Concluding Remarks:
Tutorial on Knowledge-Augmented Methods for Natural Language Processing

Meng Jiang (University of Notre Dame)

Microsoft | University of Notre Dame | USC University of Southern California | Ai2 Allen Institute for AI

Why do NLP Models need Knowledge?

- **The job of most NLP models (or most of machine learning models) is to bridge the gap between input and output.**
  - Sentiment analysis: Between a review sentence and a sentiment category (happy / unhappy)
  - Machine translation: Between a source-language sentence and a target-language
  - Fact verification: Between a factual statement and True / False (and explanation)
  - Question answering: Between a question and an answer
- **We seek perfection and never feel satisfied, so we want to use everything, if possible, to bridge the gap.**
  - Input-output pairs in training data (target task)
  - Input-output pairs in large-scale corpora (pre-training task, *not* the target task)
  - What else? Something that works with the input to better infer the output
- **That's Knowledge in the context of this tutorial.**

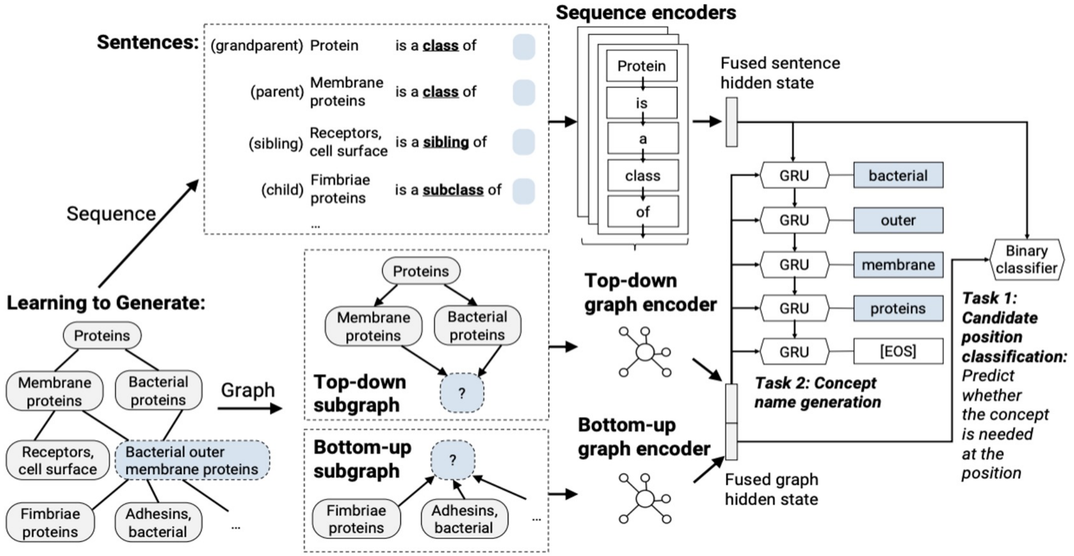
Where can the Knowledge come from?

- **Depends on Task domain:**
  - Commonsense: OMCS, ConceptNet, etc.
  - Encyclopedia: Wikipedia, WikiData, Wiktionary, etc.
  - General domain: Freebase, DBpedia, YAGO, etc.
  - Specific domains: UMLS, ArnetMiner, DBLP, etc.
  - Large language model (LLM): GPT-3, ChatGPT, etc.
- **Structured and unstructured knowledge: Knowledge graph, Text, etc.**
- **Yeah, they are basically extra data. How to make it work with Input?**
  - **Retrieval:** to find a piece of text related to the input
  - **LLM generation:** to create a piece of text
  - **Learning over knowledge graph:** to find a piece of structured data
  - **Pre-trained memory network:** to find a piece of embeddings
  - … These are the ***knowledge augmentation*** methods that we've introduced ☺

Future Directions 1

- **Use Knowledge to address Challenges besides/inside the gap:**
  1. Language models may hallucinate their output: Verified and/or edited by Knowledge
     - E.g., Factual correctness in abstractive summarization
  2. Event extraction models may limit to local context: Enhanced by structured global Knowledge
     - E.g., Document-level or even corpus-level information extraction
  3. Taxonomy construction models may limit to concept pool: Fusing info from corpus and taxo.
     - E.g., Generating concepts word by word on a taxonomy / knowledge graph

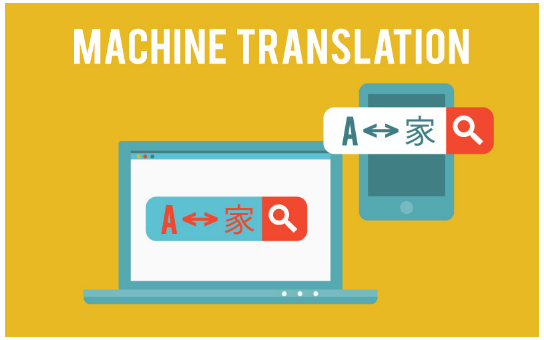
- *In abstractive summarization, 30% of generated summaries from state-of-the-art model contain unfaithful information. - Cao et al., ACL 2018.*
- *In dialogue system, 68% of generated responses from BART-large contain "hallucination" problems. -- Shuster et al., EMNLP 2021.*



Future Directions 1 (cont.)

- **Use Knowledge to address Challenges besides/inside the gap:**

4. Retrieval augmentation models may not be efficient as they need a great number of passages
    - E.g., Knowledge graph may filter out less related passages
5. Text generation models may not be able to create diverse outputs: Exploration on Knowledge
    - E.g., Leveraging one-to-many-words dictionaries, one-to-many-neighbors knowledge graph
6. NLP models may not be able to explain their decisions: Interpreting with Knowledge
    - E.g., Generating explanations along with a decision



Machine Translation
(Chinese <-> English)

尽管 → although
尽管 → despite

English-to-Chinese Dictionary
(one Chinese word has two corresponding English words)

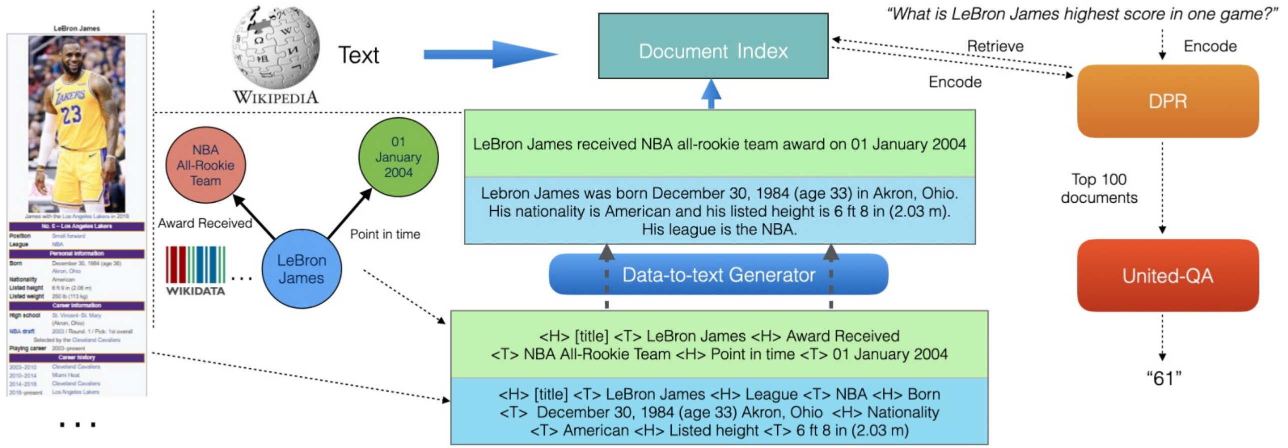
Story Generation
(Title, keywords -> Story)

One start → Happy end
One start → Sad end

Knowledge Graph
(a concept in starts can be connected to different relevant concepts in endings)

Future Directions 2

- **Improving the Methods of Knowledge Augmentation:**
  - Increasing knowledge coverage:
    - From Wikipedia to Web-scale corpus: Indexing and searching in large text corpus is computationally expensive.
    - From retrieving Wikipedia to Google search: Noisy information may be included.
  - Integrating heterogeneous knowledge:
    - Text, dictionaries, relational databases, fact triplets, knowledge graphs, taxonomies, ontologies, image / audio / video data, etc.
    - Different algorithms have been designed for different types of data. Unified approach?



Find us! Join us!

Survey paper:

- Yu et al. *A survey of knowledge-enhanced text generation.* ACM Computing Surveys, 2022.
- GitHub: https://github.com/wyu97/KENLG-Reading

Tutorials:

- Knowledge-enriched natural language generation. EMNLP 2021. https://kenlg-tutorial.github.io/
- Knowledge-Augmented Methods for Natural Language Processing. ACL 2022. https://github.com/zcgzcgzcg1/ACL2022_KnowledgeNLP_Tutorial
- Knowledge-Augmented Methods for Natural Language Processing. WSDM 2023.

Workshop:

- **The first workshop on Knowledge Augmented Methods for NLP (KnowledgeNLP-AAAI), 2023. https://knowledge-nlp.github.io/aaai2023/**