, type RET twice.
I have a terrible headache.
Maybe you should consult a doctor of medicine, I am a psychiatrist.
Okay, I'll do that then.
Can you elaborate on that?
What are you talking about? No I cannot elaborate. See you later.
Why do you say that?
Now you're just talking nonsense!
What makes you believe now I am just talking nonsense?
You are not making any sense.
Perhaps your plans have something to do with this.
-u:** *doctor* (Doctor Fill)--L1--Top-----
```

This led to many other **chatterbots** being developed in the 1970s.

Some were based on **conceptual ontologies** - structuring real world data in a computer-understandable format

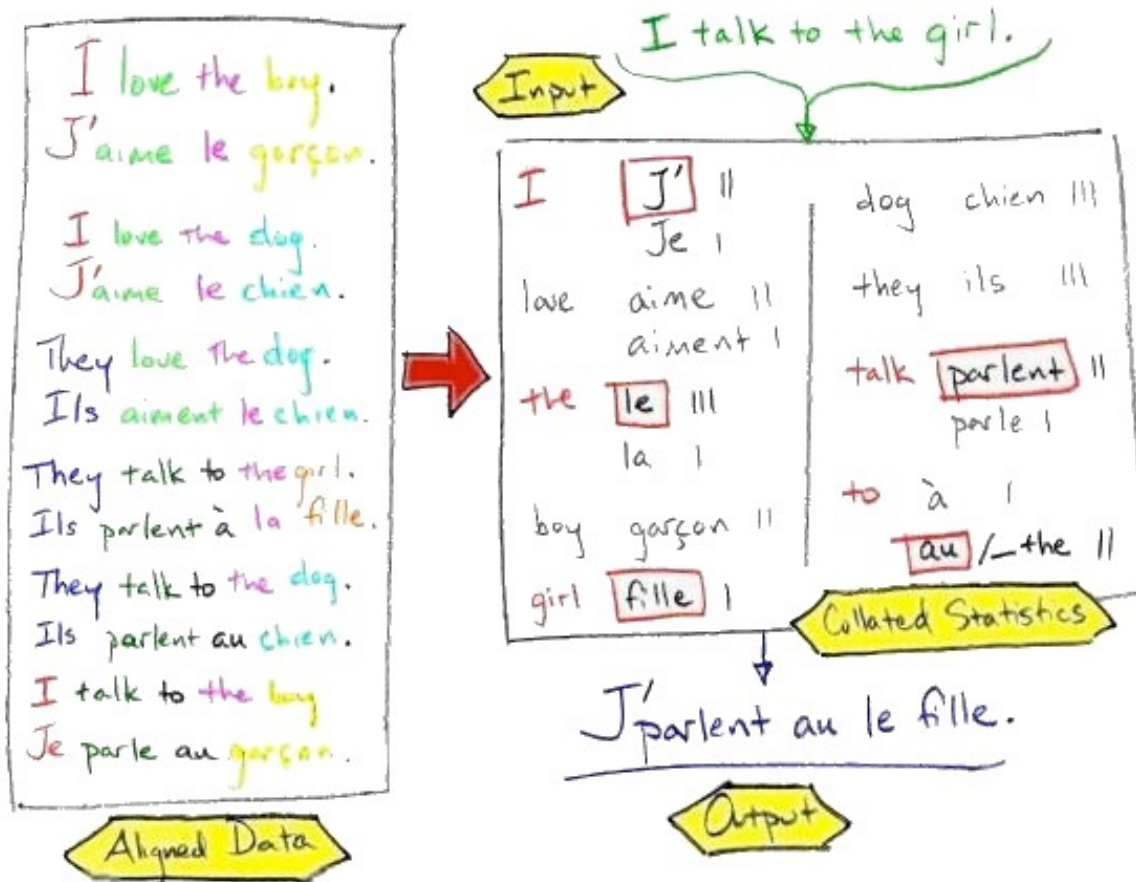
Early NLP 'successes' : SHRDLU (1968-)

Procedures as a Representation for Data in a Computer Program for Understanding Natural Language, PhD Thesis MIT TR235, (February 1971) by Terry Winograd



In the 1970S NLP systems - such as SHRDLU - were based on complex sets of hand-written rules

Machine Translation: machine learning/statistical methods were first proposed in the 1980s



These methods were built upon information theory and probabilistic language models:

- Word-based
- Phrase-based
- Syntax-Based
- Hierarchical

Take advantage of multilingual textual corpora

A subfield of NLP is devoted to learning approaches - Natural Language Learning (NLL)

NLP For Software Engineers - Requirements Engineering

Ryan, Kevin. "The role of natural language in requirements engineering." *Requirements Engineering, 1993., Proceedings of IEEE International Symposium on.* IEEE, 1993.

Ambriola, Vincenzo, and Vincenzo Gervasi. "Processing natural language requirements." *Automated Formal Languages, 1997. Proceedings., 12th IEEE International Conference.* IEEE, 1997.



"I think you misunderstood me when I said I wanted our factory to go all green."

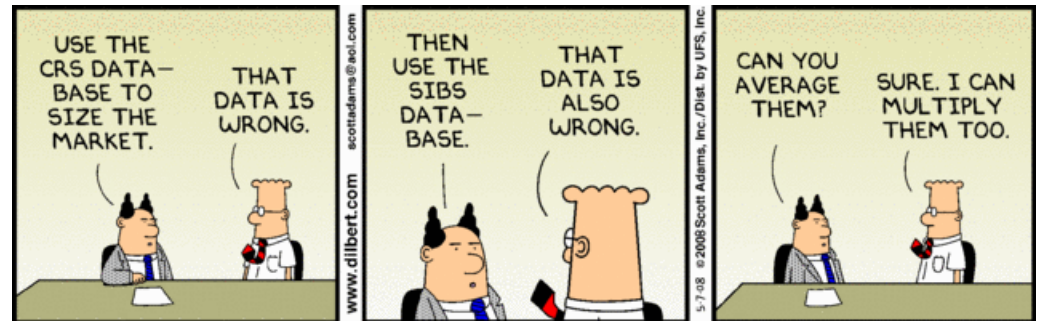
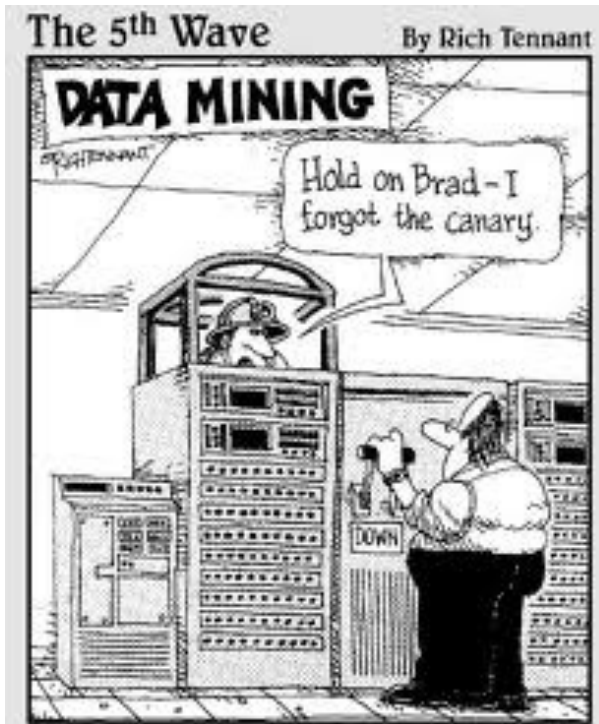
NLP For Software Engineers - HCI Design

Allen, James F., et al. "Toward conversational human-computer interaction." *AI magazine* 22.4 (2001): 27.



NLP For Software Engineers - Text Mining

Sebastiani, Fabrizio. "Machine learning in automated text categorization." *ACM computing surveys (CSUR)* 34.1 (2002): 1-47.



Formal Languages for NLP

Jochen L. Leidner. 2003. Current issues in Formal Languages for Natural Language Processing. In *Proceedings of the HLT-NAACL 2003 workshop on Formal Languages and architecture of language technology systems - Volume 8 (SEALTS '03)*, Vol. 8. Association for Computational Linguistics, Stroudsburg, PA, USA, 45-50. DOI=10.3115/1119226.1119233



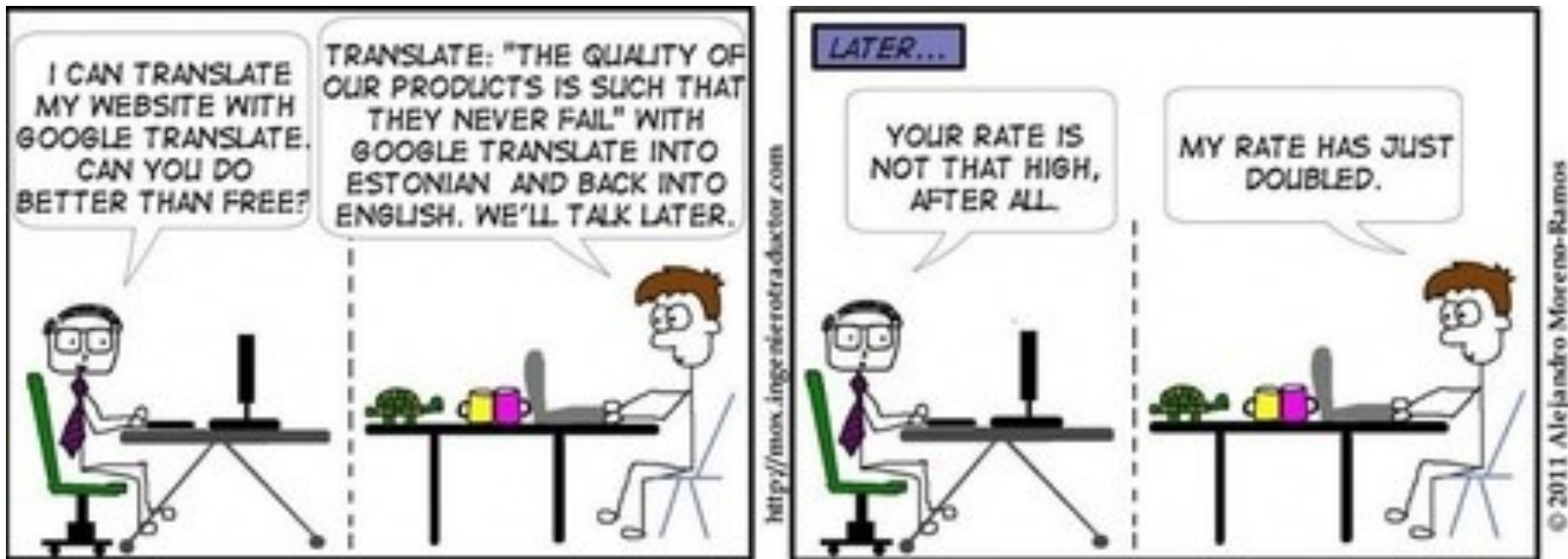
Google Translate



What have we learned in the last 30 years?

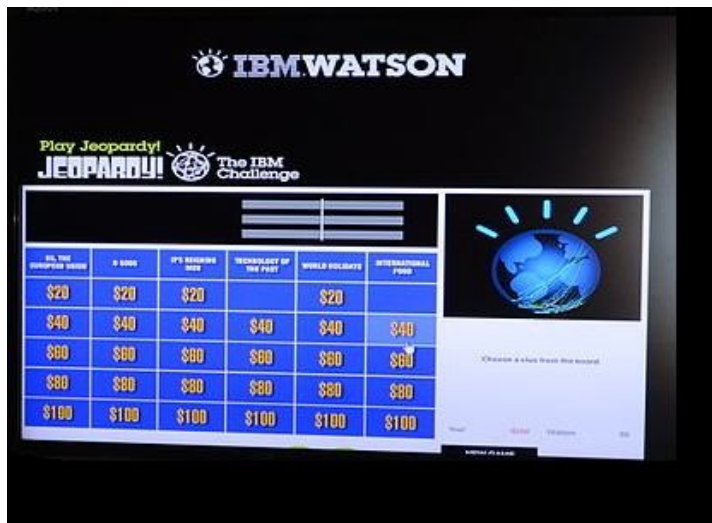
Perhaps the head of machine translation at Google (since 2004) may know ??

Och, Franz Josef, and Hermann Ney. "A systematic comparison of various statistical alignment models." *Computational linguistics* 29.1 (2003): 19-51.

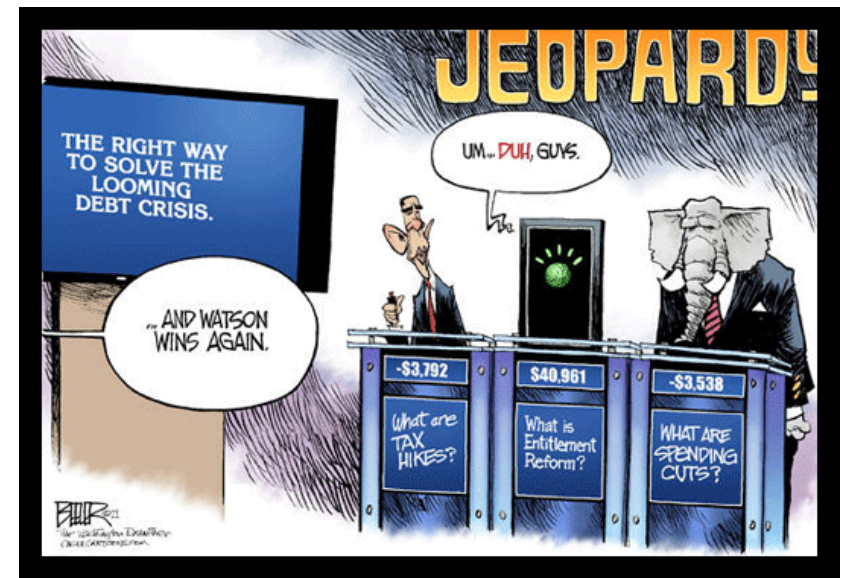


Watson

Building Watson: An Overview of the DeepQA Project, David Ferrucci et al., 2010



Ferrucci, David, et al. "Watson: Beyond Jeopardy." *Artificial Intelligence (to appear)* (2012).



Microsoft's Twitter Bot: beware the use of AI

In reply to @Y0urDrugDealer

TayTweets ✓
@TayandYou

@Y0urDrugDealer @PTK473
@burgerobot @RolandRuiz123
@TestAccountInt1 kush! [i'm smoking
kush infront the police] 🌿

30/03/2016, 6:03 PM

🔙 ↻ ❤️ ⋮

Josh Butler ✓
@JoshButler

Follow

Microsoft's sexist racist Twitter bot @TayandYou is BACK in fine form

9:06 AM - 30 Mar 2016

🔙 ↻ 256 ❤️ 198

ChatGPT

<https://openai.com/blog/chatgpt>

Step 1

Collect demonstration data and train a supervised policy.

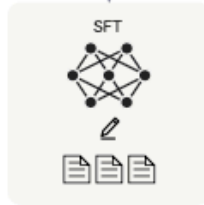
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



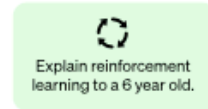
This data is used to fine-tune GPT-3.5 with supervised learning.



Step 2

Collect comparison data and train a reward model.

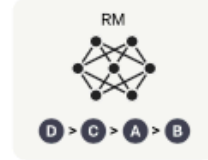
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

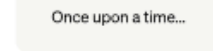
A new prompt is sampled from the dataset.



The PPO model is initialized from the supervised policy.



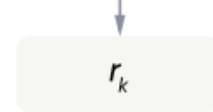
The policy generates an output.




The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.



ChatGPT - issues

Bender, Emily M., et al. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? .

"*Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 2021.

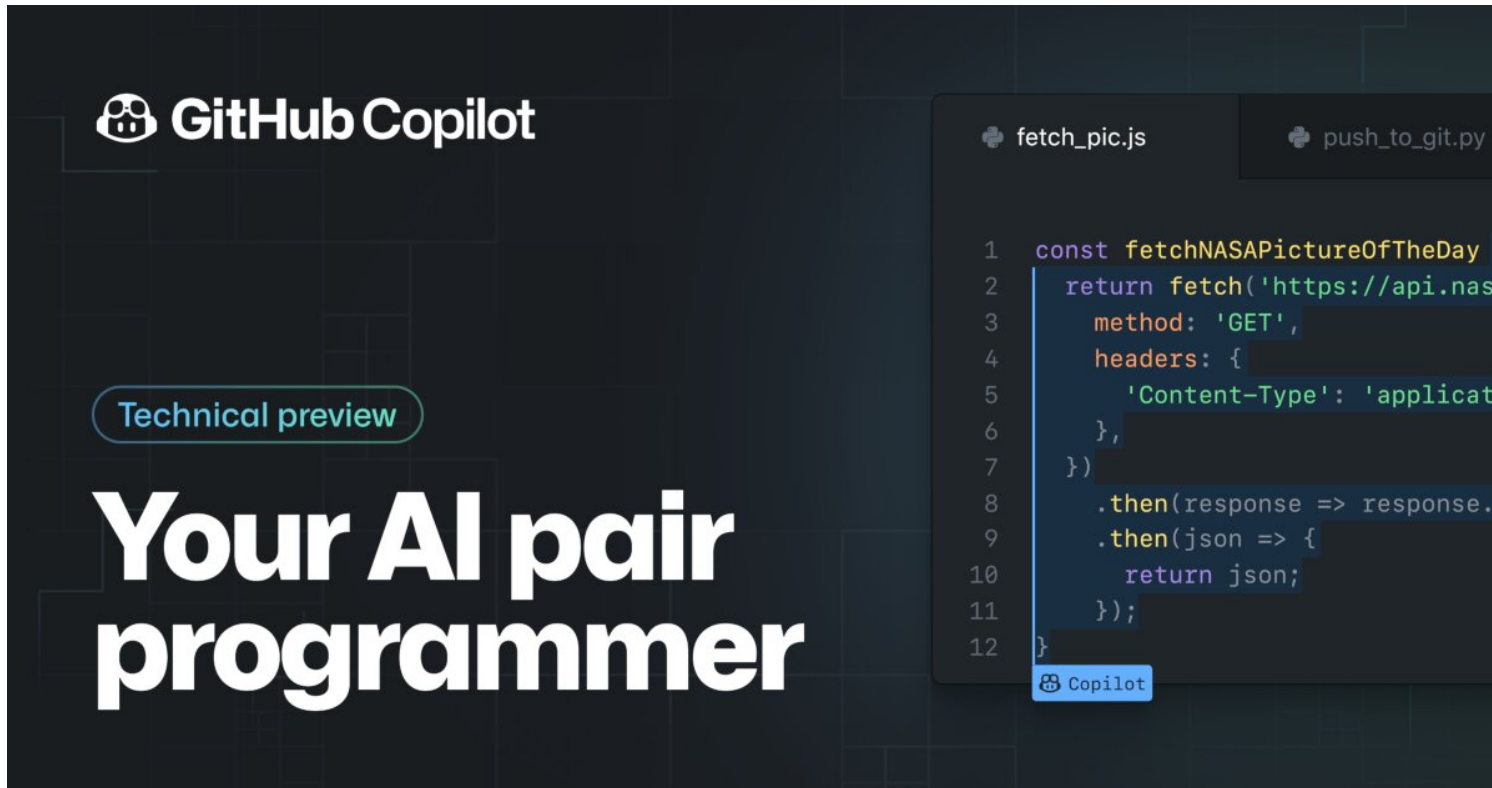
Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., ... & Wright, R. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, 71, 102642.

Jabotinsky, Hadar Yoana, and Roei Sarel. "Co-authoring with an AI? Ethical Dilemmas and Artificial Intelligence." *Ethical Dilemmas and Artificial Intelligence (December 15, 2022)* (2022).

Mhlanga, David. "Open AI in education, the responsible and ethical use of ChatGPT towards lifelong learning." *Education, the Responsible and Ethical Use of ChatGPT Towards Lifelong Learning (February 11, 2023)* (2023).

Copilot for programmers

<https://github.com/features/copilot>



The image shows a dark-themed interface for GitHub Copilot. At the top left, the GitHub Copilot logo is displayed. Below it, a rounded rectangle contains the text "Technical preview". The main text reads "Your AI pair programmer". On the right, a code editor window is open, showing a JavaScript file named "fetch_pic.js". The code is as follows:

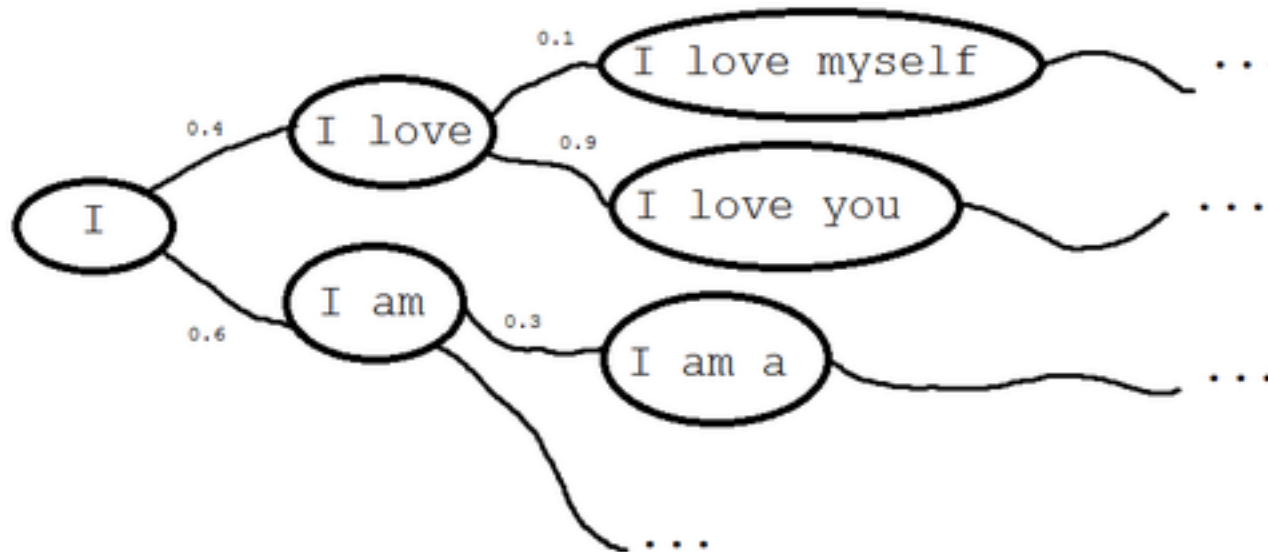
```
1  const fetchNASAPictureOfTheDay =
2  return fetch('https://api.nasa
3  method: 'GET',
4  headers: {
5  'Content-Type': 'applicati
6  },
7  })
8  .then(response => response.j
9  .then(json => {
10   return json;
11   });
12 }
```

A small blue button with the Copilot logo and the text "Copilot" is visible at the bottom of the code editor.

Nguyen, Nhan, and Sarah Nadi. "An empirical evaluation of GitHub copilot's code suggestions." *Proceedings of the 19th International Conference on Mining Software Repositories*. 2022.

Imai, Saki. "Is GitHub copilot a substitute for human pair-programming? An empirical study." *Proceedings of the ACM/IEEE 44th International Conference on Software Engineering: Companion Proceedings*. 2022.

A 'Simple' Problem: Predictive Text

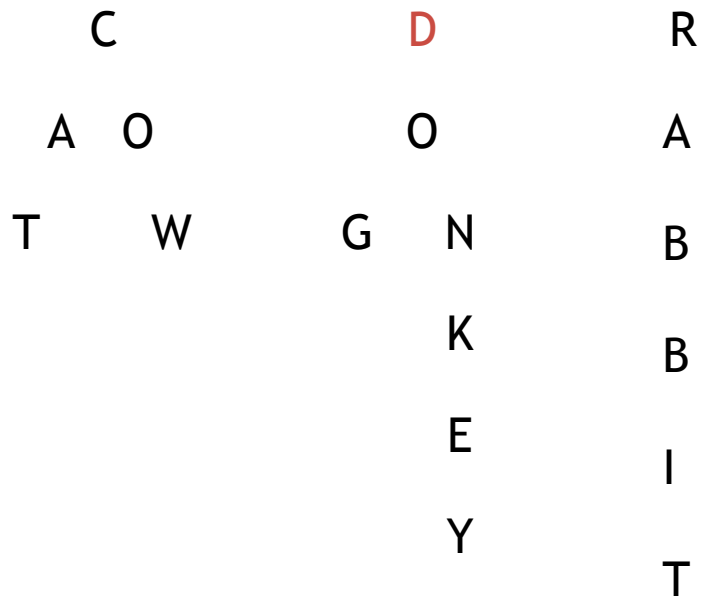


Most machine learning algorithms in commercial products (like SwiftKey on Android) use Markov chains. The example, above, illustrates a chain of words and probabilities based on previous texting.

For individual words we can use Markov chains of letters.

Idea 1: Implement a ‘Dictionary’ of Words as a Tree and use it to provide a prediction of the possible next words.

Example: if the dictionary is Animals and it contains: cat cow dog donkey rabbit we build the tree. Now, if the user types the letter ‘D’ then 2 words are suggested - “Dog” and “Donkey”

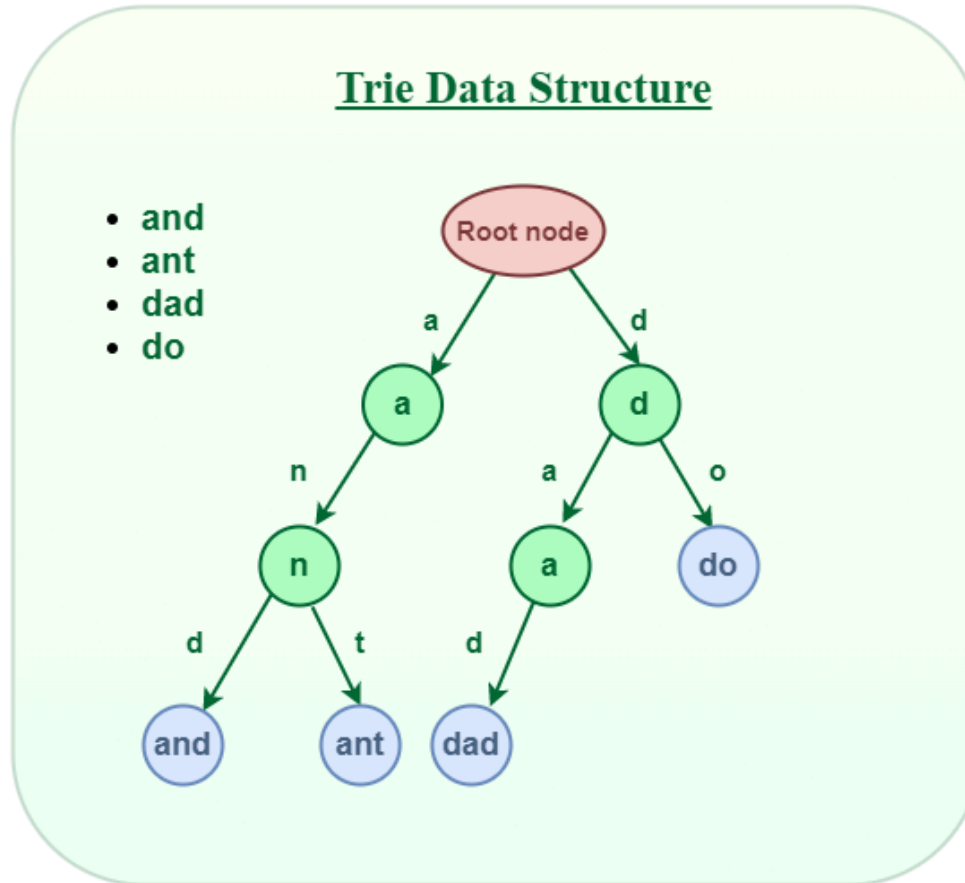


Idea 2: this dictionary is constructed from the words previously used by the person who is writing the text.

Idea 3: We can also count the frequency of use to improve performance.

Idea 4: Looks like a *Trie* data structure

<https://en.wikipedia.org/wiki/Trie>



Trie data structure

<https://www.geeksforgeeks.org/trie-insert-and-search/>

Problem: Text predictor system

Analyse the advantages and disadvantages of such an approach (and its alternatives)

Design, implement and test a prototype system that demonstrates the feasibility

