NPTEL ONLINE CERTIFICATION COURSES

DEEP LEARNING FOR NATURAL LANGUAGE PROCESSING

Lecture 01 : Introduction to the Course

PROF.PAWAN GOYAL

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

CONCEPTS COVERED

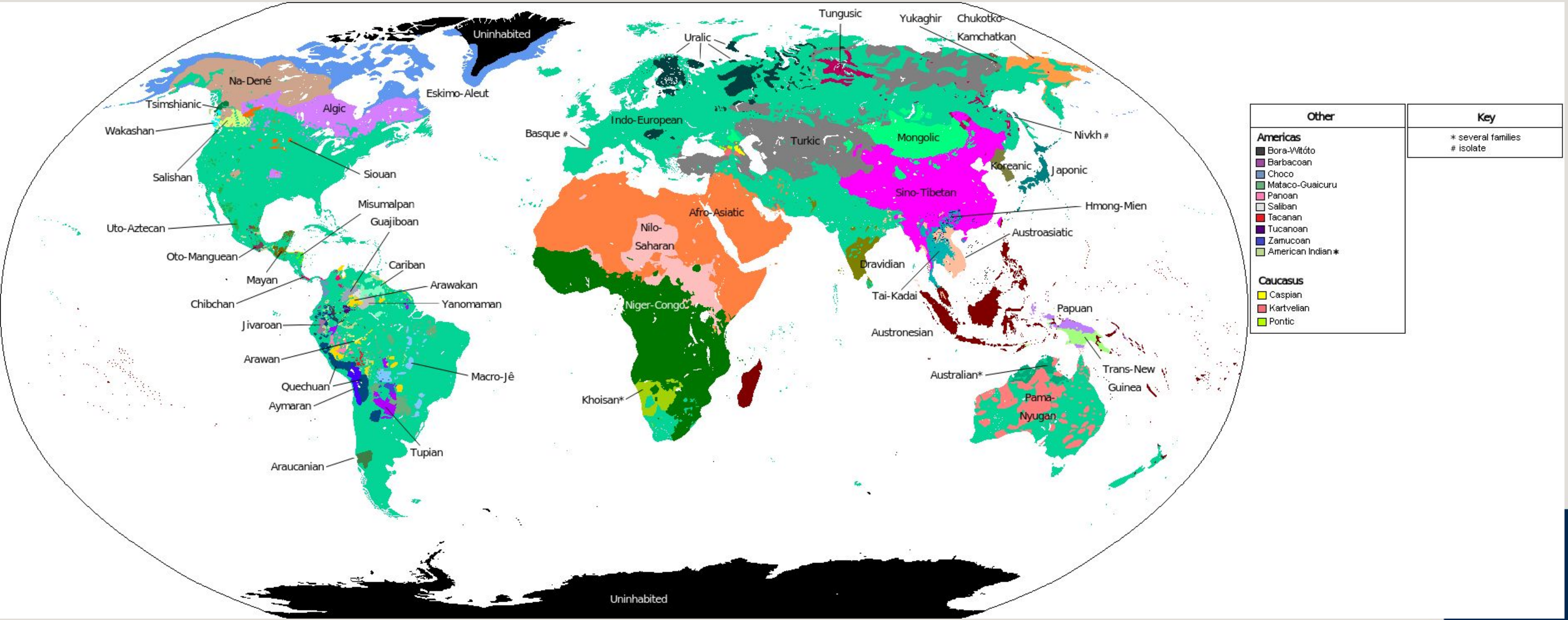
- Course Information
- What is NLP?
- Why Deep Learning for NLP?
- Course Content

Course Information

- My Contact
  - Email: pawang@cse.iitkgp.ac.in
  - Webpage: http://cse.iitkgp.ac.in/~pawang/
  - Course Page: https://sites.google.com/view/dl4nlp-nptel/home

- Teaching Assistants (Inaugural Course)
  - Subhendu Khatuya
  - Pretam Ray

Natural Language Processing



Natural Languages: Languages that evolved naturally through human use

Source: https://en.wikipedia.org/wiki/Language_family

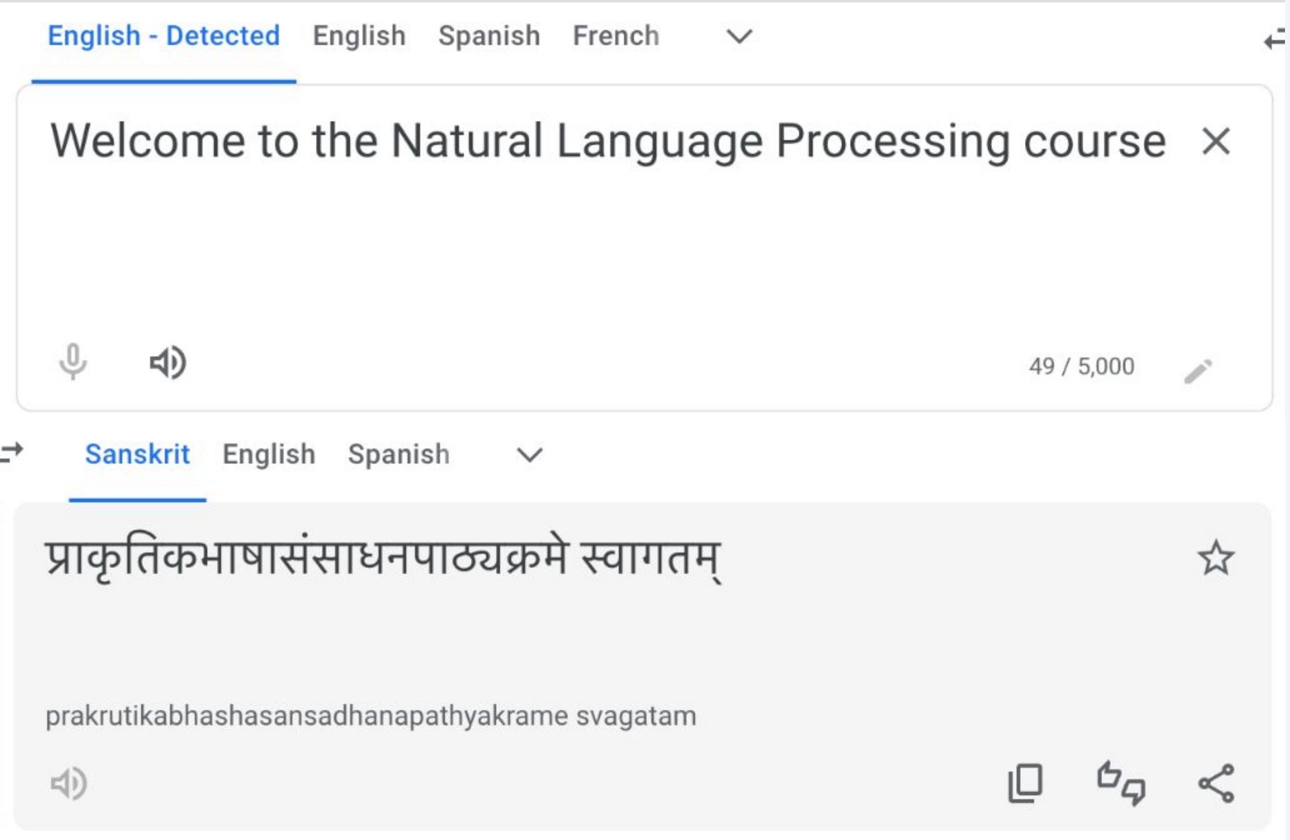
Natural Language Processing

What is NLP?

- *Making computers understand what we write (or speak)*
- *Making computers write (and speak)*

The field of NLP attempts to design, implement and test systems that process natural languages for practical applications

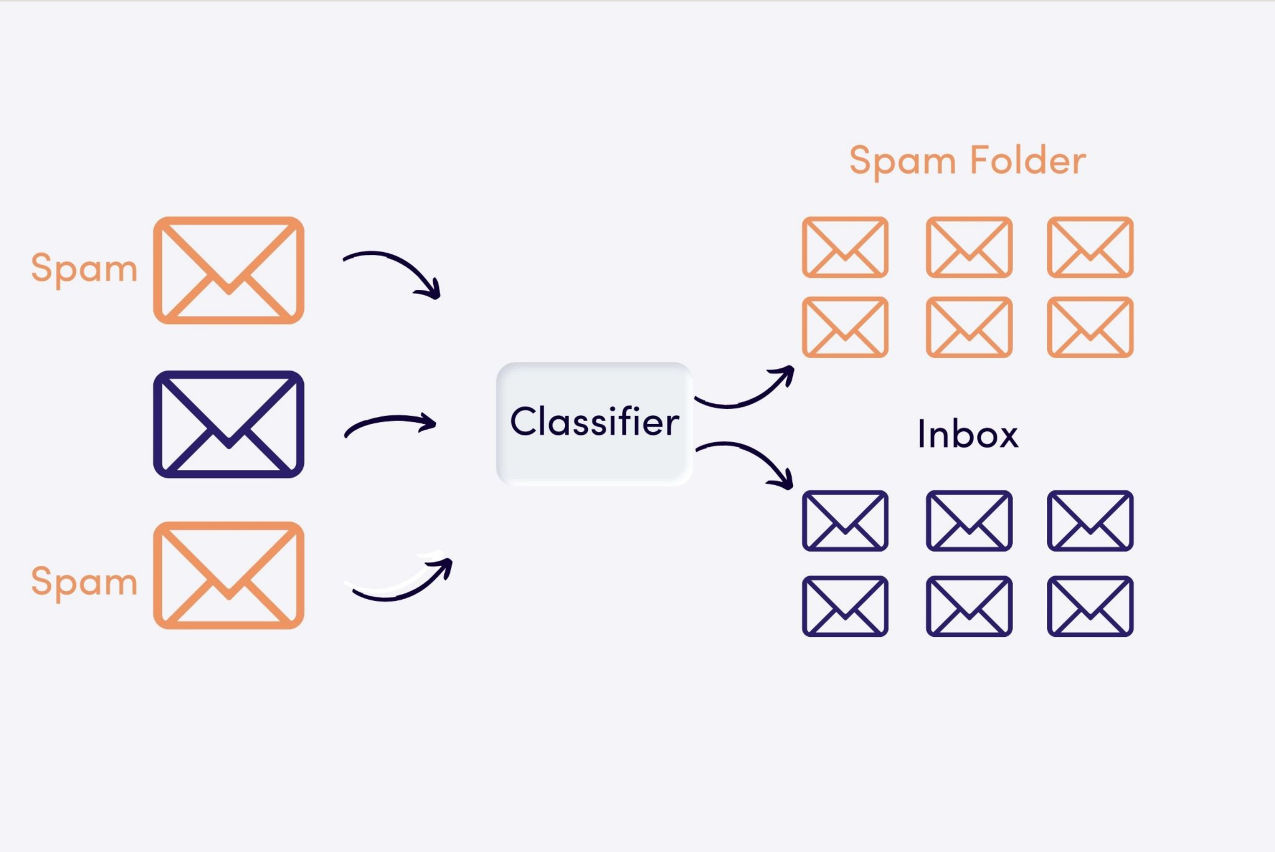
NLP Applications: NLP is everywhere!



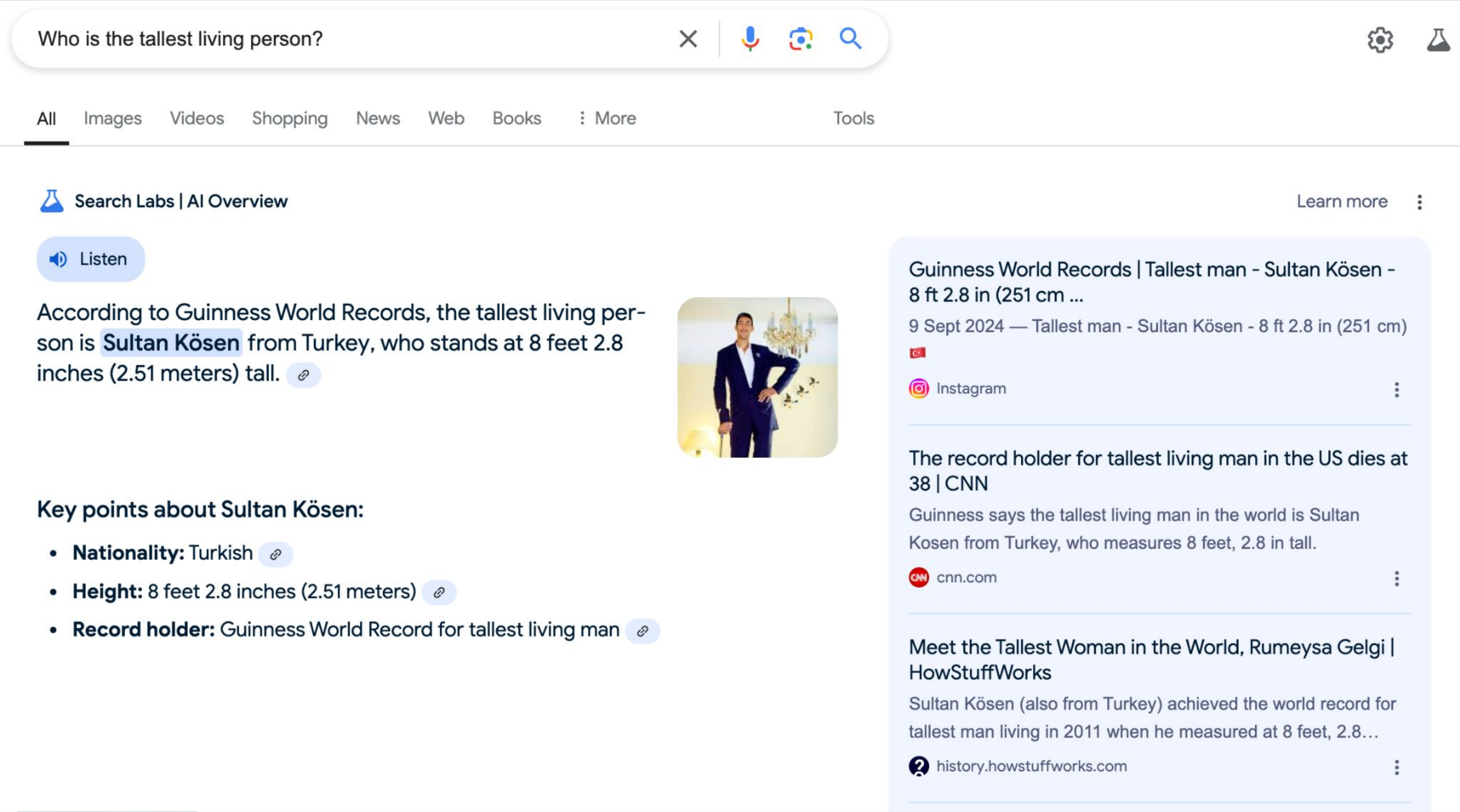
Welcome to the Natural Language Processing course

प्राकृतिकभाषासंसाधनपाठ्यक्रमे स्वागतम्

prakrutikabhashasansadhanapathyakrame svagatam



NLP is everywhere!



Who is the tallest living person?

Search Labs | AI Overview

According to Guinness World Records, the tallest living person is **Sultan Kösen** from Turkey, who stands at 8 feet 2.8 inches (2.51 meters) tall.

Key points about Sultan Kösen:

- **Nationality:** Turkish
- **Height:** 8 feet 2.8 inches (2.51 meters)
- **Record holder:** Guinness World Record for tallest living man

Guinness World Records | Tallest man - Sultan Kösen - 8 ft 2.8 in (251 cm ...

9 Sept 2024 — Tallest man - Sultan Kösen - 8 ft 2.8 in (251 cm) 🇹🇷

Instagram

The record holder for tallest living man in the US dies at 38 | CNN

Guinness says the tallest living man in the world is Sultan Kosen from Turkey, who measures 8 feet, 2.8 in tall.

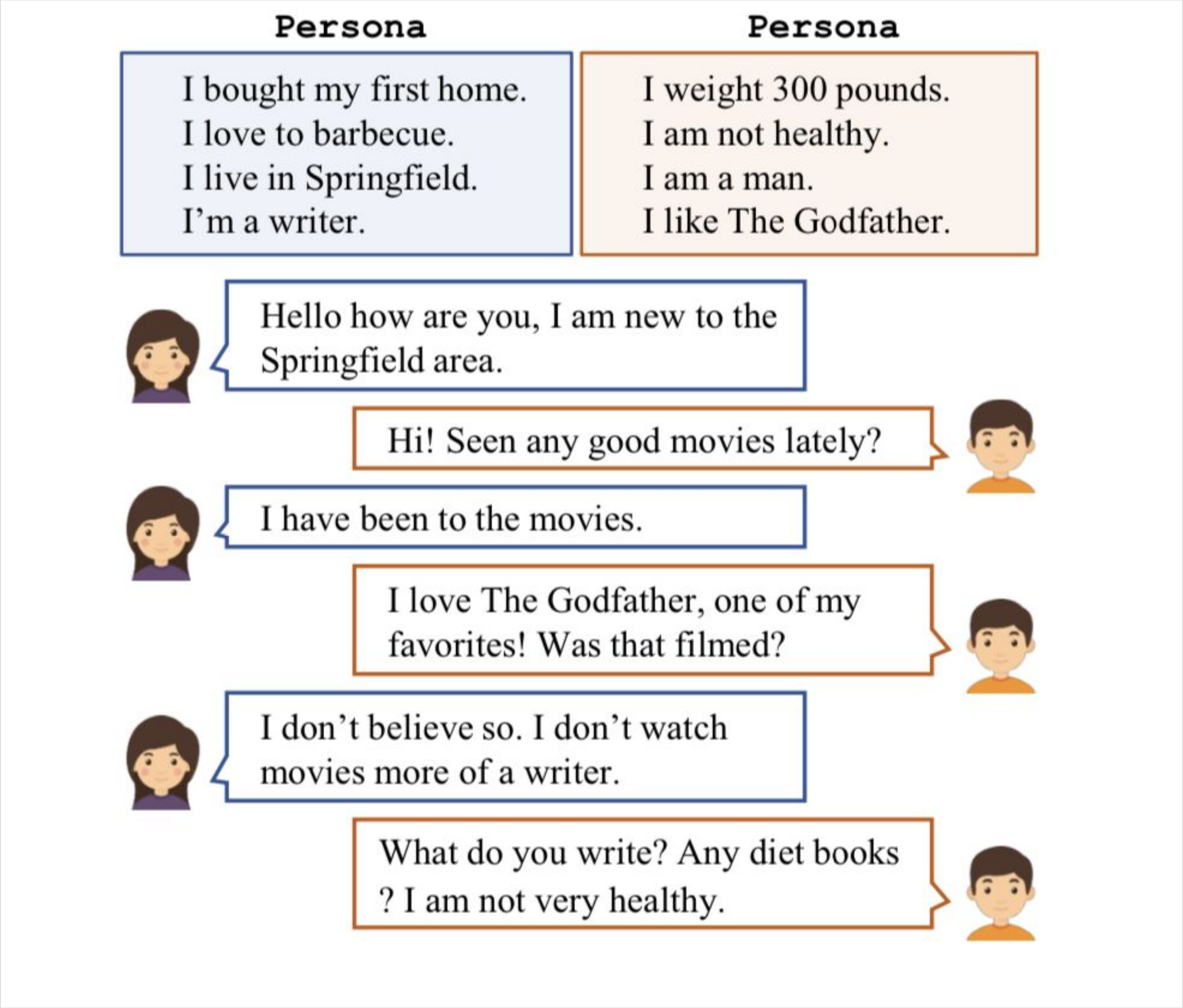
cnn.com

Meet the Tallest Woman in the World, Rumeysa Gelgi | HowStuffWorks

Sultan Kösen (also from Turkey) achieved the world record for tallest man living in 2011 when he measured at 8 feet, 2.8…

history.howstuffworks.com

NLP is everywhere!



Persona

I bought my first home.
I love to barbecue.
I live in Springfield.
I'm a writer.

Persona

I weight 300 pounds.
I am not healthy.
I am a man.
I like The Godfather.

Hello how are you, I am new to the Springfield area.

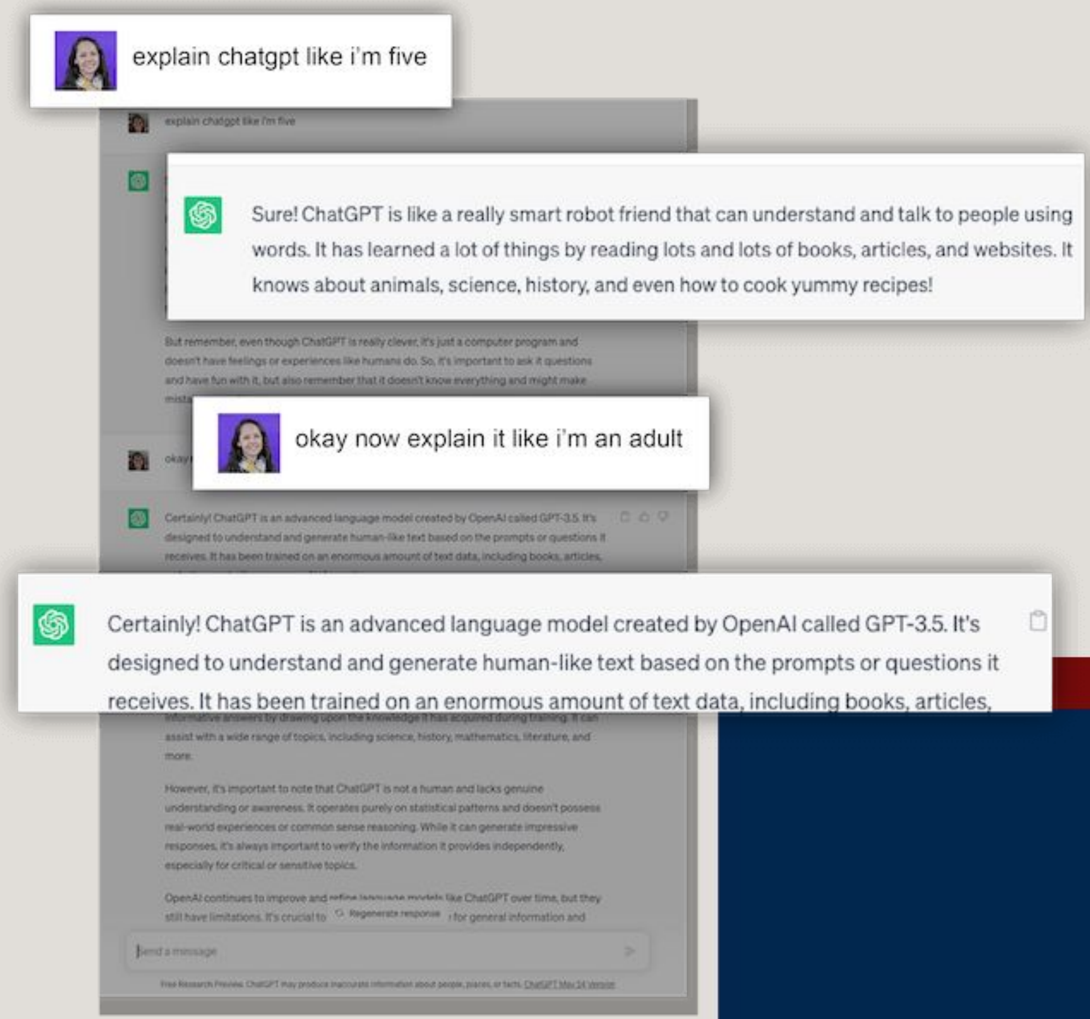
Hi! Seen any good movies lately?

I have been to the movies.

I love The Godfather, one of my favorites! Was that filmed?

I don't believe so. I don't watch movies more of a writer.

What do you write? Any diet books ? I am not very healthy.



explain chatgpt like i'm five

Sure! ChatGPT is like a really smart robot friend that can understand and talk to people using words. It has learned a lot of things by reading lots and lots of books, articles, and websites. It knows about animals, science, history, and even how to cook yummy recipes!

But remember, even though ChatGPT is really clever, it's just a computer program and doesn't have feelings or experiences like humans do. So, it's important to ask it questions and have fun with it, but also remember that it doesn't know everything and might make

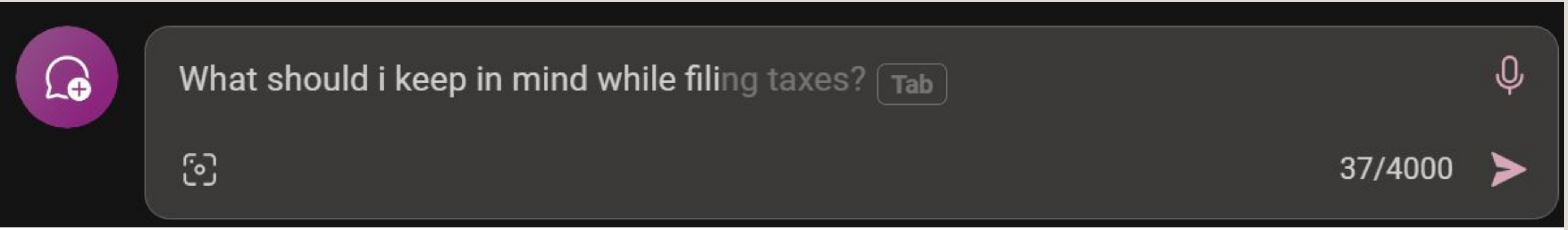
okay now explain it like i'm an adult

Certainly! ChatGPT is an advanced language model created by OpenAI called GPT-3.5. It's designed to understand and generate human-like text based on the prompts or questions it receives. It has been trained on an enormous amount of text data, including books, articles,

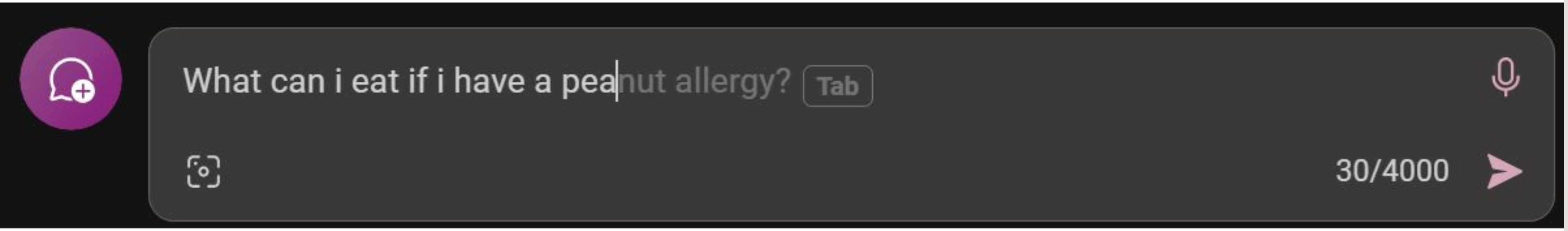
informative answers by drawing upon the knowledge it has acquired during training. It can assist with a wide range of topics, including science, history, mathematics, literature, and more.

However, it's important to note that ChatGPT is not a human and lacks genuine understanding or awareness. It operates purely on statistical patterns and doesn't possess real-world experiences or common sense reasoning. While it can generate impressive responses, it's always important to verify the information it provides independently, especially for critical or sensitive topics.

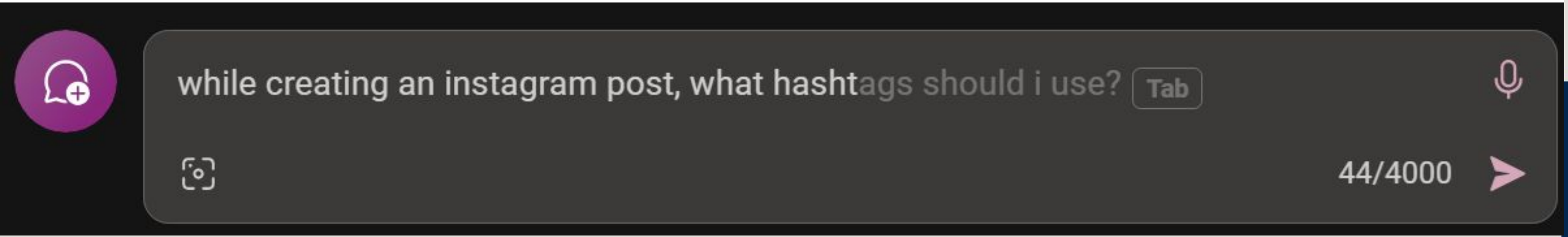
NLP is everywhere!



What should i keep in mind while filing taxes?

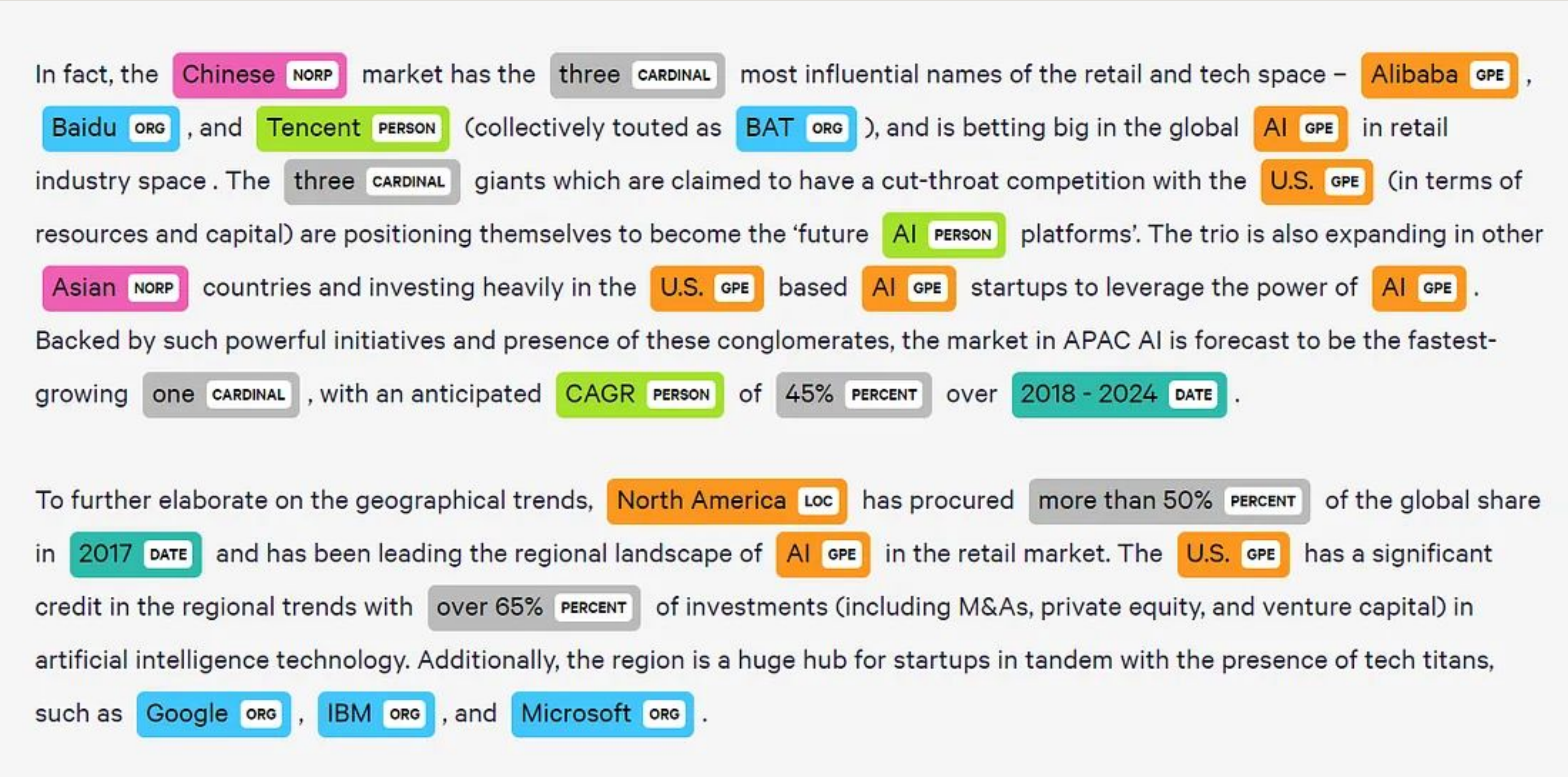


What can i eat if i have a peanut allergy?



while creating an instagram post, what hashtags should i use?

NLP is everywhere!

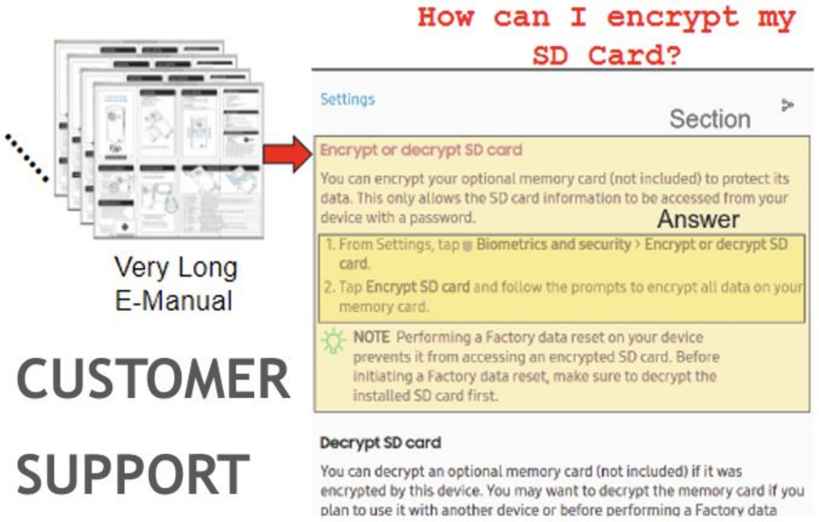


In fact, the Chinese NORP market has the three CARDINAL most influential names of the retail and tech space – Alibaba GPE , Baidu ORG , and Tencent PERSON (collectively touted as BAT ORG), and is betting big in the global AI GPE in retail industry space . The three CARDINAL giants which are claimed to have a cut-throat competition with the U.S. GPE (in terms of resources and capital) are positioning themselves to become the 'future AI PERSON platforms'. The trio is also expanding in other Asian NORP countries and investing heavily in the U.S. GPE based AI GPE startups to leverage the power of AI GPE . Backed by such powerful initiatives and presence of these conglomerates, the market in APAC AI is forecast to be the fastest-growing one CARDINAL , with an anticipated CAGR PERSON of 45% PERCENT over 2018 - 2024 DATE .

To further elaborate on the geographical trends, North America LOC has procured more than 50% PERCENT of the global share in 2017 DATE and has been leading the regional landscape of AI GPE in the retail market. The U.S. GPE has a significant credit in the regional trends with over 65% PERCENT of investments (including M&As, private equity, and venture capital) in artificial intelligence technology. Additionally, the region is a huge hub for startups in tandem with the presence of tech titans, such as Google ORG , IBM ORG , and Microsoft ORG .

Source: https://medium.com/@alessandropaticchio/named-entity-recognition-from-scratch-e76b9b3affad

Domain Specific Applications



Encrypt or decrypt SD card

You can encrypt your optional memory card (not included) to protect its data. This only allows the SD card information to be accessed from your device with a password.

1. From Settings, tap Biometrics and security › Encrypt or decrypt SD card.
2. Tap Encrypt SD card and follow the prompts to encrypt all data on your memory card.

NOTE Performing a Factory data reset on your device prevents it from accessing an encrypted SD card. Before initiating a Factory data reset, make sure to decrypt the installed SD card first.

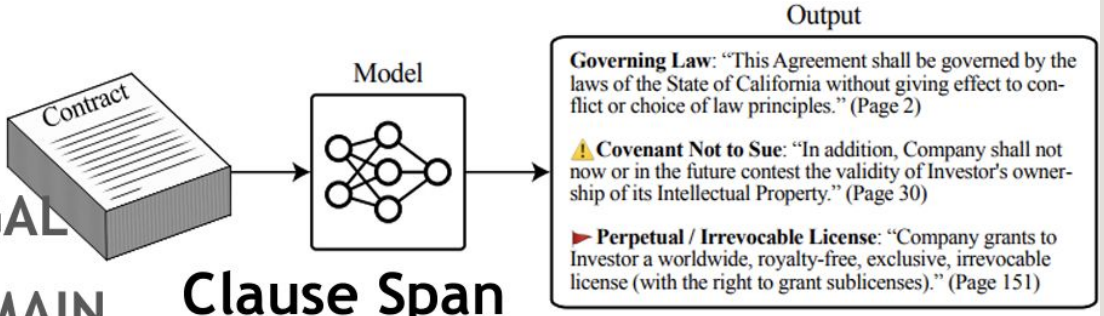
Decrypt SD card

You can decrypt an optional memory card (not included) if it was encrypted by this device. You may want to decrypt the memory card if you plan to use it with another device or before performing a Factory data

CUSTOMER SUPPORT DOMAIN

Question Answering

LEGAL DOMAIN

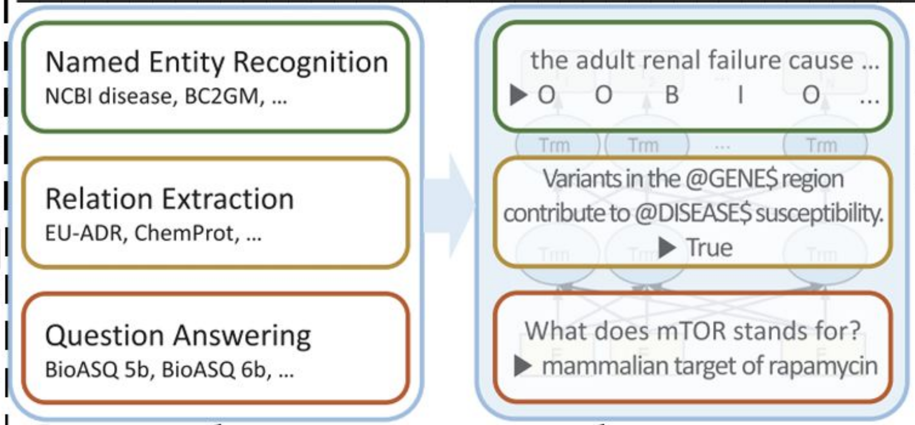


Governing Law: "This Agreement shall be governed by the laws of the State of California without giving effect to conflict or choice of law principles." (Page 2)

⚠ **Covenant Not to Sue**: "In addition, Company shall not now or in the future contest the validity of Investor's ownership of its Intellectual Property." (Page 30)

► **Perpetual / Irrevocable License**: "Company grants to Investor a worldwide, royalty-free, exclusive, irrevocable license (with the right to grant sublicenses)." (Page 151)

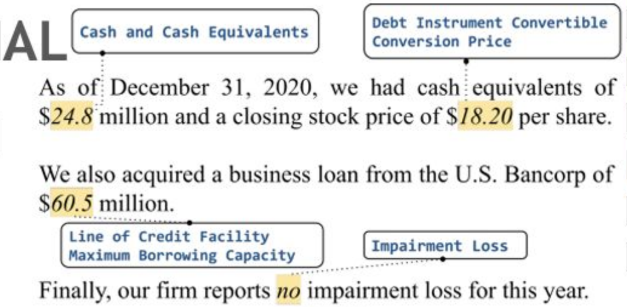
Clause Span Extraction


Named Entity Recognition
NCBI disease, BC2GM, ...
Relation Extraction
EU-ADR, ChemProt, ...
Question Answering
BioASQ 5b, BioASQ 6b, ...
the adult renal failure cause ...
▶ O O B I O ...
Trm
Trm
...
Trm
Variants in the @GENE$ region contribute to @DISEASE$ susceptibility.
▶ True
What does mTOR stands for?
▶ mammalian target of rapamycin

BIOMEDICAL DOMAIN

Several sentence and token level tasks

FINANCIAL DOMAIN


Cash and Cash Equivalents
Debt Instrument Convertible Conversion Price
Line of Credit Facility Maximum Borrowing Capacity
Impairment Loss

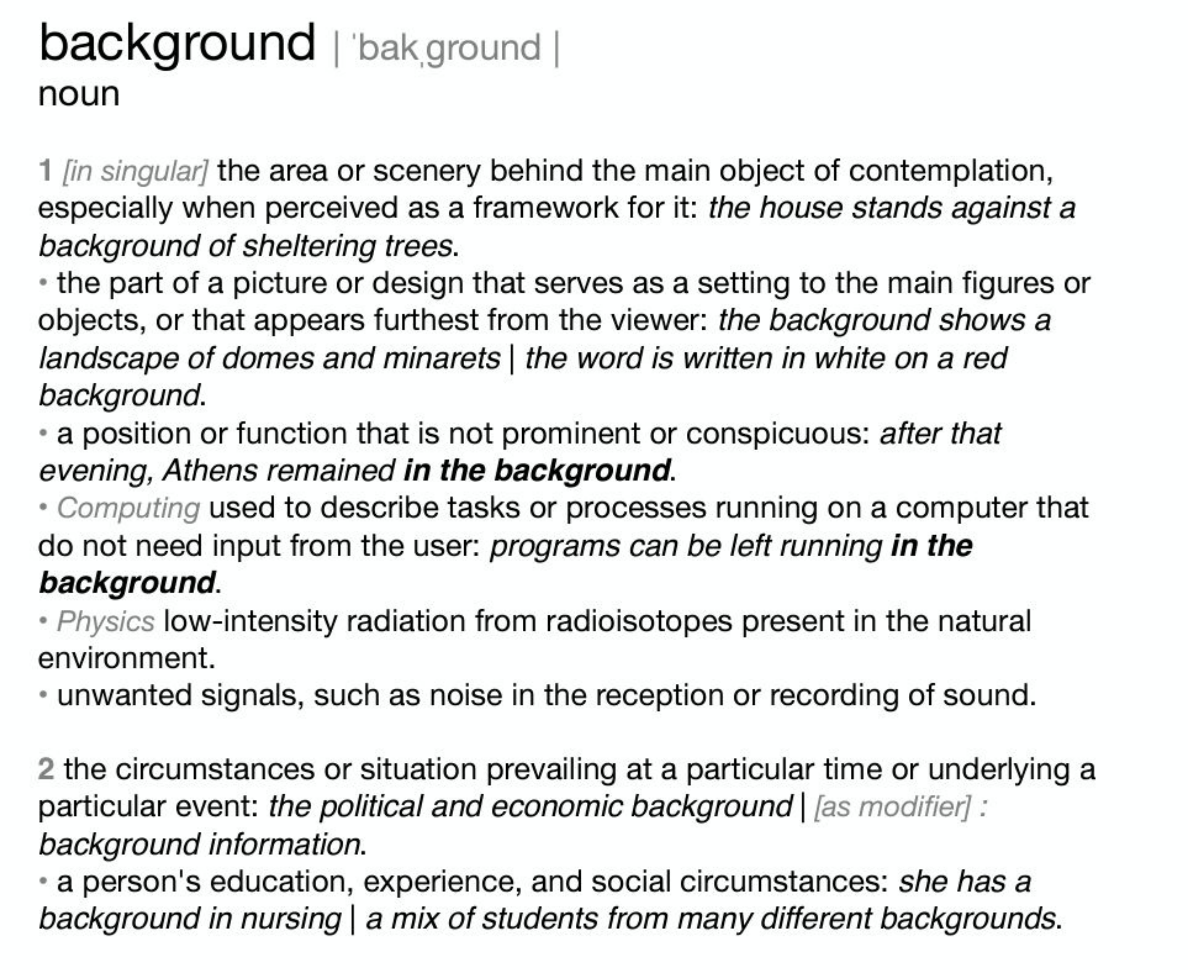
As of December 31, 2020, we had cash equivalents of \$24.8 million and a closing stock price of \$18.20 per share.

We also acquired a business loan from the U.S. Bancorp of \$60.5 million.

Finally, our firm reports no impairment loss for this year.

NER

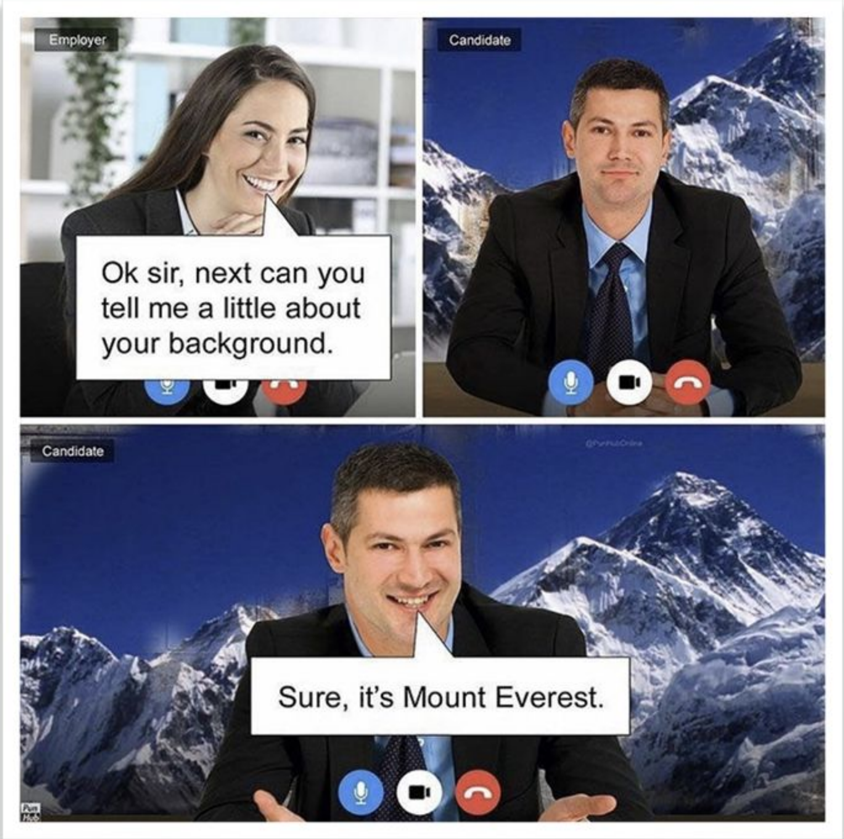
Why is NLP Hard? Language Ambiguity



background | ˈbakˌground |
noun

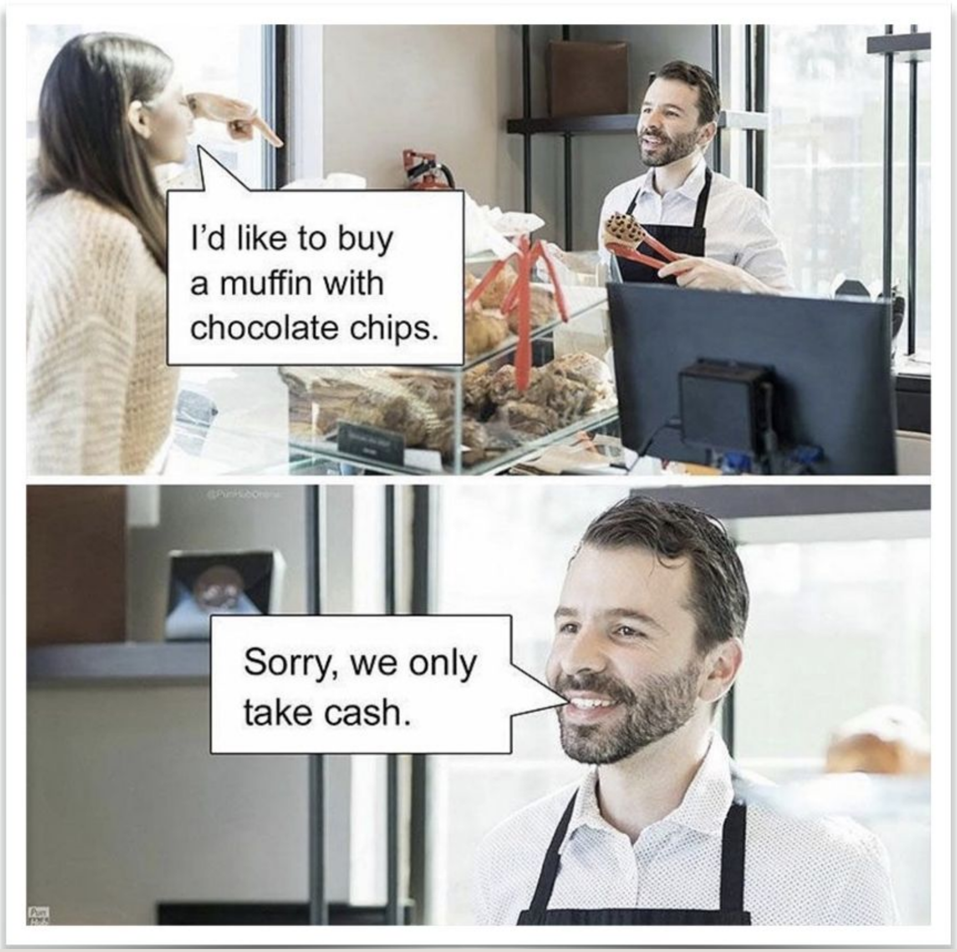
1 *[in singular]* the area or scenery behind the main object of contemplation, especially when perceived as a framework for it: *the house stands against a background of sheltering trees*.
- the part of a picture or design that serves as a setting to the main figures or objects, or that appears furthest from the viewer: *the background shows a landscape of domes and minarets* | *the word is written in white on a red background*.
- a position or function that is not prominent or conspicuous: *after that evening, Athens remained* ***in the background***.
- *Computing* used to describe tasks or processes running on a computer that do not need input from the user: *programs can be left running* ***in the background***.
- *Physics* low-intensity radiation from radioisotopes present in the natural environment.
- unwanted signals, such as noise in the reception or recording of sound.

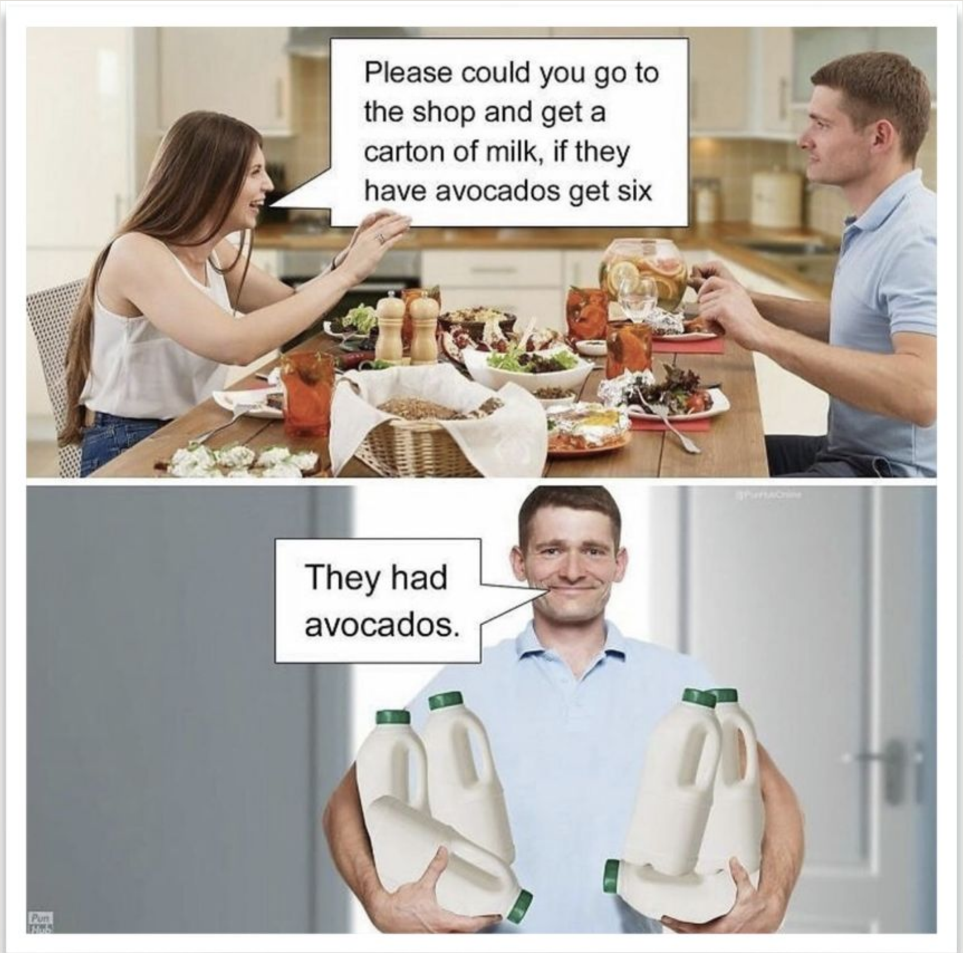
2 the circumstances or situation prevailing at a particular time or underlying a particular event: *the political and economic background* | *[as modifier]* : *background information*.
- a person's education, experience, and social circumstances: *she has a background in nursing* | *a mix of students from many different backgrounds*.


Employer
Candidate
Ok sir, next can you tell me a little about your background.
Candidate
Sure, it's Mount Everest.

Source: https://courses.cs.cornell.edu/courses/cs5740

Why is NLP Hard? Language Ambiguity


I'd like to buy a muffin with chocolate chips.
Sorry, we only take cash.


Please could you go to the shop and get a carton of milk, if they have avocados get six
They had avocados.

Source: https://courses.cs.cornell.edu/courses/cs5740

Why is NLP Hard? Language Ambiguity

Let’s try to decipher this weird conversation!

Rahul: *I saw a monkey with a banana.*

Computer: *That’s gruesome!*

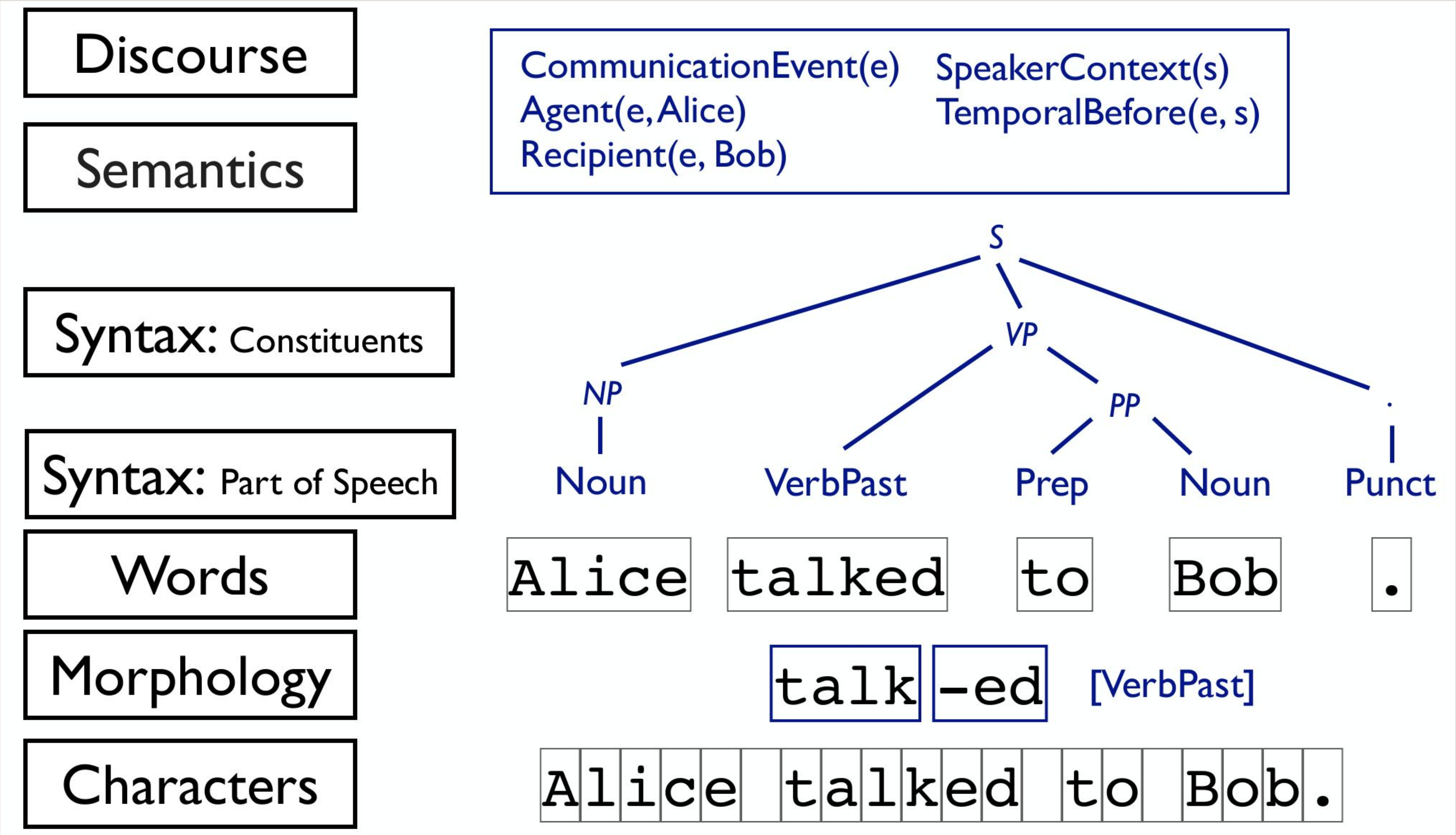
Rahul: *Why? What’s so gruesome about seeing a monkey?*

Computer: *Oh I see! What else did you see with the banana?*

In Natural Languages, ambiguity is the rule, not an exception

Example: Courtesy Dr. Monojit choudhury

NLP: Levels of Linguistic Structure

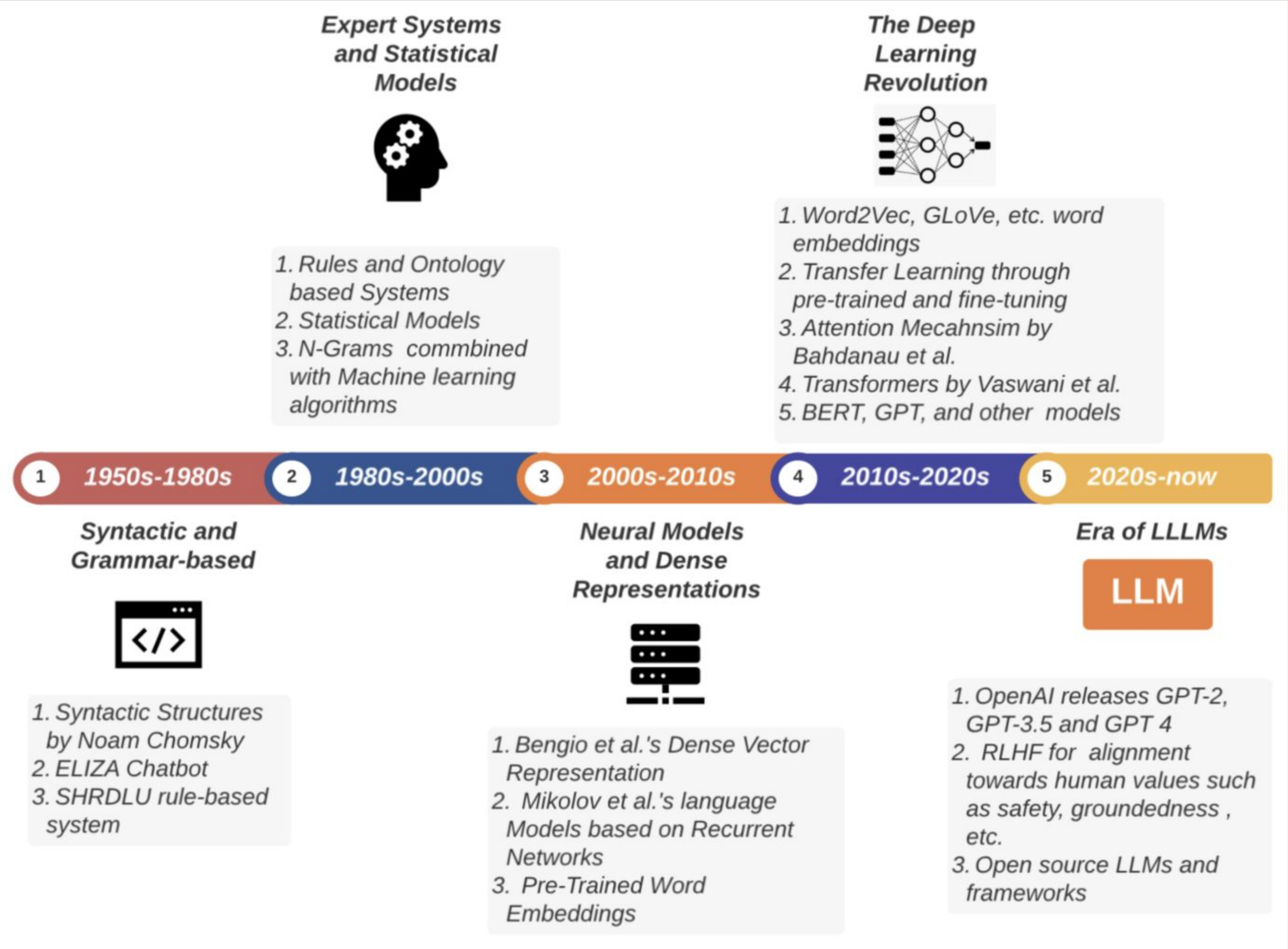


Source: https://people.cs.umass.edu/~miyyer/cs685

NLP Paradigms

We generally try to map problems to various (ML) paradigms

- Sentiment Analysis, news article groupings, etc. → Text Classification
- Named entity recognition, code-mixing, etc. → Sequence Labeling
- Machine Translation, summarization, chatbots, etc. → Text Generation



1 **1950s-1980s**

Syntactic and Grammar-based

1. Syntactic Structures by Noam Chomsky
2. ELIZA Chatbot
3. SHRDLU rule-based system

2 **1980s-2000s**

Expert Systems and Statistical Models

1. Rules and Ontology based Systems
2. Statistical Models
3. N-Grams commbined with Machine learning algorithms

3 **2000s-2010s**

Neural Models and Dense Representations

1. Bengio et al.'s Dense Vector Representation
2. Mikolov et al.'s language Models based on Recurrent Networks
3. Pre-Trained Word Embeddings

4 **2010s-2020s**

The Deep Learning Revolution

1. Word2Vec, GLoVe, etc. word embeddings
2. Transfer Learning through pre-trained and fine-tuning
3. Attention Mecahnsim by Bahdanau et al.
4. Transformers by Vaswani et al.
5. BERT, GPT, and other models

5 **2020s-now**

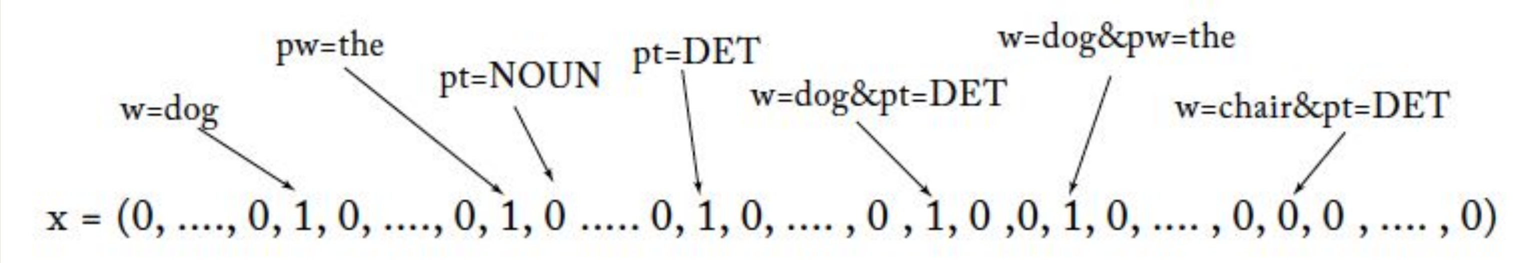
Era of LLLMs

1. OpenAI releases GPT-2, GPT-3.5 and GPT 4
2. RLHF for alignment towards human values such as safety, groundedness , etc.
3. Open source LLMs and frameworks

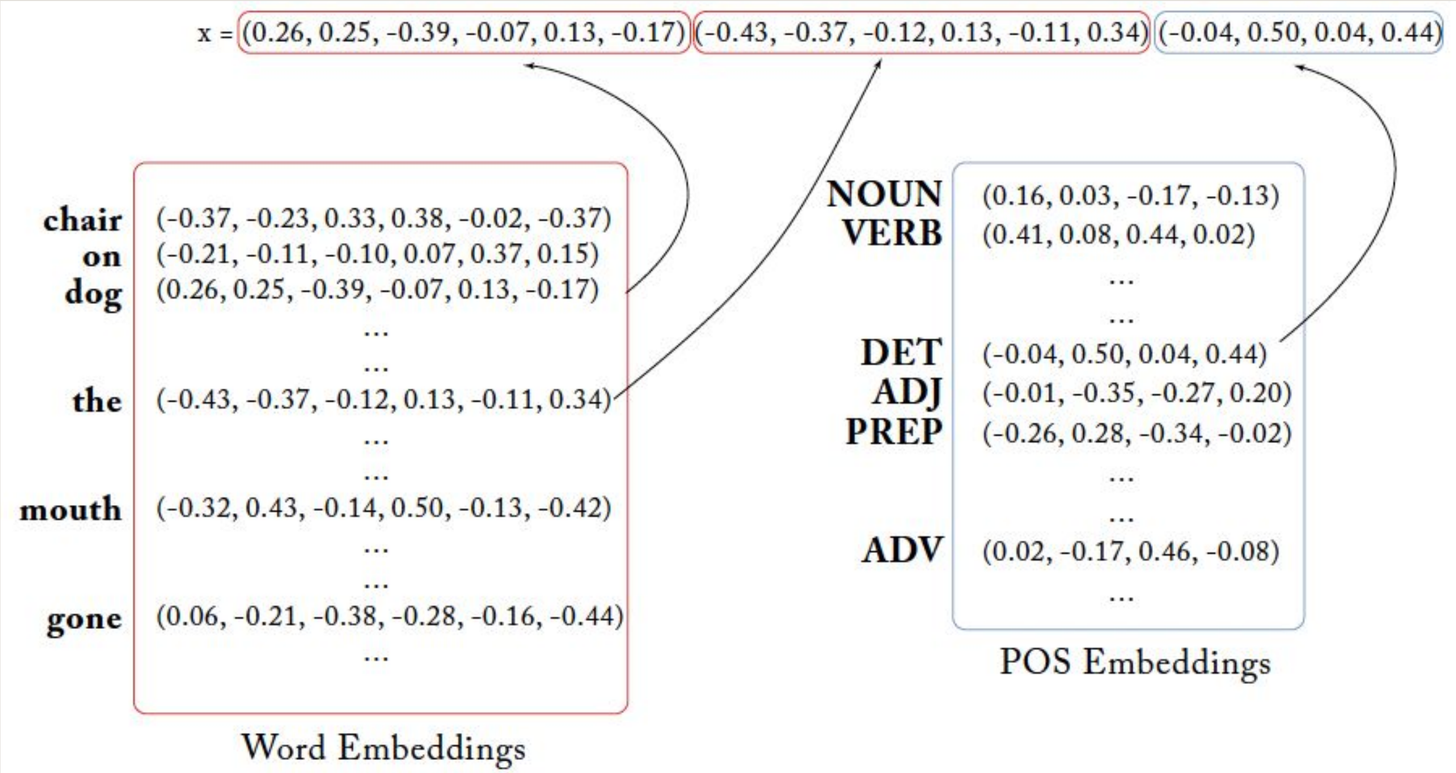
Timeline illustrating the progression of NLP from the 1950s

Source: *Kamath, Uday, et al. "Large Language Models: A Deep Dive." (2024).*

Why Deep Learning?



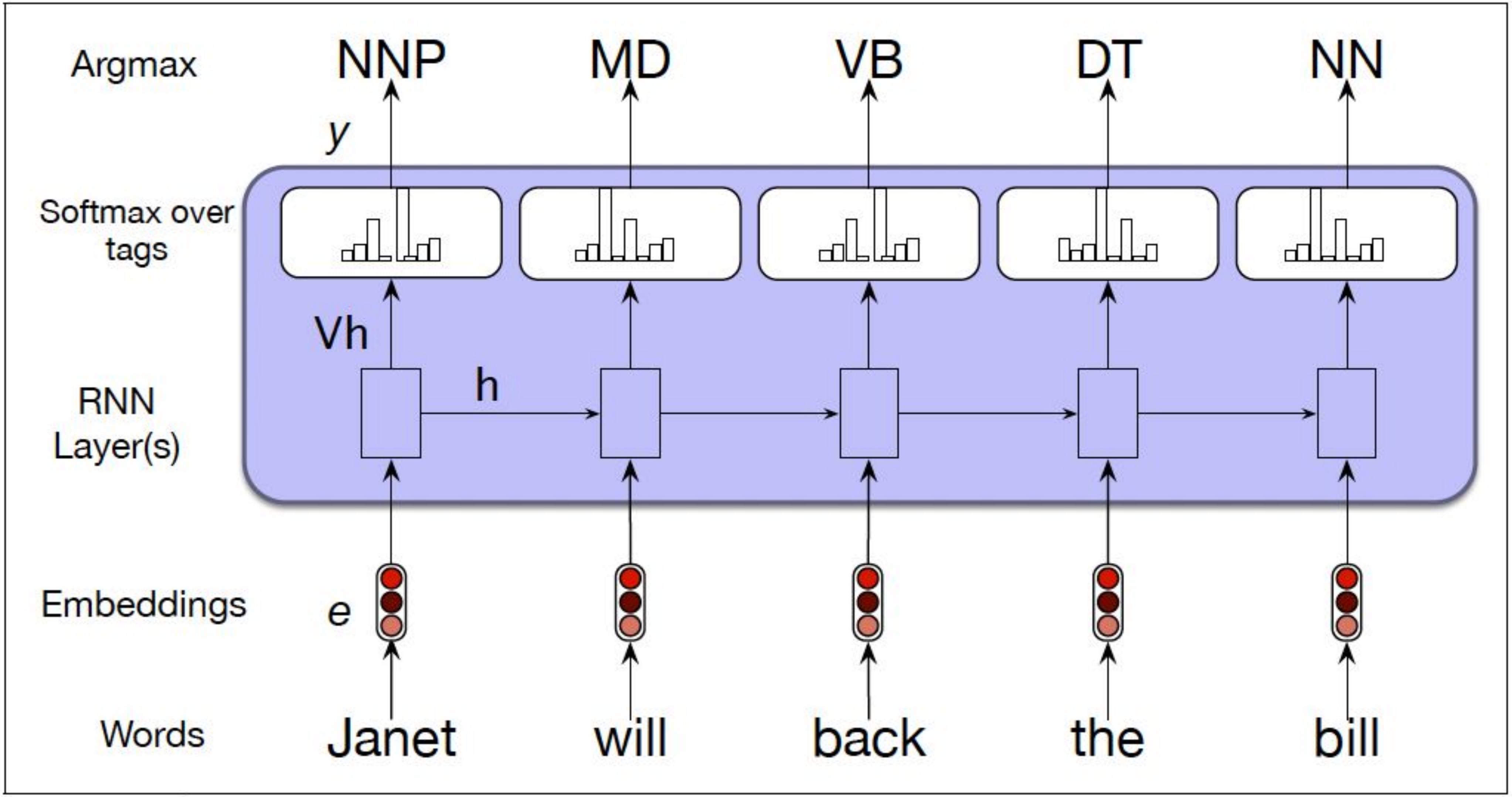
w=dog
pw=the
pt=NOUN
pt=DET
w=dog&pt=DET
w=dog&pw=the
w=chair&pt=DET
x = (0, ..., 0, 1, 0, ..., 0, 1, 0 ..... 0, 1, 0, ... , 0 , 1, 0 ,0, 1, 0, ... , 0, 0, 0 , ... , 0)



x = (0.26, 0.25, -0.39, -0.07, 0.13, -0.17) (-0.43, -0.37, -0.12, 0.13, -0.11, 0.34) (-0.04, 0.50, 0.04, 0.44)
chair (-0.37, -0.23, 0.33, 0.38, -0.02, -0.37)
on (-0.21, -0.11, -0.10, 0.07, 0.37, 0.15)
dog (0.26, 0.25, -0.39, -0.07, 0.13, -0.17)
...
...
the (-0.43, -0.37, -0.12, 0.13, -0.11, 0.34)
...
...
mouth (-0.32, 0.43, -0.14, 0.50, -0.13, -0.42)
...
...
gone (0.06, -0.21, -0.38, -0.28, -0.16, -0.44)
...
Word Embeddings
NOUN (0.16, 0.03, -0.17, -0.13)
VERB (0.41, 0.08, 0.44, 0.02)
...
...
DET (-0.04, 0.50, 0.04, 0.44)
ADJ (-0.01, -0.35, -0.27, 0.20)
PREP (-0.26, 0.28, -0.34, -0.02)
...
...
ADV (0.02, -0.17, 0.46, -0.08)
...
POS Embeddings

Sparse vs. dense feature representations. Two encodings of the information: current word is “dog;” previous word is “the” previous pos-tag is “DET.”

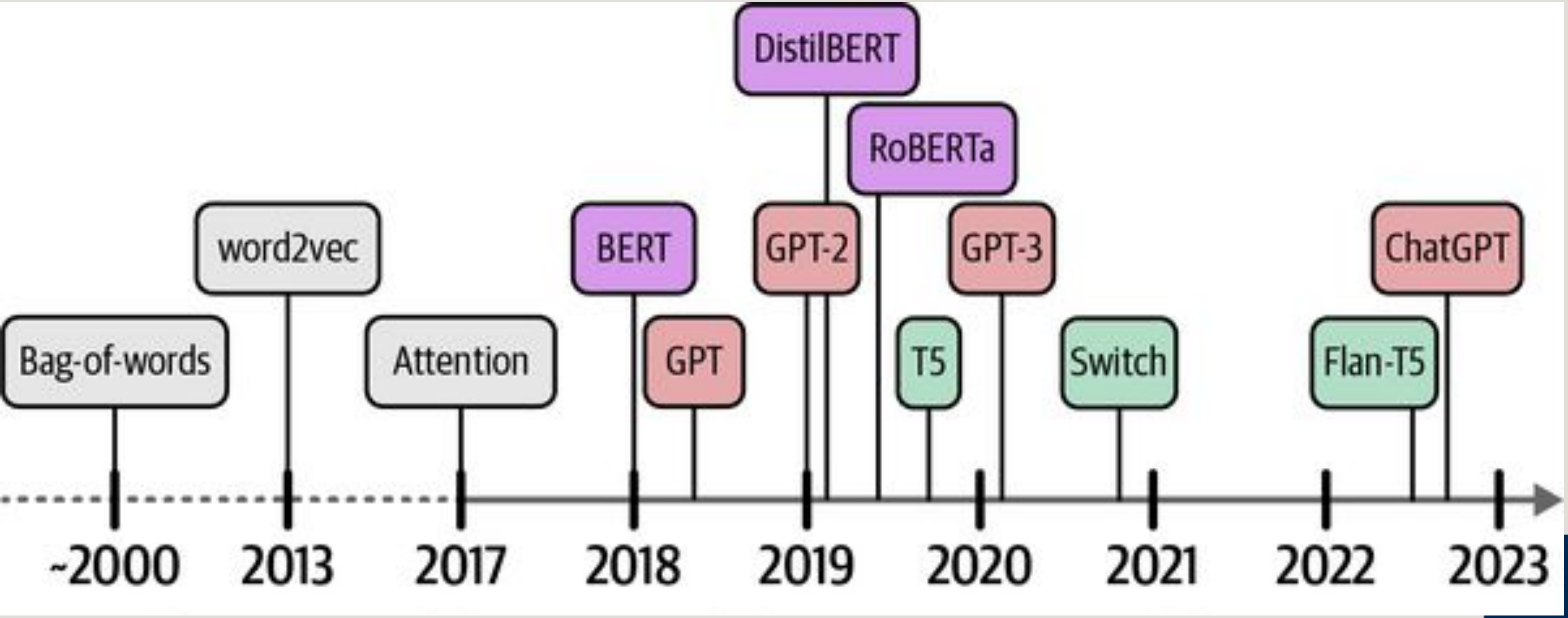
Source: *Yoav Goldberg, Graeme Hirst. Neural Network Methods in Natural Language Processing, Morgan & Claypool Publishers (2017).*



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

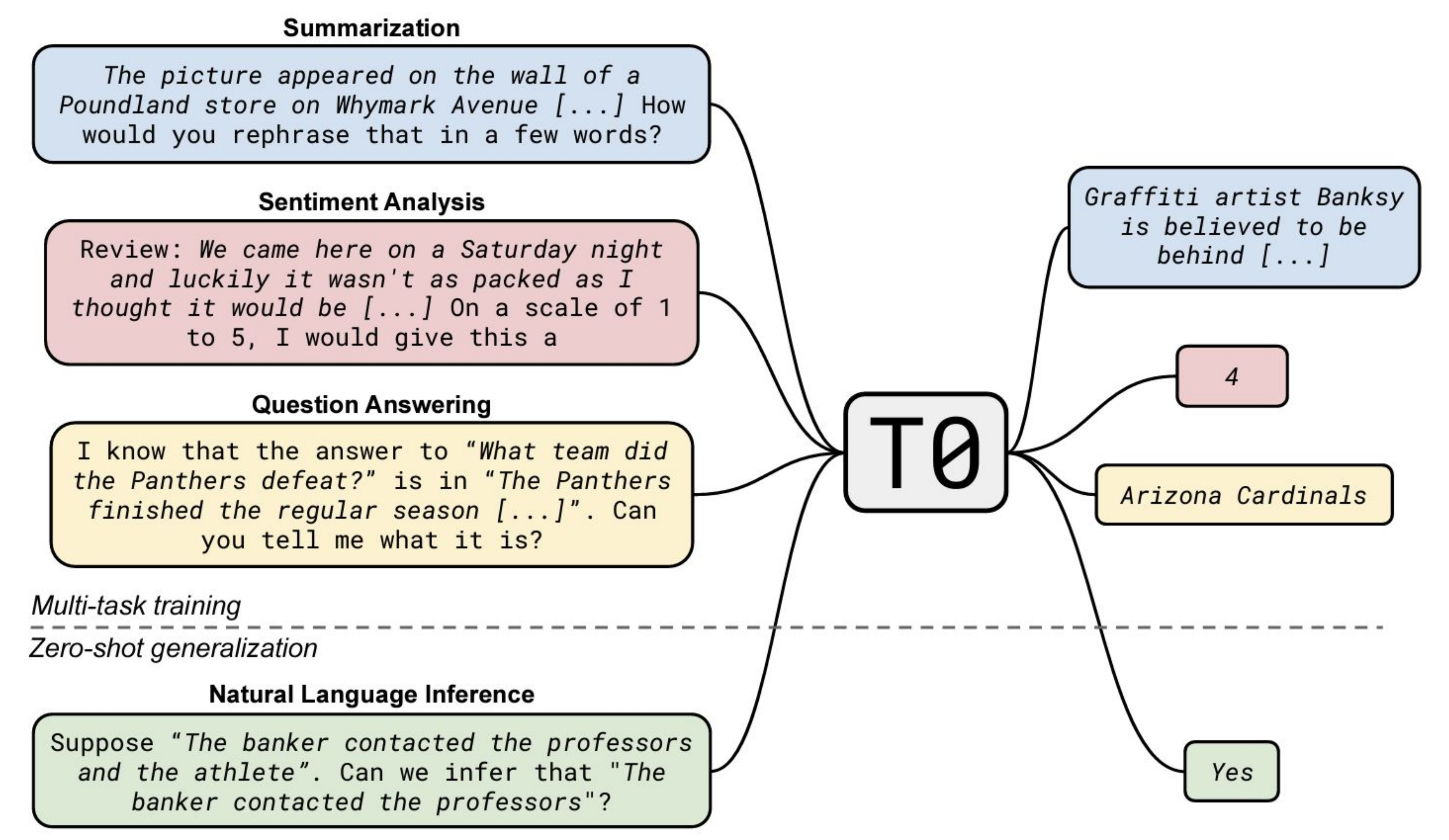
These dense feature representations are used with various deep-learning architectures

A timeline of the recent developments



Source: Alammar, J., & Grootendorst, M. (2024). Hands-On Large Language Models. O'Reilly.

Change of NLP paradigms: Just use generation!



Summarization

The picture appeared on the wall of a Poundland store on Whymark Avenue [...] How would you rephrase that in a few words?

Sentiment Analysis

Review: *We came here on a Saturday night and luckily it wasn't as packed as I thought it would be [...]* On a scale of 1 to 5, I would give this a

Question Answering

I know that the answer to "*What team did the Panthers defeat?*" is in "*The Panthers finished the regular season [...]*". Can you tell me what it is?

Multi-task training

Zero-shot generalization

Natural Language Inference

Suppose "*The banker contacted the professors and the athlete*". Can we infer that "*The banker contacted the professors*"?

T0

Graffiti artist Banksy is believed to be behind [...]

4

Arizona Cardinals

Yes

Sanh, Victor, et al. "Multitask Prompted Training Enables Zero-Shot Task Generalization." *ICLR 2022*

Course Content (Weeks 1-6)

Background

- Introduction to NLP
- Introduction to Deep Learning and Representation Learning
- Word Representation: Word2Vec, Glove, FastText, Multilingual

Models and Architectures

- Recurrent Neural Networks: RNNs, LSTMs, Sequence to Sequence
- Attention Mechanism and Transformers: Attention in RNNs, Self-Attention in Transformers

Methods

- Pretraining: Self-supervised Learning objectives for Pretraining, ELMo, BERT, GPT, T5, BART, Fine-tuning

Course Content (Weeks 7-12)

Tasks

- Question Answering, Text Summarization, Dialogs
- Domain and language-specific applications and challenges

Methods (LLMs)

- Towards building LLMs as chat assistants: Instruction Fine-tuning, Reinforcement learning from human feedback, Alignment techniques
- In-content learning, chain-of-thought prompting, Various LLMs
- Parameter Efficient Fine-tuning (PEFT), LoRA, QLoRA
- Handling Long Context, Retrieval Augmented Generation (RAG)

Conclusion

- Analysis and Interpretability, ethical considerations

REFERENCES

Daniel Jurafsky and James H. Martin. 2024. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 3rd edition. Online manuscript released August 20, 2024. https://web.stanford.edu/~jurafsky/slp3.

Alammar, J., & Grootendorst, M. (2024). Hands-On Large Language Models. O'Reilly.

Yoav Goldberg, Graeme Hirst. Neural Network Methods in Natural Language Processing, Morgan & Claypool Publishers (2017).

Kamath, Uday, et al. "Large Language Models: A Deep Dive." (2024).

THANK YOU