Rethinking Search: Making Domain Experts out of Dilettantes

Donald Metzler
Google Research
metzler@google.com

Yi Tay
Google Research
yitay@google.com

Dara Bahri
Google Research
dbahri@google.com

Marc Najork
Google Research
najork@google.com

Jungwoo Lim, Seonmin Koo

Index

1. Motivation

2. Introduction

3. Model-Based Information Retrieval

4. Properties of Envisioned Model

5. Conclusions

Motivation

Motivation

- Motivation From Pre-trained LMs
  - Traditional IR 이나 QA system과 달리 pre-trained LMs은 prose 을 directly 생성할 수 있음

  - Traditional IR 엔진 구성
    - Corpus의 각 document에 대한 효율적인 쿼리가 가능하도록 index 구축
    - 주어진 쿼리에 대한 후보(candidate)들 검색
    - 각 후보에 대한 relevance score 계산

Ex. 고려대학교 자연어처리 연구실이 어디야?

- Traditional IR

→ 답변의 reference 제시

- Domain Experts

“ 애기능 생활관에 있는데 좋다고 하더라 “

→ 실제적이고 구체적인 답변

Motivation

- Motivation From Pre-trained LMs
  - 그러나 아래와 같은 dilettantes 문제
    - they do not have a true understanding of the world
    - they are prone to hallucinating
    - they are incapable of justifying their utterances by referring to supporting documents in the corpus they were trained over

✓ Sequences of terms과 documents 사이의 간격을 줄이지 못한다는 것을 근본적인 문제로 제시

Motivation

- Motivation From Pre-trained LMs

  - 최근 동향을 고려할 때 앞으로는 기존 연구를 통합하여 발전시켰을 때의 가능성에 대해 생각해야 함

  - 만약 Classical IR과 pre-trained LMs 결합하여 연구하면 domain expert quality responses에 가까워지지 않을까? 라는 motivation임

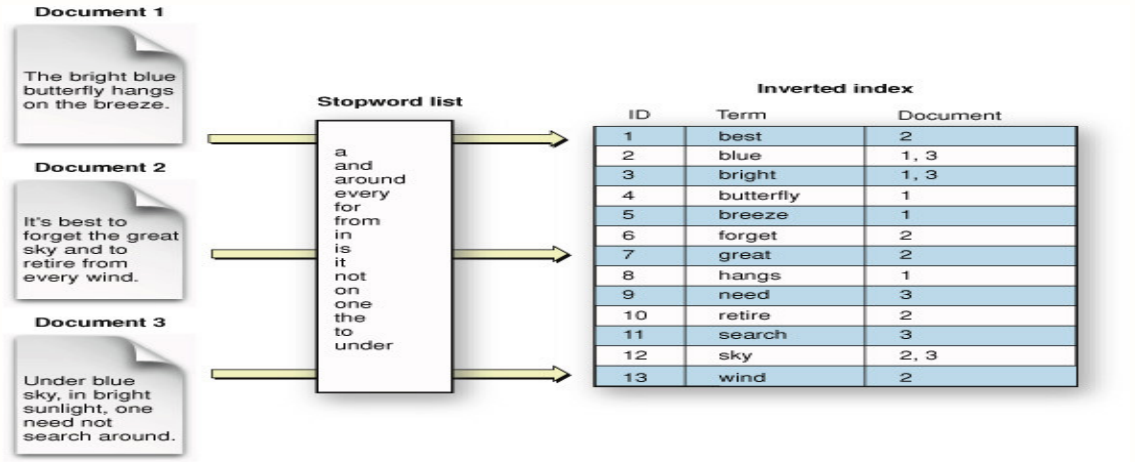
Introduction

Introduction

- **Domain Experts**
  - Actual domain expert는 주어진 topic을 “true understanding”
  - 본 논문에서 정의하는 domain experts의 기준
    - **주어진 domain에서 system이 human expert와 같은 quality의 결과를 생성할 수 있는 것**
    - 실제로 “understanding”하고 생성 했는지 유무는 기준 아님
  - Domain experts 를 달성하기 위한 방법으로 classical IR과 pre-trained LMs 결합하는 방법 탐구

Introduction

- Previous Research
  - Inverted Index
    - 정의: 키워드를 통해 문서를 찾아내는 방식



  - 수십 년 동안 대부분의 modern search engine의 핵심 역할
    - Word를 uninterpreted tokens로 취급, semantic 파악 안함
    - 구체적으로, Morphology, term similarity, grammar 인식 못함

https://ciir.cs.umass.edu/downloads/SEIRiP.pdf

Introduction

- Previous Research

  - Dense vector-based index
    - Inverted indexing 의 vocabulary mismatch problem 해결
    - Recall 개선하는데 도움되는 semantically- rich document representations 인코딩

  - IR researchers는 language understanding advances를 잘 활용하고 있음
    - Representation learning은 retrieval 목적으로 사용
    - pre-trained LMs는 scoring에 활용

https://arxiv.org/pdf/2004.13969.pdf
https://arxiv.org/pdf/2004.04906.pdf
https://arxiv.org/pdf/2004.12832.pdf
https://arxiv.org/pdf/2010.01195.pdf

https://aclanthology.org/P19-1612.pdf
https://arxiv.org/pdf/2010.11386.pdf
https://arxiv.org/pdf/2007.00808.pdf

Introduction

- Critical Look
  - 모든 발전에도 불구하고 최신 IR 시스템과 classical IR 시스템은 근본적으로 다르지 않음
  - Index-retrieve-then-rank framework 는 거의 변하지 않음

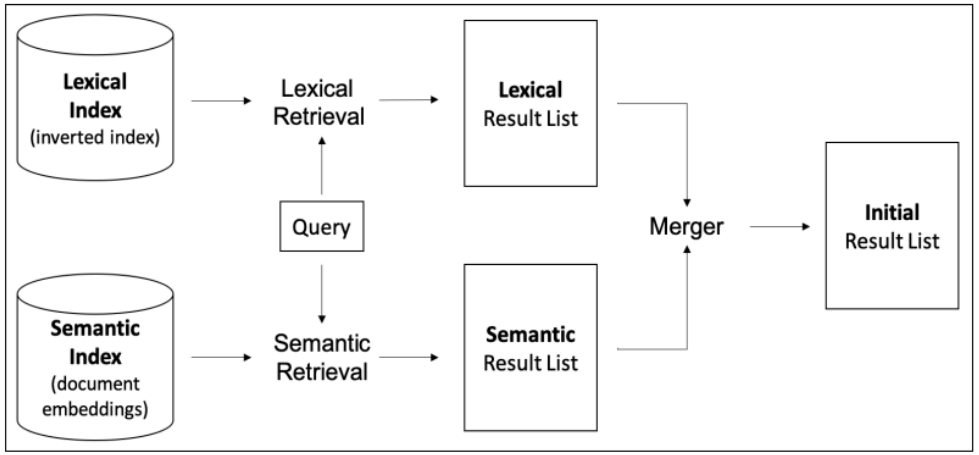
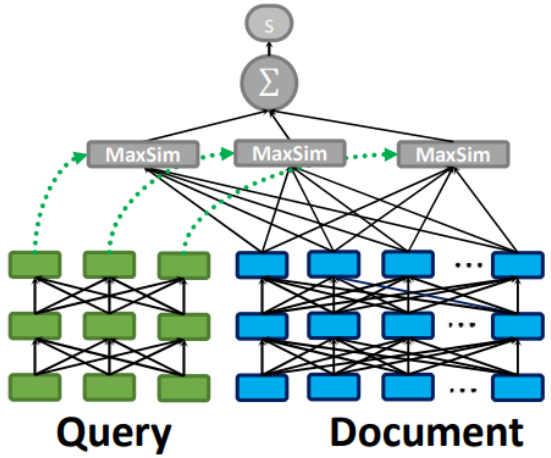


Figure 1: The hybrid retrieval approach.



(d) Late Interaction
(i.e., the proposed ColBERT)

Introduction

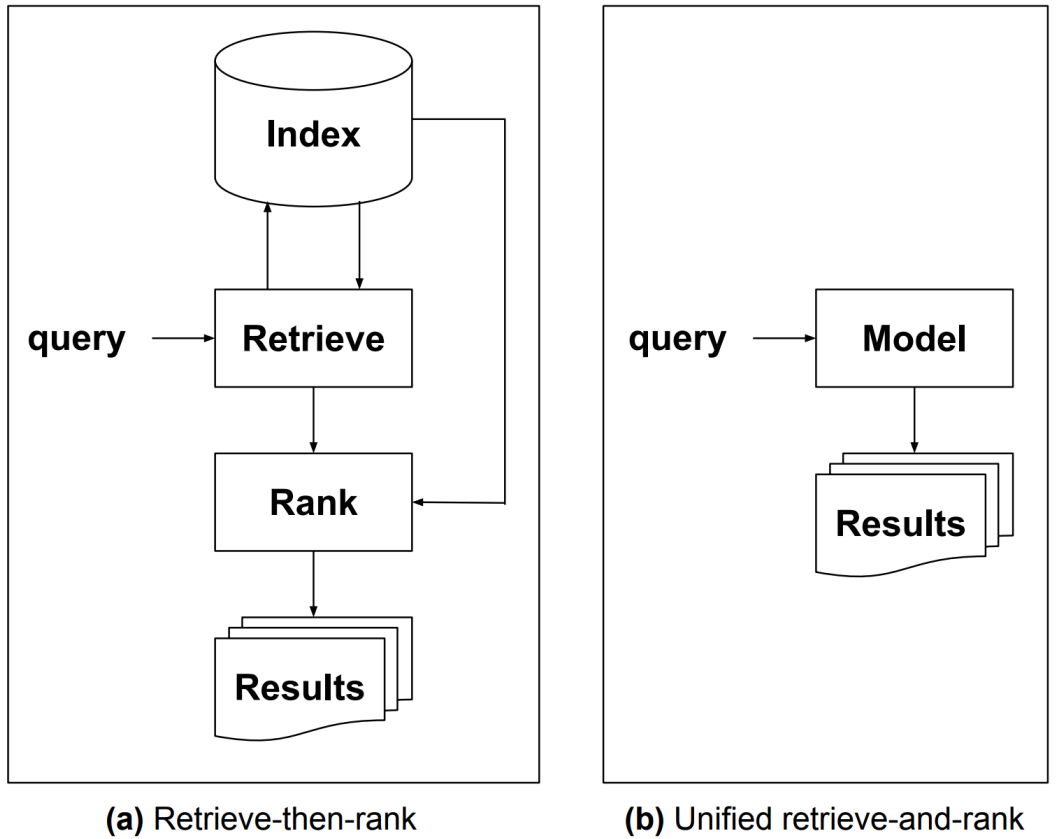
- Proposal
  - Document retrieval systems의 elements와 pre-trained 언어 모델을 결합한 retrieval 시스템 개발 case 제시
  - 주어진 Corpus에 대한 모든 knowledg를 인코딩하여 index의 필요성을 제거한 IR 시스템 구상

  → Model-Based information Retrieval에 대한 필요성

Model-Based Information Retrieval

Model-Based Information Retrieval

- Envisioned Model



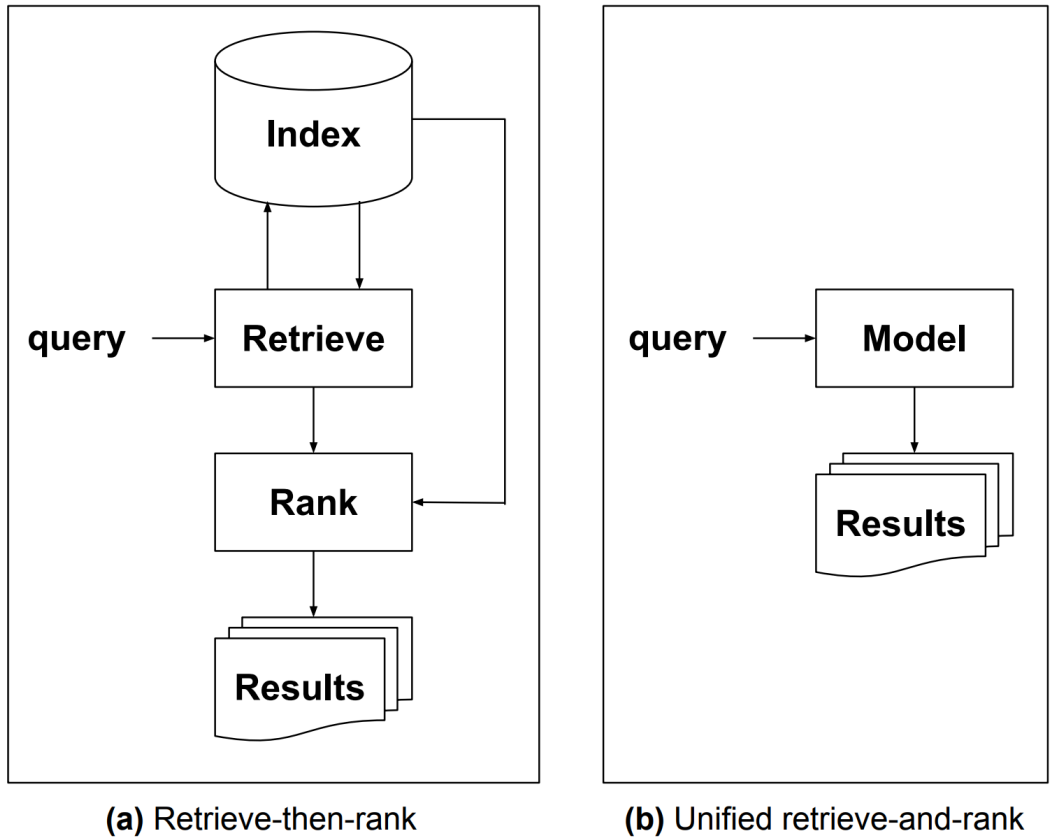
(a) Retrieve-then-rank **(b)** Unified retrieve-and-rank

- is meant to replace the long-lived "retrieve-then-rank" paradigm by collapsing the indexing, retrieval, and ranking components of traditional IR systems into a single consolidated model

- consolidated model replaces the indexing, retrieval, and ranking components. In essence, it is referred to as model-based because there is nothing but a model

Model-Based Information Retrieval

- Beyond Language Models



(a) Retrieve-then-rank

(b) Unified retrieve-and-rank

- 언어모델은 단순히 task를 위하여 sequence를 입력 받고 sequence를 출력하는 모델이기 때문에 관련되어 있는document와 sequence 내의
- 토큰들간의 관계성을 무시하게 됨

- 이를 위하여 document id같은 high-level entities 들을 input으로 같이 입력해주면서 이러한 관계성을 학습 시켜야할 것이라고 주장함

Properties of Envisioned Model

Multi-Task Learning

- Envisioned Model by Multi-task Learning

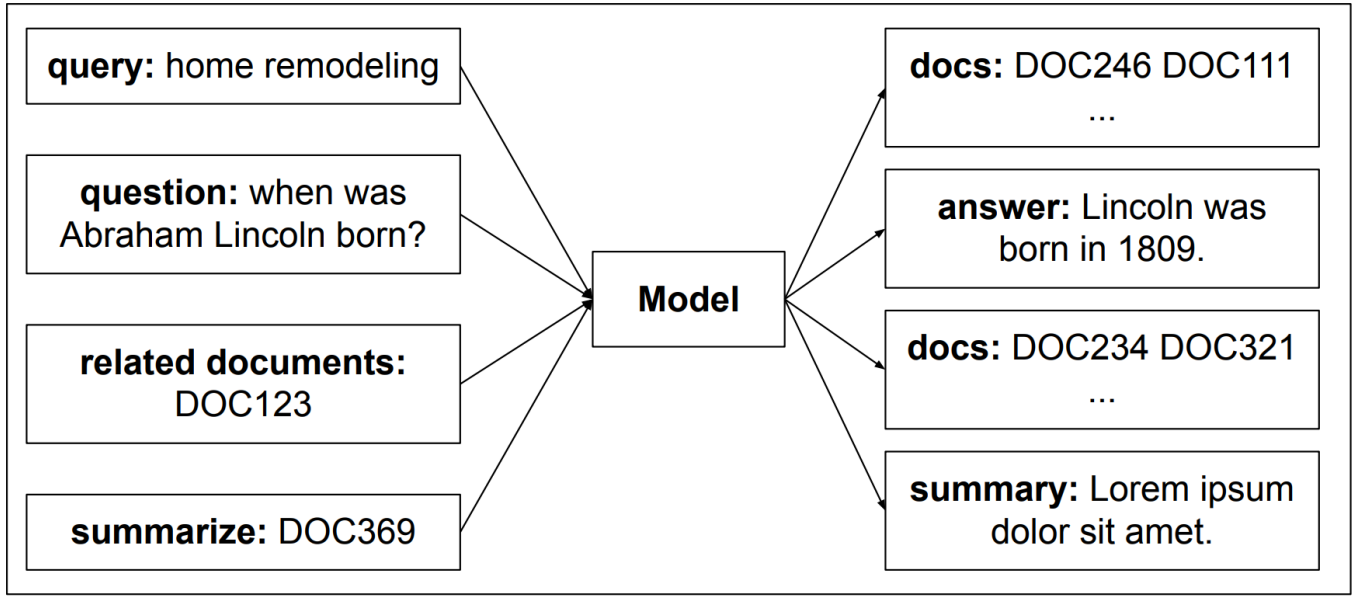


Figure 2: Example of how a single consolidated model can be leveraged to solve a wide range of IR tasks. This example shows a model that handles document retrieval, question answering, related document retrieval, and document summarization tasks.

Zero- and Few-shot Learning

- Envisioned Model by Multi-task Learning

Ad Hoc Retrieval (zero-shot)

- **Input:** $query$
- **Output:** $reldoc_1, \ldots, reldoc_n$

Query Understanding (few-shot)

- **Input:** $(query_1, intent_1), \ldots, (query_n, intent_n)\ query$
- **Output:** $intent$

Pseudo-relevance feedback (few-shot)

- **Input:** $(query_1, doc_1), \ldots, (query_n, doc_n)\ query$
- **Output:** $reldoc_1, \ldots, reldoc_n$

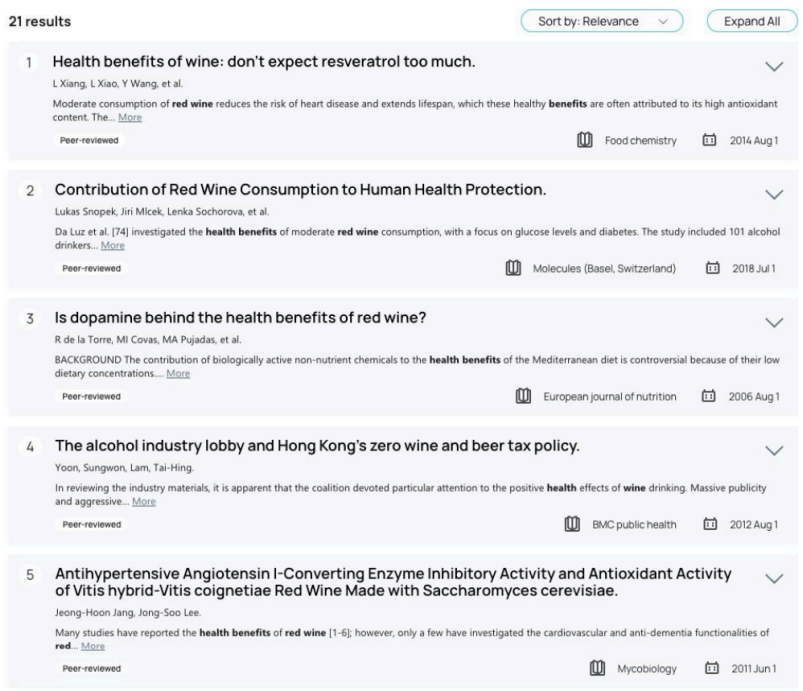
Document Understanding (few-shot)

- **Input:** $(doc_1, label_1), \ldots, (doc_n, label_n)\ doc$
- **Output:** $label$

- having a **consolidated multi-task model** that understands the connections between sequences of terms and document identifiers opens up a wide range of straightforward and powerful use cases.

Envisioned Domain Expert

- Comparison on Search Engine



What are the health benefits and risks of red wine?
Well red wine definitely has health benefits, like promoting heart health, anti-bacterial properties, lowering your risk of certain cancers and much more. On the other hand it may stain your teeth and cause the more than occasional hang over.

What are the health benefits and risks of red wine?
According to WebMD, red wine's benefits include promoting heart health, anti-bacterial properties, and lowering your risk of certain cancers [webmd.com]. On the other hand, the Mayo Clinic reports that red wine may stain your teeth and cause the occasional hang over [mayoclinic.org].

Figure 3: Example domain-specific search engine (left), pre-trained language model (middle), and envisioned domain expert (right) responses for the query "What are the health benefits and risks of red wine?".

Response Generation

- **Desired aspects to have in the response**
  - Authoritative
    : Responses should generate content by pulling from highly authoritative sources
  - Transparent
    : Whenever possible, the provenance of the information being presented to the user should be made available to them.
  - Unbiased
    : Pre-trained LMs are trained to maximize their predictive power on their training data, and thus they may reflect societal biases in that data

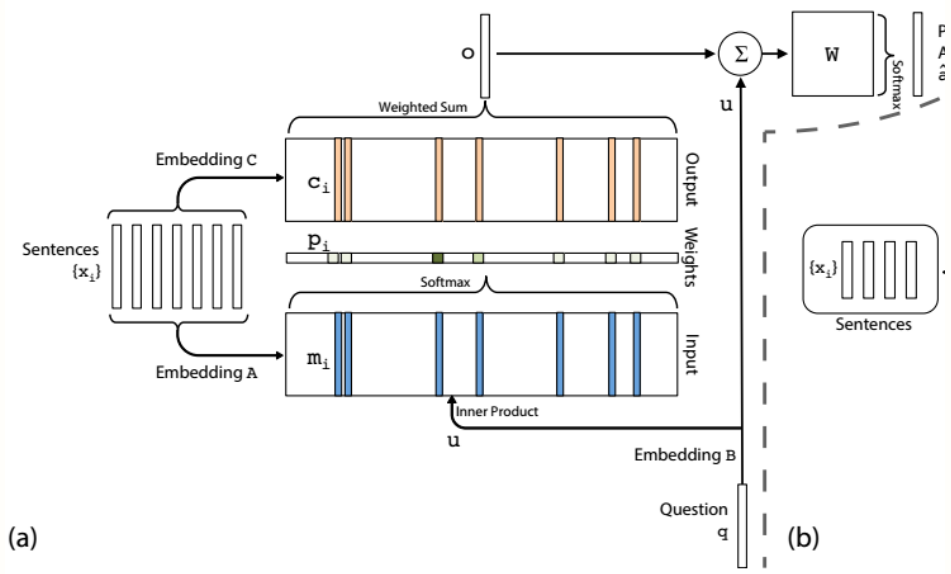
Response Generation

- **Desired aspects to have in the response**
  - Diverse perspective
    : Generated responses should represent a range of diverse perspectives but should not be polarizing

  - Accessible
    : Written in terms that are understandable to the user

Reasoning Capabilities

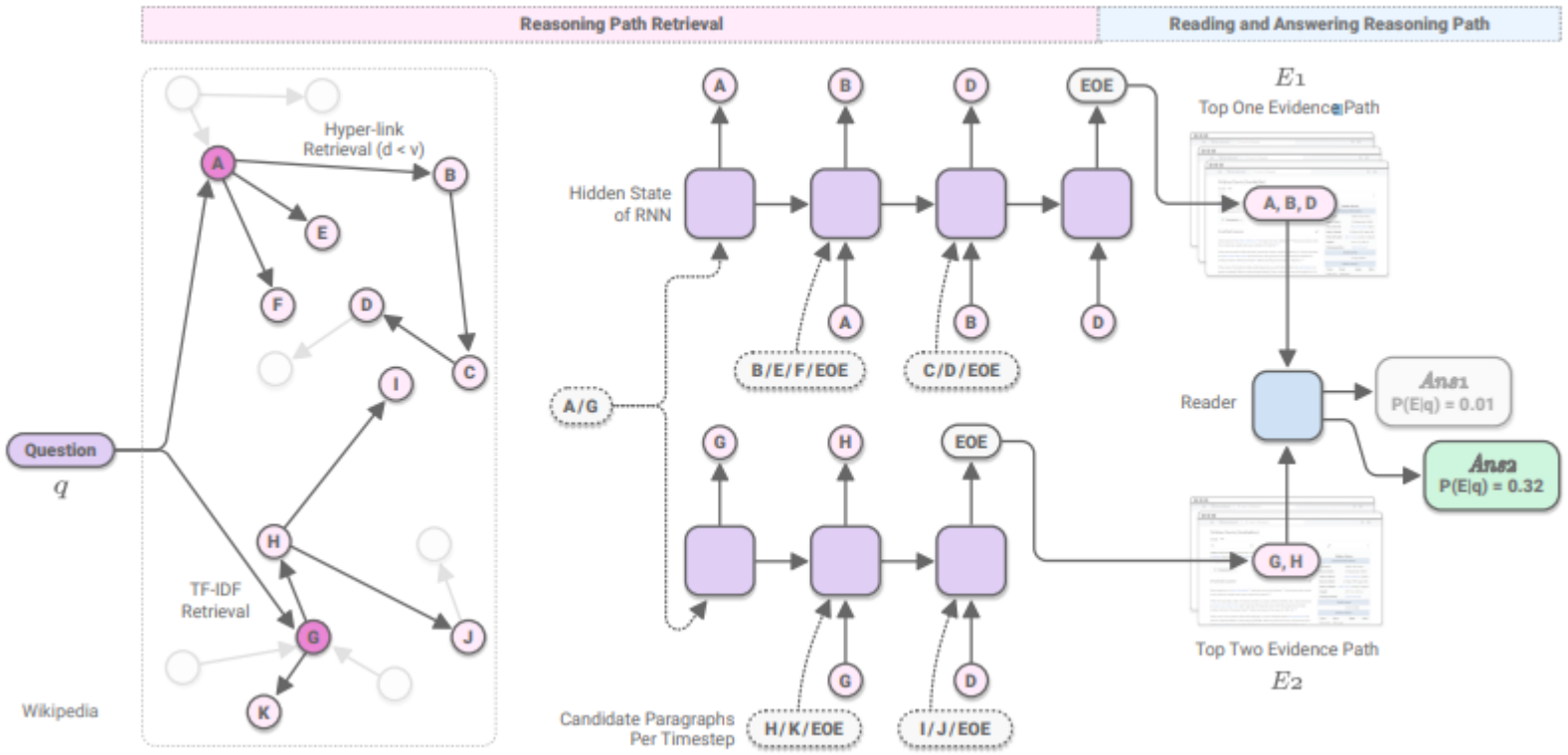
- Desired aspects in modularity
  - Memory-like Inductive bias: Memory Network… etc.



https://arxiv.org/pdf/1910.02610.pdf
https://www.tutorialexample.com/understand-end-to-end-memory-networks-part-1-a-simple-tutorial-for-nlp-beginners/

Reasoning Capabilities

- Desired aspects in modularity
  - Relational Inductive bias: Retrieving Reasoning Paths



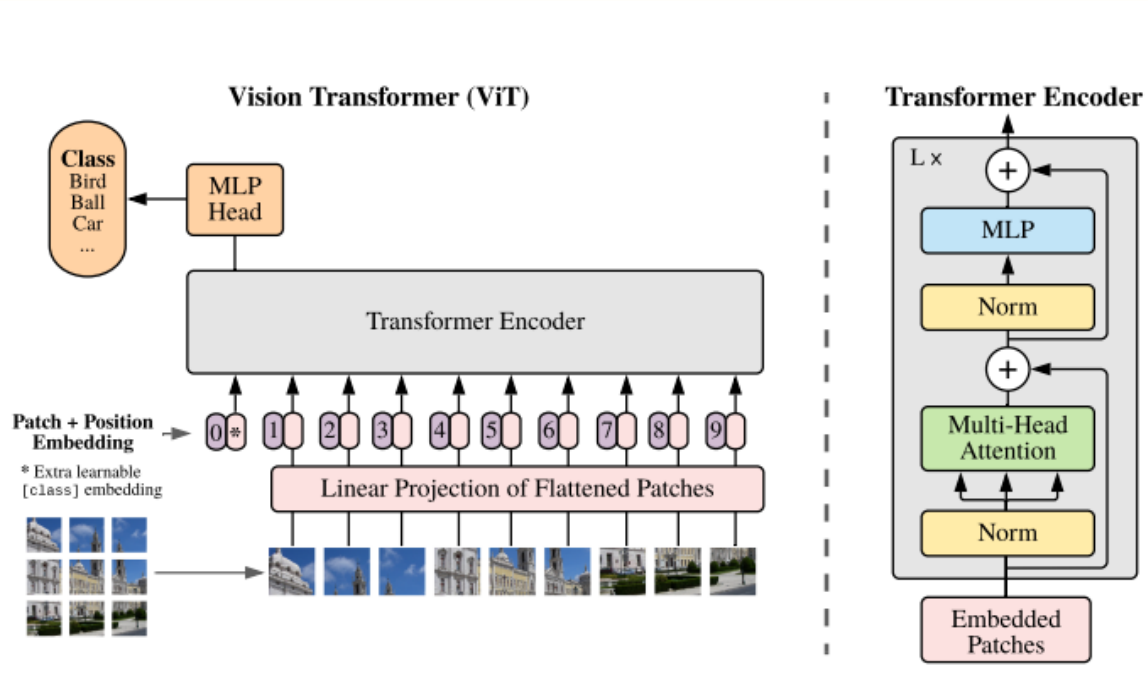
https://arxiv.org/pdf/1911.10470.pdf

Reasoning Capabilities

- Desired aspects in modularity
  - Arithmetic Reasoning
    e.g. 36,500 USD to pounds
  - Logical Reasoning
    e.g. All men are mortal. Socrates is a man. Therefore, Socrates is mortal.
  - Temporal Reasoning
    e.g. 2am PST to GMT-2
  - Geographical Reasoning
    e.g. how far is California from New York City

Combining Modalities

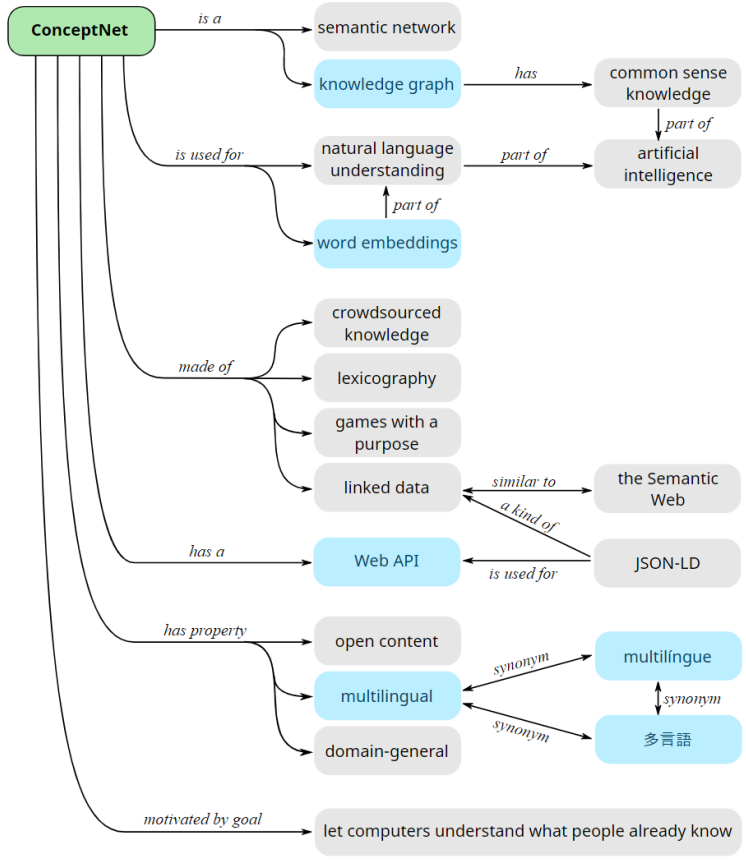
- Expecting Modalities
  - Metadata
  - Media content (images, video, and audio)



https://arxiv.org/pdf/2010.11929.pdf

Document and Corpus Structure

- Expecting Structures
  - Open corpus (web)
    - ConceptNet originated from the crowdsourcing project Open Mind Common Sense, which was launched in 1999 at the MIT Media Lab. It has since grown to include knowledge from other crowdsourced resources, expert-created resources, and games with a purpose.
  - Graph structure



Multiple Languages

- Expecting Properties
  - Cross-lingual IR
  - Balancing proportions between the training data from diverse languages

https://arxiv.org/pdf/1906.01502.pdf

Scale

- **Expecting Difficulties**
  - Model Capacity
  - Size/Length of Documents
  - Computation

Learning

- Desired way of Learning
  - Incremental Learning
    : incremental learning is a method of machine learning in which input data is continuously used to extend the existing model's knowledge i.e. to further train the model.

  - Continual Learning: studies the problem of learning from an infinite stream of data, with the goal of gradually extending acquired knowledge and using it for future learning

    - Online Learning: without demanding offline training of large batches or separate tasks introduces fast acquisition of new information

https://arxiv.org/abs/1909.08383
https://arxiv.org/pdf/1802.07569.pdf

Other Important Aspects

- Model Interpretability
  - it is well-known that modern deep neural networks suffer from interpretability issues
- Model Controllability
  - model designer should know how to control the behavior of the trained model
- Model Robustness
  - “the” → “teh”

Conclusions

Conclusions

- Model-based information retrieval framework
  - breaks away from the traditional index-retrieve-then-rank paradigm by encoding the knowledge contained in a corpus in a consolidated model that replaces the indexing, retrieval, and ranking components of traditional systems
  - adapt to new low resource tasks and corpora (via zero- and few-shot learning), and can be used to synthesize high quality responses that go well beyond what today's search and question answering systems are capable of.

Thank you