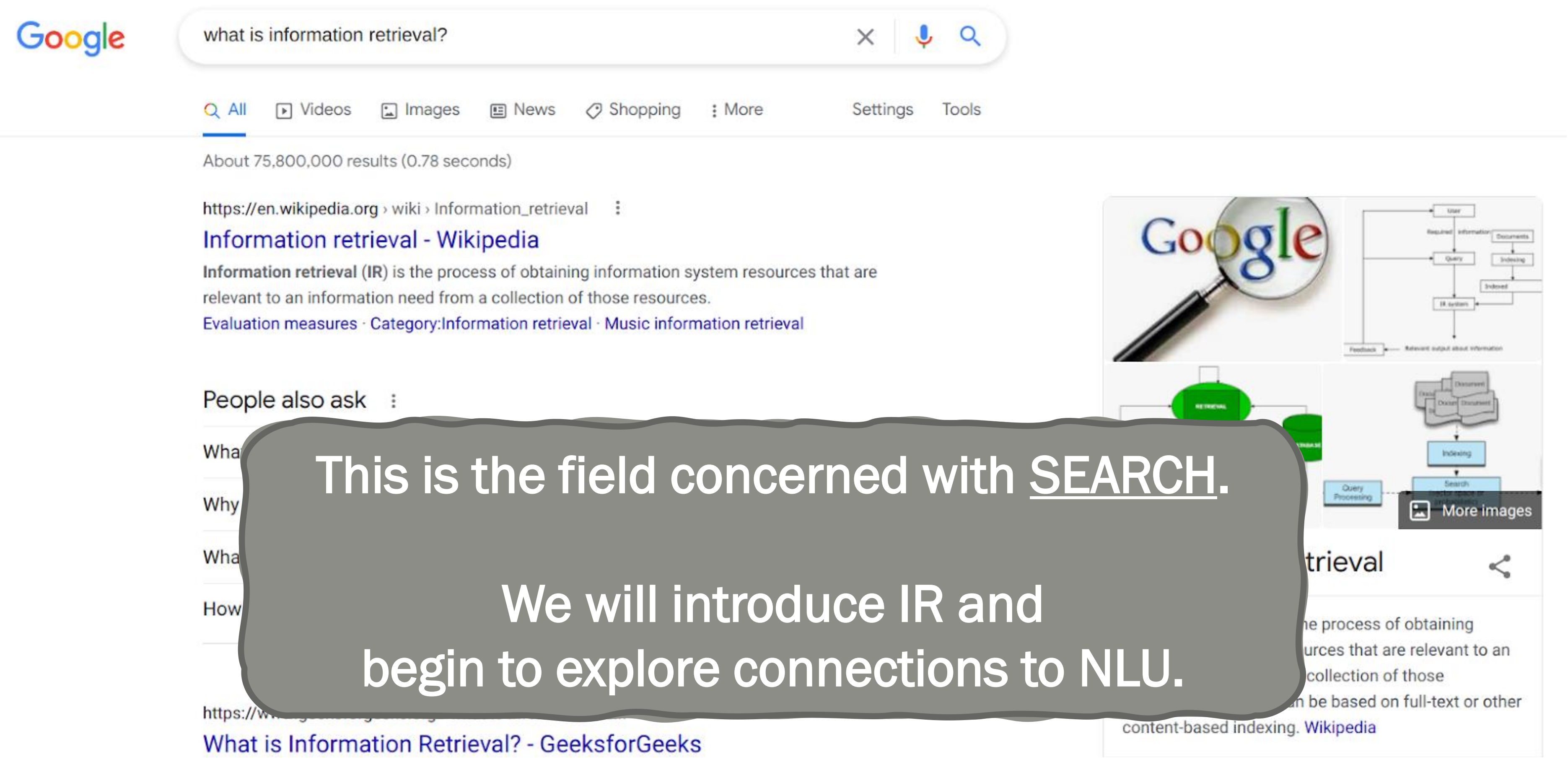
NLU & IR: OVERVIEW

Omar Khattab

CS224U: Natural Language Understanding

Spring 2021

What is information retrieval?



About 75,800,000 results (0.78 seconds)

https://en.wikipedia.org › wiki › Information_retrieval

Information retrieval - Wikipedia

Information retrieval (**IR**) is the process of obtaining information system resources that are relevant to an information need from a collection of those resources.

Evaluation measures · Category:Information retrieval · Music information retrieval

People also ask

Wha

Why

Wha

How

https://w

What is Information Retrieval? - GeeksforGeeks

trieval

he process of obtaining
urces that are relevant to an
collection of those
n be based on full-text or other
content-based indexing. Wikipedia

This is the field concerned with SEARCH.

We will introduce IR and
begin to explore connections to NLU.

What is information retrieval?

Finding material that fulfills an **information need** from within a **large collection** of **unstructured documents**.

Simplified definition from IIR Book
(Manning, Raghavan, and Schütze)

What is information retrieval?

Finding material that fulfills an **information need** from within a **large collection** of **unstructured documents**.

SEARCH is essential to IR!

But search over *highly-structured* data (e.g., database records) is NOT typically considered an IR problem.

Simplified definition from IIR Book (Manning, Raghavan, and Schütze)

Relevance — and the "Information Need"

- The goal of a search system is to satisfy an **information need**.
  - Material we retrieve is **relevant** only if it advances this goal.

- In many (most) tasks, the user will express a **query**.
  - But queries can be ambiguous, incomplete, or inaccurate.
  - We must rely on our knowledge of the <u>task</u> and the <u>user</u>.

Typical information needs vary by task

- Beyond Web pages and files, popular types of collections include digital libraries, media items, products, online conversations, etc.

Expression of Information Need	Potential Query	Potential Collection
Find related literature	The full text of the BERT paper	ACL anthology; arXiv CL

Typical information needs vary by task

- Beyond Web pages and files, popular types of collections include digital libraries, media items, products, online conversations, etc.

Expression of Information Need	Potential Query	Potential Collection
Find related literature	The full text of the BERT paper	ACL anthology; arXiv CL
Recommend me a TV show to watch	[no explicit query!]	Netflix shows

Typical information needs vary by task

- Beyond Web pages and files, popular types of collections include digital libraries, media items, products, online conversations, etc.

Expression of Information Need	Potential Query	Potential Collection
Find related literature	The full text of the BERT paper	ACL anthology; arXiv CL
Recommend me a TV show to watch	[no explicit query!]	Netflix shows
Find every relevant patent	Boolean query with technical terms	U.S. Patents

Typical information needs vary by task

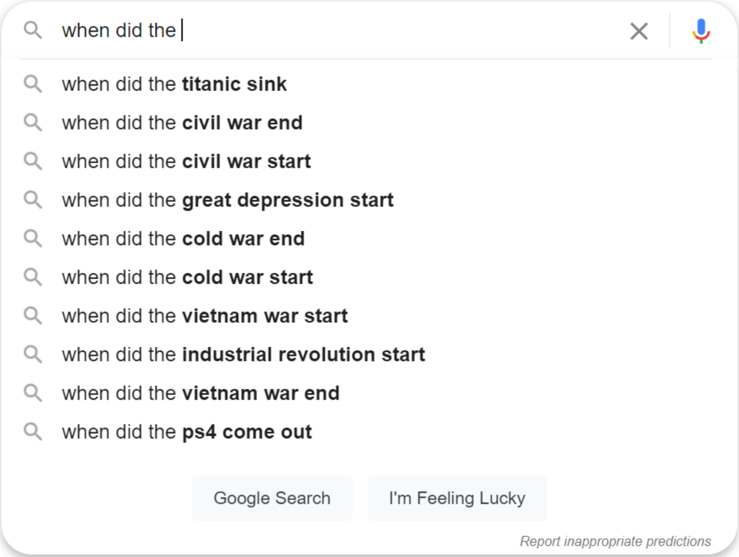
- Beyond Web pages and files, popular types of collections include digital libraries, media items, products, online conversations, etc.

Expression of Information Need	Potential Query	Potential Collection
Find related literature	The full text of the BERT paper	ACL anthology; arXiv CL
Recommend me a TV show to watch	[no explicit query!]	Netflix shows
Find every relevant patent	Boolean query with technical terms	U.S. Patents
Buy a new laptop	Short conversation: system asks questions to ascertain your criteria	E-commerce platforms

Typical information needs vary by task

- Each search task poses unique challenges!
  - Many of them lack key features that make Web search work.

- Unlike, say, Slack search, Web search can often rely on lots of:
  - Popular “head” queries
  - Redundant documents on common topics
  - Explicit (hyper)links between documents



when did the

- when did the **titanic sink**
- when did the **civil war end**
- when did the **civil war start**
- when did the **great depression start**
- when did the **cold war end**
- when did the **cold war start**
- when did the **vietnam war start**
- when did the **industrial revolution start**
- when did the **vietnam war end**
- when did the **ps4 come out**

Where does NLU fit in **IR**?

- Queries and documents are often expressed in natural language.
- Due to **vocabulary mismatch**, lexical matching doesn't suffice!

Jimmy Lin
@lintool

> "IR makes NLP [more!] useful.
> NLP makes IR interesting."

what compounds protect the digestive system against viruses

In the **stomach**, gastric acid and proteases serve as powerful **chemical defenses** against ingested **pathogens**.

Where does IR fit into **NLU**?

- Advanced models often have information needs too!

- Retrieval in NLU can contribute to:
  - **Creating new challenging NLU tasks**
  - Improving model efficiency and quality for existing NLU tasks
  - Evaluating NLU systems whenever the output domain is large

Retrieval supports “open-domain” NLU tasks

- We’ve briefly introduced SQuAD before…

Context: Chemical barriers also protect against infection. The skin and respiratory tract secrete antimicrobial peptides such as the β-defensins. […] In the stomach, gastric acid and proteases serve as powerful chemical defenses against ingested pathogens.

Question: What compounds in the stomach protect against ingested pathogens?

Answer: gastric acid and proteases

Standard Question Answering (e.g., SQuAD)

Rajpurkar, Pranav, et al. “SQuAD: 100,000+ questions for machine comprehension of text." EMNLP’16

From standard QA to open-domain QA

- Drop the passage hint!

Context: All of [English] Wikipedia, with no special hints about the answer

Question: What compounds in the stomach protect against ingested pathogens?

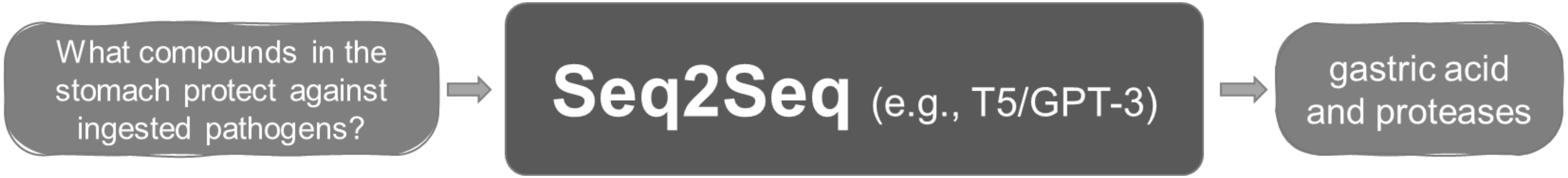
Answer: gastric acid and proteases

Open-Domain Question Answering
(e.g., this “Open-SQuAD”)

Chen, Danqi, et al. "Reading Wikipedia to answer open-domain questions." ACL’17.

Open-Domain QA: **Closed-Book Approaches**

- Feed the question to a monolithic black-box generative model!
  - Knowledge is stored *implicitly* in the model parameters
  - Often as a byproduct of language-model pretraining
  - Need more “knowledge”? Train a larger model on more data!

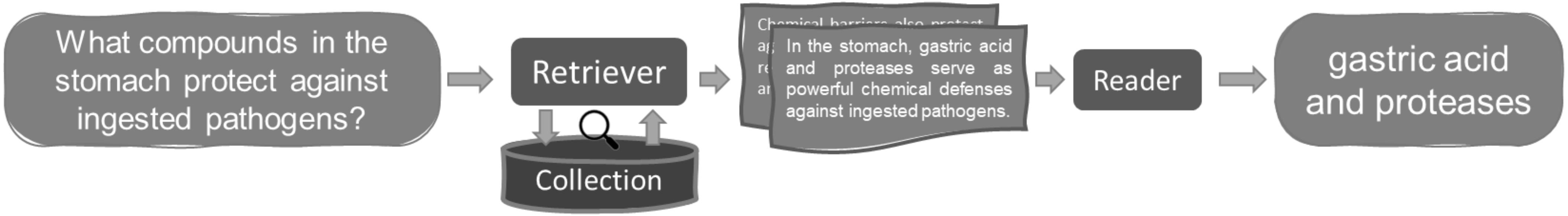

What compounds in the stomach protect against ingested pathogens?
Seq2Seq (e.g., T5/GPT-3)
gastric acid and proteases

Roberts, Adam, Colin Raffel, and Noam Shazeer. "How Much Knowledge Can You Pack Into the Parameters of a Language Model?." *EMNLP*’20.

Open-Domain QA: Open-Book Approaches

- Feed the question to a modular **retrieve-and-read** architecture
  - Knowledge is stored *explicitly* in the collection
  - We decouple **reasoning** and **knowledge**

The reader has an ***information need.*** *The retriever's task is to satisfy it efficiently and accurately.*



What compounds in the stomach protect against ingested pathogens?

Retriever

Collection

In the stomach, gastric acid and proteases serve as powerful chemical defenses against ingested pathogens.

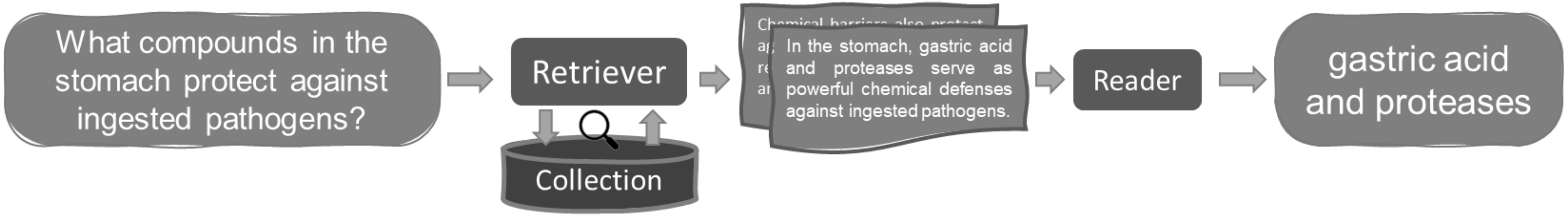
Reader

gastric acid and proteases

Open-Domain QA: **Open-Book Approaches**

- ✓ Models can be much smaller
- ✓ Knowledge can be updated (or customized) without retraining
- ✓ Model predictions might become more explainable

- ✗ We now need to worry about the interactions between a **retriever** and **reader**



What compounds in the stomach protect against ingested pathogens?

Retriever

Collection

Chemical barriers also protect

In the stomach, gastric acid and proteases serve as powerful chemical defenses against ingested pathogens.

Reader

gastric acid and proteases

A few retrieval-based NLP tasks

Task Name	Input	Output
Open-Domain QA	Question	Answer

A few retrieval-based NLP tasks

Task Name	Input	Output
Open-Domain QA	Question	Answer
Fact Checking	Claim	Binary Label & Justification

A few retrieval-based NLP tasks

Task Name	Input	Output
Open-Domain QA	Question	Answer
Fact Checking	Claim	Binary Label & Justification
Query-Focused Summarization	Topic	Summary
Informative Dialogue	Conversation Turns	Response

A few retrieval-based NLP tasks

Task Name	Input	Output
Open-Domain QA	Question	Answer
Fact Checking	Claim	Binary Label & Justification
Query-Focused Summarization	Topic	Summary
Informative Dialogue	Conversation Turns	Response
Entity Linking	Utterance	Mapping from spans to entities in a knowledge base

Retrieval-based NLP tasks

- KILT is a recent benchmark that brings together several datasets for **knowledge-intensive** language tasks.

KILT

- These are tasks that explicitly have a knowledge component.

Open Question: Can retrieval dramatically improve performance for standard NLU tasks too?

Accurate knowledge matters for most (all?) tasks!

“Bring your own book!”

Next...

- The remainder is structured as small crash courses into:
  - Classical Information Retrieval
  - Neural Information Retrieval
  - Open-Domain Question Answering

References

Manning, Christopher, Prabhakar Raghavan and Schutze, H. “Introduction to Information Retrieval.” (2008).

Manning, Christopher, and Pandu Nayak (2019). CS276 Information Retrieval and Web Search: Inverted Indices [Class handout]. Retrieved from http://web.stanford.edu/class/cs276/19handouts/lecture2-intro-boolean-6per.pdf

Elsayed, Tamer. CMPT621 Information Retrieval: Introduction to IR [Class handout]. Retrieved from https://www.dropbox.com/sh/8oeivk53ymsj3oy/AABt9M6ve5qYCZShkS7a5BkZa/Lecture%20Slides?dl=0&preview=1-CMPT621-S21-Session1-Intro+to+IR.pdf

Rajpurkar, Pranav, et al. “SQuAD: 100,000+ questions for machine comprehension of text." EMNLP’16.

Chen, Danqi, et al. "Reading Wikipedia to answer open-domain questions." ACL’17.

Roberts, Adam, Colin Raffel, and Noam Shazeer. "How Much Knowledge Can You Pack Into the Parameters of a Language Model?." EMNLP’20.

Petroni, Fabio, et al. "KILT: a benchmark for knowledge intensive language tasks." NAACL’21.