WSDM 2023 Tutorial

Knowledge-Augmented Methods for Natural Language Processing

Chenguang Zhu[1], Yichong Xu[1], Xiang Ren[2], Bill Yuchen Lin[2], Meng Jiang[3], Wenhao Yu[3]

[1] Microsoft Cognitive Services Research [2] University of Southern California [3] University of Notre Dame

Presenters

Chenguang Zhu

Principal Research Manager

Microsoft Cognitive Services Research

Yichong Xu

Senior Researcher

Microsoft Cognitive Services Research

Xiang Ren

Assistant Professor

Dept. of Computer Science

University of Southern California

Bill Yuchen Lin

Ph.D. candidate

Dept. of Computer Science

University of Southern California

Meng Jiang

Assistant Professor

Dept. of Computer Science and Engineering

University of Notre Dame

Wenhao Yu

Ph.D. candidate

Dept. of Computer Science and Engineering

University of Notre Dame

Disclaimer: This tutorial is our own opinions

- Not Microsoft's, USC's or Univ. of Notre Dame's
- To access mentioned models + datasets, please refer to corresponding licensing information
- We're not promoting the use of any particular model and/or datasets
- There are slides / figures borrowed from respective papers
- This tutorial is by no means exhaustive: we've tried our best to include relevant materials

How to access tutorial materials

- Detailed information about our tutorial can be found at: https://www.wsdm-conference.org/2023/program/tutorials

- Talk slides are at: https://github.com/zcgzcgzcg1/WSDM2023_Knowledge_NLP_Tutorial/

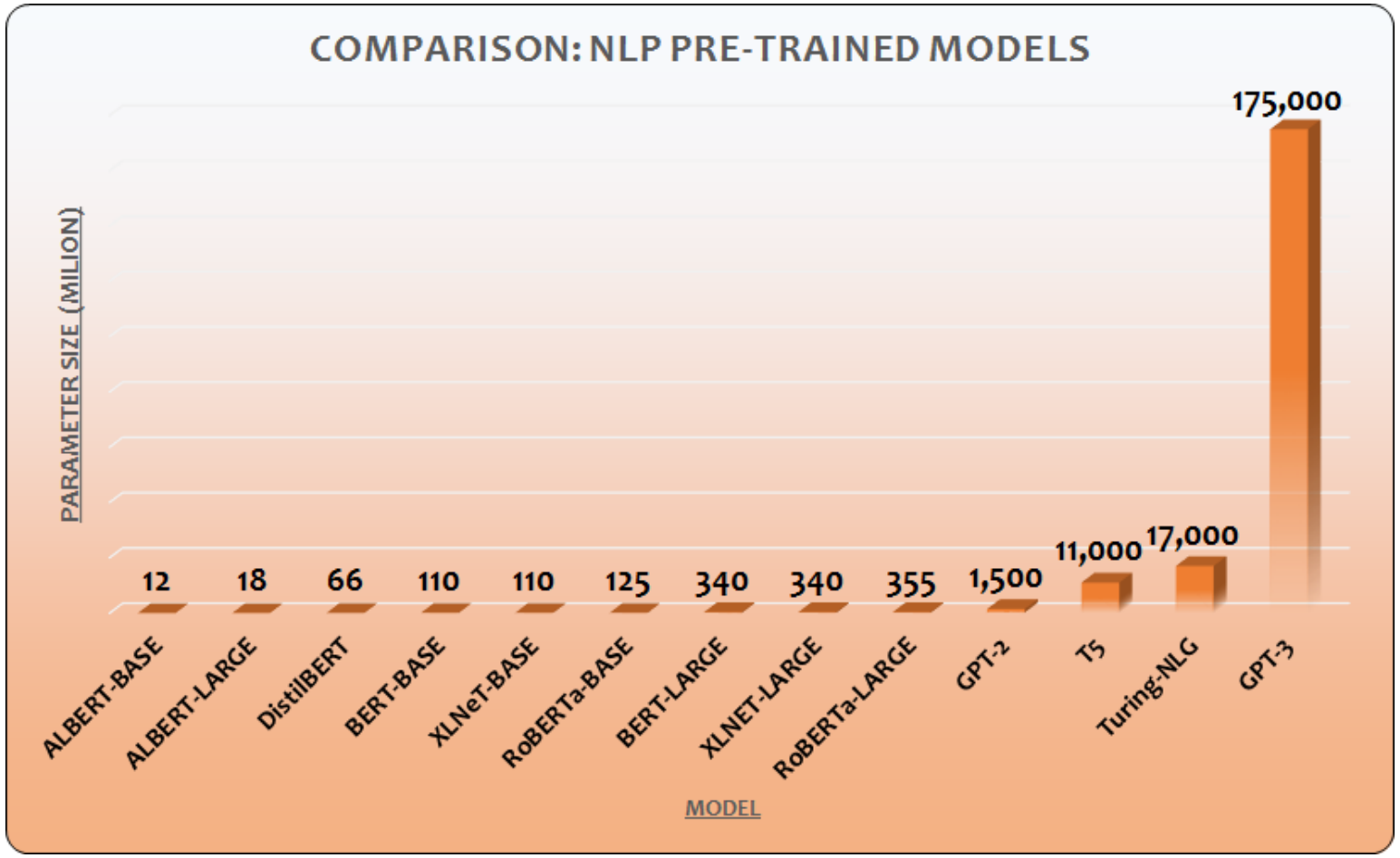
What is this tutorial about?

- **How to fuse knowledge and common sense into natural language processing**
- Knowledge in natural language understanding (NLU)
  - Natural language inference, sentence classification, sequence labeling, etc.
- Knowledge in natural language generation (NLG)
  - Text summarization, dialogue response generation, story generation, etc.
- Commonsense reasoning
  - Commonsense Q&A, commonsense generation

Schedule

Local time (GMT+8)	Content	Presenter
09:30-09:45	Motivation and Introduction of Knowledge in NLP	Chenguang Zhu
09:45-10:35	Knowledge in Natural Language Understanding	Yichong Xu
10:35-11:00	Knowledge in Natural Language Generation	Wenhao Yu / Meng Jiang
11:00-11:30	Coffee Break	
11:30-11:55	Knowledge in Natural Language Generation	Wenhao Yu / Meng Jiang
11:55-12:45	Commonsense Knowledge and Reasoning for NLP	Yuchen Lin / Xiang Ren
12:45-13:00	Summary and Future Direction	Meng Jiang / Xiang Ren

Where is NLP heading?



- Large, Huge, Gigantic Language models
- Training cost affordable only by few large companies
- Even fine-tuning is impossible for a majority of researchers and practitioners
- Does model size solve everything?
  - *Unfortunately, no*
- Then why are we doing it?

Diagram from https://medium.com/analytics-vidhya/openai-gpt-3-language-models-are-few-shot-learners-82531b3d3122

Integration of External Knowledge

Knowledge in NLP

- A language model (LM) learns **how to express**

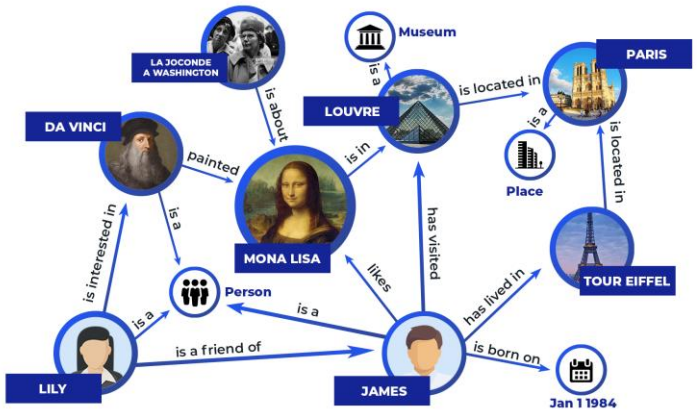
  I go school to to want. ✗

  I want to go to school. ✓

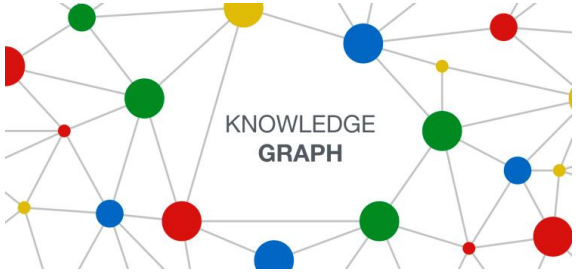
- Knowledge indicates **what to express**

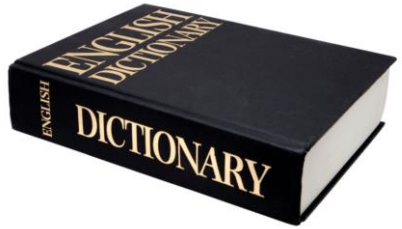
  Q: Where is the painting **Mona Lisa**?

  A: It is in **Louvre, Paris**.



Knowledge sources





Structured

- **Knowledge graph**: A meta-representation of knowledge, common sense, entities, relations
- **Dictionary**: explanation of words and phrases

Unstructured

- **Text data**: Knowledge from data without a predefined format, e.g., documents, emails
- **Large language models**, e.g., GPT-3

Knowledge is any external information absent from the input but helpful for generating the output

Integrate Knowledge into LM

- Step 1: **Ground** language into related knowledge
- Step 2: **Represent** knowledge
- Step 3: **Fuse** knowledge representation into language model

Integrate Knowledge into LM

- **Ground** language into related knowledge
  - String matching, NER, Entity linking, information retrieval
  - Identify concepts and relations in the knowledge source

The **pen** is on the **desk**.

Integrate Knowledge into LM

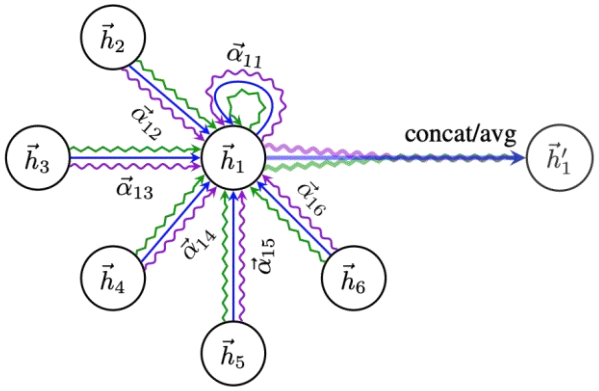
- **Represent** knowledge
  - Concept names

Desk

  - Description of concepts

Desk: A table, frame, or case, now usually with a flat top, for writers and readers. It often has a drawer or repository underneath.

  - Graph embeddings



Integrate Knowledge into LM

- **Fuse** knowledge representation into language model
  - Concatenate concept names/descriptions into input

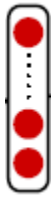
The pen is on the desk. [SEP] desk: a table, …

  - Append/add concept embeddings into input embeddings

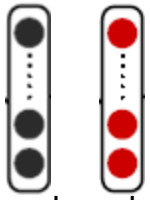
The pen is on the desk.

  - Attention

Graph embedding of pen



Graph embedding of desk



Graph embeddings

LM Transformer