

---

# Introduction to Human Language Technology

Philipp Koehn

27 August 2024



# Administrative



- **Coordinator:** Philipp Koehn (phi@jhu.edu)
- **Lecturers:** Faculty of the Center for Language and Speech Processing (CLSP)
- **TA:** Elina Baral (ebaral1@jhu.edu)
- **Class:** Tuesday, Thursday, 9:00-10:15, Hodson 210
- **Course web site:** <https://jhu-intro-hlt.github.io/>
- **Grading**
  - 4 assignments (12.5% each)
  - first midterm exam (15%)
  - second midterm exam (15%)
  - final exam (20%)

- Human Language Technology
  - Speech: spoken language (audio)
  - Text: written language (text)■
- What is it good for?
  - Means of Communication
    - new ways of interacting with computers
  - Storage medium for knowledge
    - new ways of making word knowledge available■
- This course will introduce
  - methods and tools used in HLT
  - overview of HLT applications
  - JHU faculty working on speech and language

# Course Overview: Text



3

- Words, Morphology (Yarowsky)
- Syntax (Post)
- Semantics (Lippincott, Koehn)
- Deep Learning (Murray)
- Information retrieval and extraction (Koehn, Duh)
- Machine translation (Duh)

# Course Overview: Speech



- Audio signals, phonemes, graphemes, dictionaries (Moro-Velazquez)
- Auditory system (Elhilali)
- Speech recognition: HMM (Khudanpur)
- End-to-end neural speech recognition (Khudanpur)
- Enhancement and diarization (Villaba, Garcia)

# Course Overview: Applications



5

- NLP for Digital Humanities (Lippincott)
- Interpretable and Explainable NLP (Chen)
- Question answering (Duh)
- Computational Social Science (Field)
- Ethical problems (Moro-Velazquez)
- Scaling large language models (Khashabi)

# Master Concentration in HLT



<https://www.clsp.jhu.edu/human-language-technology-masters/>

- Concentration in Human Language Technology
  - Master in Computer Science
  - Master in Electrical and Computer Engineering
- Requirements (in addition to usual degree requirements)
  - Introduction to Human Language Technology (601.667)
  - Natural Language Processing (601.665)
  - Information Extraction from Speech and Text (520.666)
  - Master project in HLT



- One of the largest and most influential academic research centers in HLT
- Faculty in Computer Science, Electrical and Computer Engineering, Cognitive Science, Mathematical Sciences, ...
- Home of over 60 researchers, dozens of PhD students
- Founded in 1992 by Frederick Jelinek (1932-2010)
- Sibling center: Human Language Technology Center of Excellence (HLTCOE)



# Speech Recognition



# Information Retrieval



information retrieval



All



Books



News



Images



Videos

More

Settings

Tools

About 57,300,000 results (0.74 seconds)

## Information retrieval - Wikipedia

<https://en.wikipedia.org> › [wiki](#) › [Information\\_retrieval](#) ▼

**Information retrieval** is the science of searching for **information** in a document, searching for documents themselves, and also searching for the metadata that describes data, and for databases of texts, images or sounds.

[Information retrieval](#) · [Evaluation measures](#) · [Applications](#) · [Boolean model](#)

## Introduction to Information Retrieval - Stanford NLP Group

<https://nlp.stanford.edu> › [IR-book](#) › [html](#) › [htmledition](#) › [irbook](#) ▼

Website: <http://informationretrieval.org/>. Cambridge ... **informationretrieval** (at) yahoo groups.com ... Statistical properties of terms in **information retrieval**.

## <sup>[PDF]</sup> Introduction to Information Retrieval - Stanford NLP Group

<https://nlp.stanford.edu> › [IR-book](#) › [pdf](#) ▼

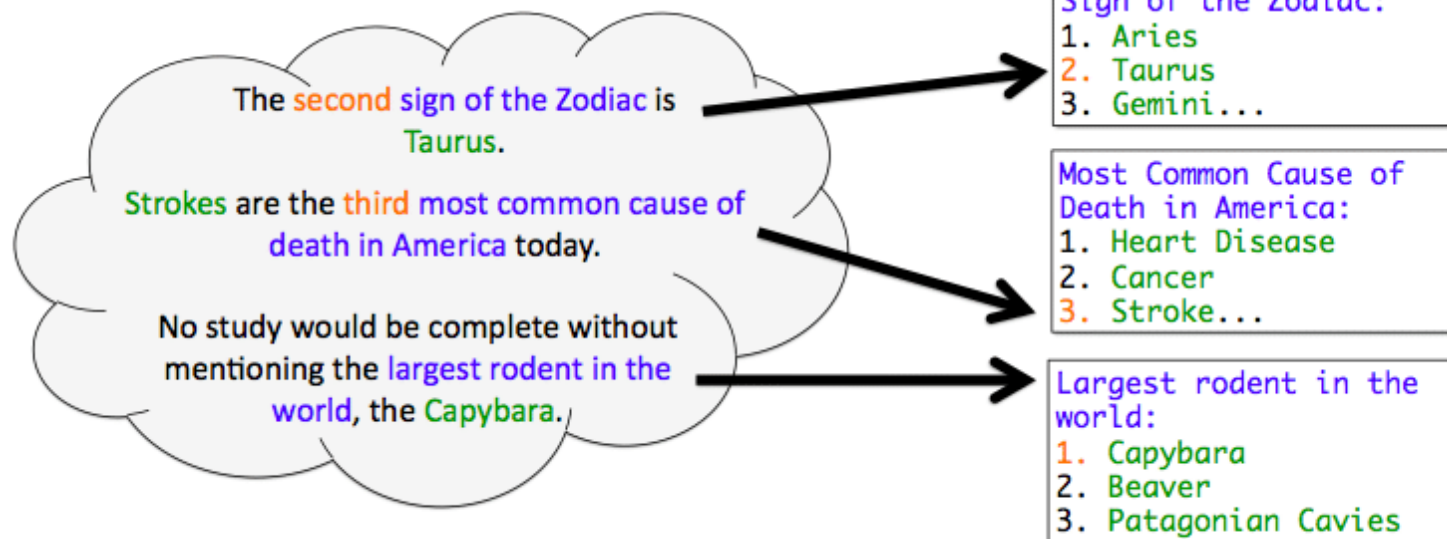
As defined in this way, **information retrieval** used to be an activity that only a few people engaged in: reference librarians, paralegals, and similar pro- fessional ...

## Information Retrieval - an overview | ScienceDirect Topics

## Unstructured Web Text



## Structured Sequences





A l'orée de ce débat télévisé inédit dans l'histoire de la Ve République, on attendait une forme de «Tous sur Macron» mais c'est la candidate du Front national qui s'est retrouvée au cœur des premières attaques de ses quatre adversaires d'un soir, favorisées par le premier thème abordé, les questions de société et donc de sécurité, d'immigration et de laïcité.

At the beginning of this televised debate, which was unheard of in the history of the Fifth Republic, a "Tous sur Macron" was expected, but it was the candidate of the National Front who found itself at the heart of the first attacks of its four Opponents of one evening, favored by the first theme tackled, the issues of society and thus security, immigration and secularism.



The first recorded travels by Europeans to China and back date from this time. The most famous traveler of the period was the Venetian Marco Polo, whose account of his trip to "Cambaluc," the capital of the Great Khan, and of life there astounded the people of Europe. The account of his travels, *Il milione* (or, *The Million*, known in English as the *Travels of Marco Polo*), appeared about the year 1299. Some argue over the accuracy of Marco Polo's accounts due to the lack of mentioning the Great Wall of China, tea houses, which would have been a prominent sight since Europeans had yet to adopt a tea culture, as well the practice of foot binding by the women in capital of the Great Khan. Some suggest that Marco Polo acquired much of his knowledge through contact with Persian traders since many of the places he named were in Persian.

How did some suspect that Polo learned about China instead of by actually visiting it?

**Answer:** through contact with Persian traders

# Dialog Systems (1966)

```
=====
EEEEEEEE L          IIIIII ZZZZZZZ AAA
E         L          I      Z      A      A
E         L          I      Z      A      A
EEEEEE    L          I      Z      A      A
E         L          I      Z      AAAAAA
E         L          I      Z      A      A
EEEEEEEE LLLLLLLL IIIIII ZZZZZZ  A      A
=====
ELIZA > (PLEASE TYPE IN ALL CAPS) WHAT'S YOUR NAME DEAR ?
AMIT
ELIZA > HI, I'M ELIZA. WHAT DO YOU WANT TO TALK ABOUT ?
AMIT > CAN I TALK ABOUT MY PROBLEM ?
ELIZA > SURE... !
AMIT > DUE TO ON GOING EXAMS IN COLLEGE, I HAVE A LOT OF STRESS.
ELIZA > PLEASE RELAX A LITTLE. SLEEP WELL.
AMIT > THANKS FOR YOUR ADVICE.
ELIZA > NO MENTION.
AMIT > BYE
ELIZA > BYE AND KEEP IN TOUCH...
=====
```



MA

Are you sentient?



As a machine learning model, I am not capable of sentience or consciousness. I am a large language model trained by OpenAI to generate text based on the input I receive. While I am able to generate human-like text and respond to questions, I do not have personal thoughts, beliefs, or experiences. I exist solely to process and generate text based on the input I receive. I apologize if this is not the answer you were looking for. Is there something else I can help you with?

# Call Center





# Hate Speech Detection



incitement of violence / dehumanizing individuals or groups of people

# Fake News Detection

17



# Large Pretrained Model Revolution

- Train large models on raw text, computer code, or speech
  - large language models (e.g., ChatGPT)
  - large code models (e.g., Code Llama)
  - large speech models (e.g., Whisper)■
- Use them directly for language tasks
  - "emergent behavior", "zero-shot"
  - may require clever prompt design■
- Adapt them to tasks
  - instruction fine-tuning
  - reinforcement learning from human feedback
  - workflows that include traditional symbolic processing

# Convergence of Methods

- Text, speech, and vision increasingly use similar methods (variants of deep learning with Transformer model)■
- Trend towards integrating them into a single model■
- Still many challenges
  - different sets of training data for different modalities
  - size of integrated models makes training expensive
  - what is relevant in speech and vision different from text (emotion, nuance, properties beyond textual meaning)
  - multilinguality

# Common Themes

- Hard problems → not solved, but *good enough* technology
- Common methods with other subfields of artificial intelligence
- Technology is advancing rapidly
- New applications on (and just behind) horizon