Entity Representations and Entity Ranking

DR. SHUBHAM CHATTERJEE

RESEARCH ASSOCIATE

DEPARTMENT OF COMPUTING SCIENCE

UNIVERSITY OF GLASGOW, UK

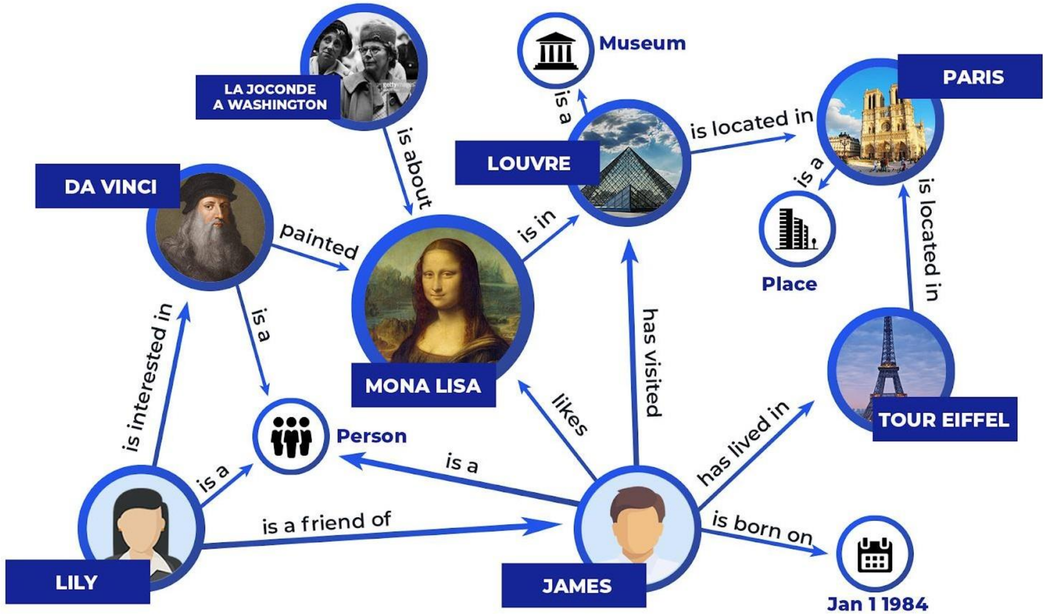
What is Entity Ranking?

What is an Entity?

- Uniquely identifiable object or thing.
- Characterized by: Name(s), Type(s), Attributes, and Relationships to other entities.

- **Name**: Food and Drug Organization
- **Type**: Organization
- **Relation**: *Located_In* (USA)

Knowledge Graph

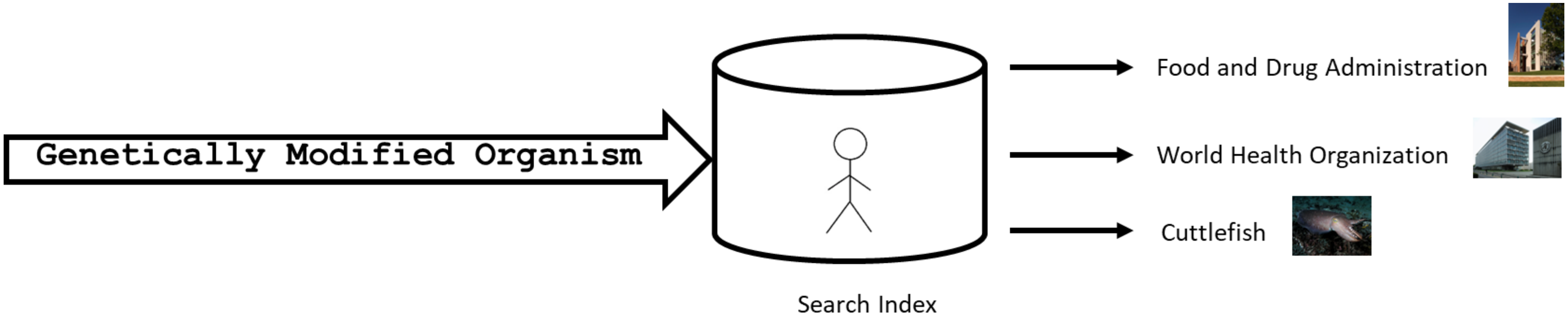


- Repository of entities.
- Nodes = entities and edges = relations.
- Contains: Types, Relations, Descriptions, etc.
- Examples: DBpedia, YAGO, Wikipedia, etc.

Picture credit: *https://community.atlassian.com/t5/Confluence-questions/Knowledge-graph/qaq-p/1565284*

Task: Entity Ranking

Given a query and a KG, retrieve entities that are relevant to the query ordered by the relevance of each entity to the query.



University of Glasgow

Entity Ranking: Over the Years

2005–2008

TREC Enterprise track: Find experts on a given topic.

2007–2009

INEX Entity Ranking track: Predict ranked list of entities.

2009–2011

TREC Entity track: Find entities of a specified type.

2010–2011

Semantic Search Challenge: Entity search over structured

2012–2013

INEX Linked Data: Combination of textual and structured data.

2017-2023

Others: TREC News, TREC Complex Answer Retrieval, TREC Product Search (New!)

How Do We Find Entities?

Types of Entity Retrieval Models

“Traditional”/Non-Neural

- ❑ **Probabilistic Unstructured Models**: Consider an entity as a single “document”.
- ❑ **Probabilistic Fielded Models**: Consider an entity as a “document” with multiple fields.
- ❑ **Learning-To-Rank Models**: Use feature vectors of (query, entity) pairs to train a ML model.
- ❑ **Semantically Enriched Models**: Consider types and relations in KG for retrieval.

Neural

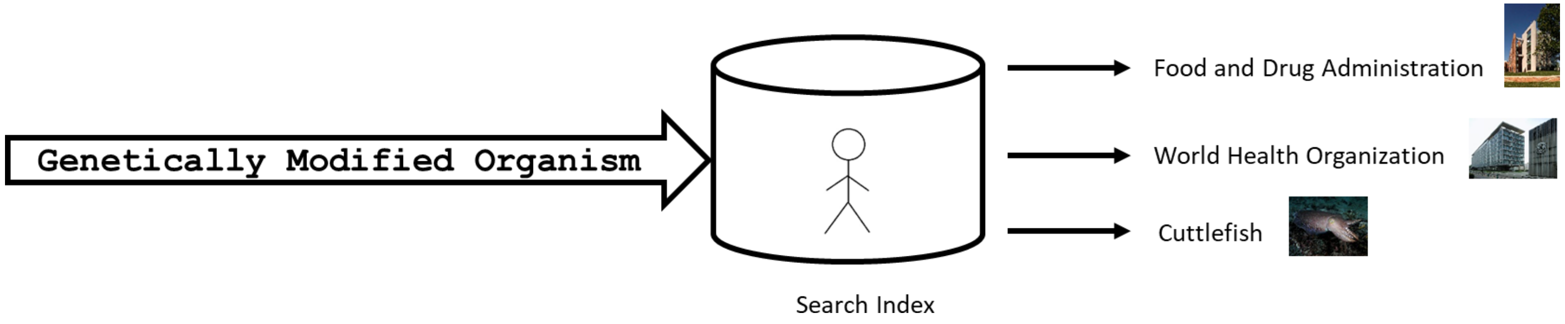
- ❑ **Graph-embedding-based Models**: Use graph embeddings for entity ranking.
- ❑ **Transformer-based Models**: Use Transformers (e.g., BERT) for entity ranking.

“Traditional”/Non-Neural Entity Ranking

How Do Systems Usually Retrieve Entities?

- Create a search index of all entities.
- Match query against representation of the entity.

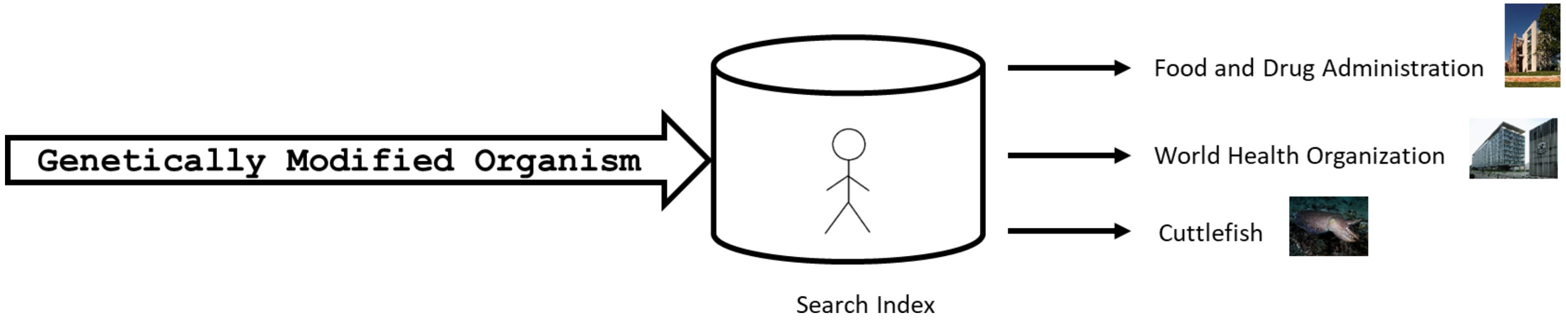
One LM for whole entity



How Do Systems Usually Retrieve Entities?

- Create a search index of all entities.
- Match query against representation of the entity.

Different LM for entity fields



Different LM for Entity Fields

Represent entities as "documents" with different fields

Entity: FDA

Name	Type
FDA	Organization
$\boldsymbol{\theta_1}$	$\boldsymbol{\theta_2}$

Different language models for different fields.

Different LM for Entity Fields

Entity: FDA

Represent entities as "documents" with different fields

Name	Type
FDA	Organization
$\boldsymbol{\theta_1}$	$\boldsymbol{\theta_2}$

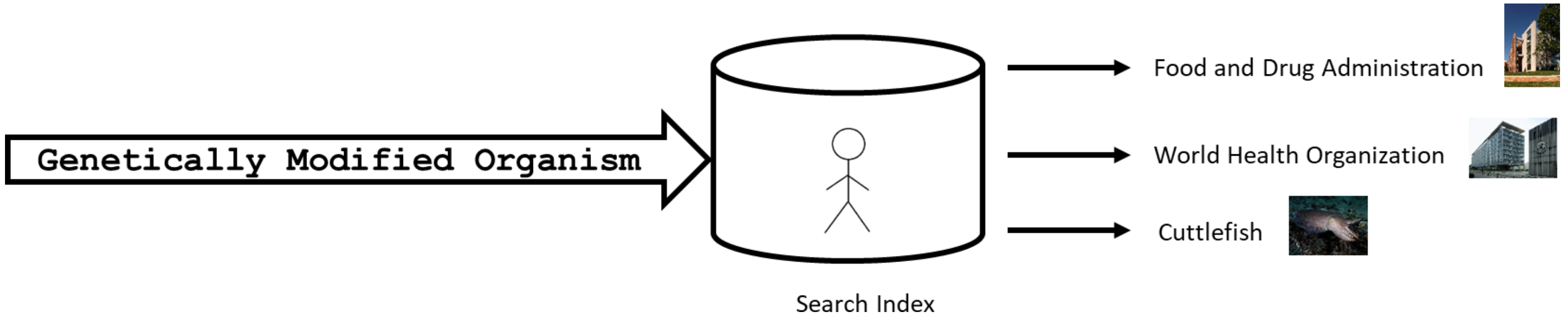
Combine using a linear mixture with field weight $\boxed{\alpha}$

Use Coordinate Ascent!

How Do Systems Usually Retrieve Entities?

- Create a search index of all entities.
- Match query against representation of the entity.

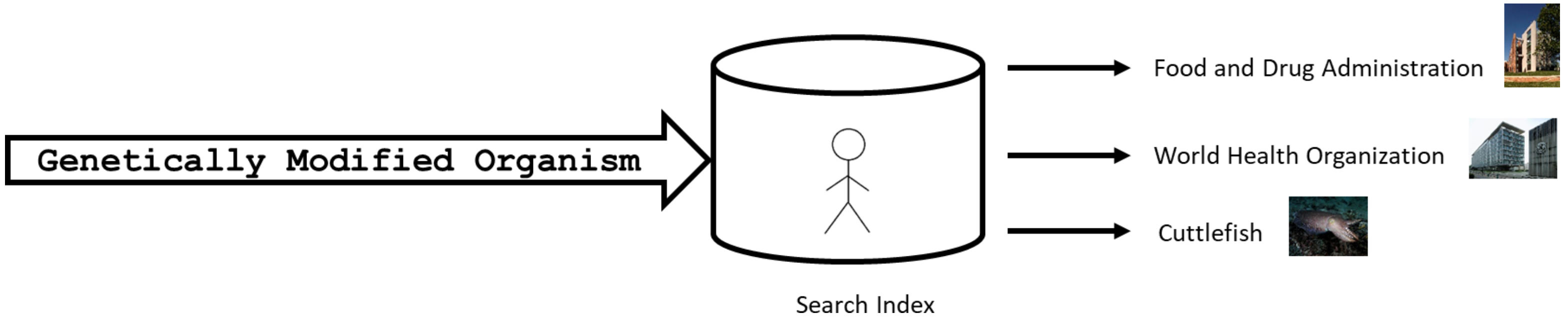
Different LM for entity fields



How Do Systems Usually Retrieve Entities?

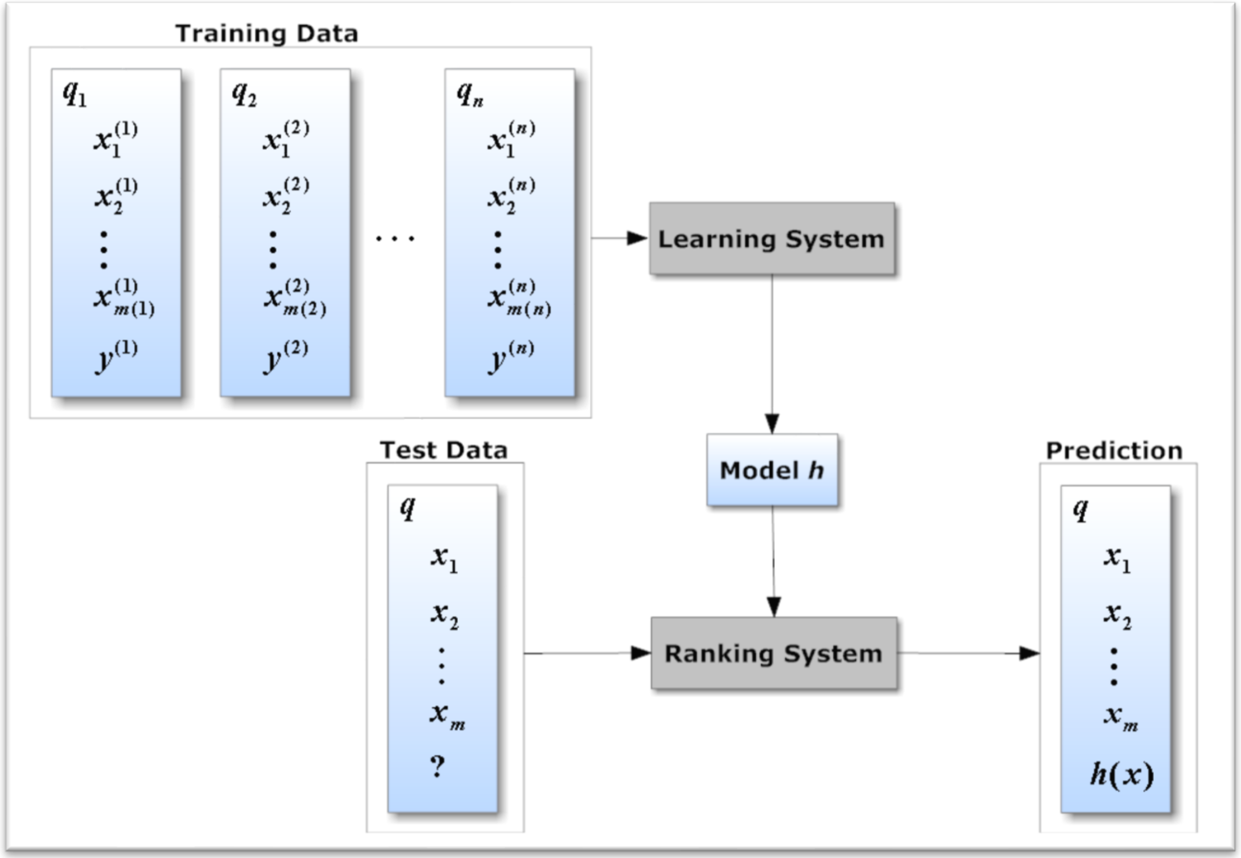
- Create a search index of all entities.
- Match query against representation of the entity.

Use features to train ML model.



Use Features to Train ML model: Learning-To-Rank

Features:
- ❑Entity retrieval from a KB using BM25+RM3.
- ❑Whether the candidate entity is contained in the query entities.
- ❑Normalized Levenshtein Distance between the query and the mention.

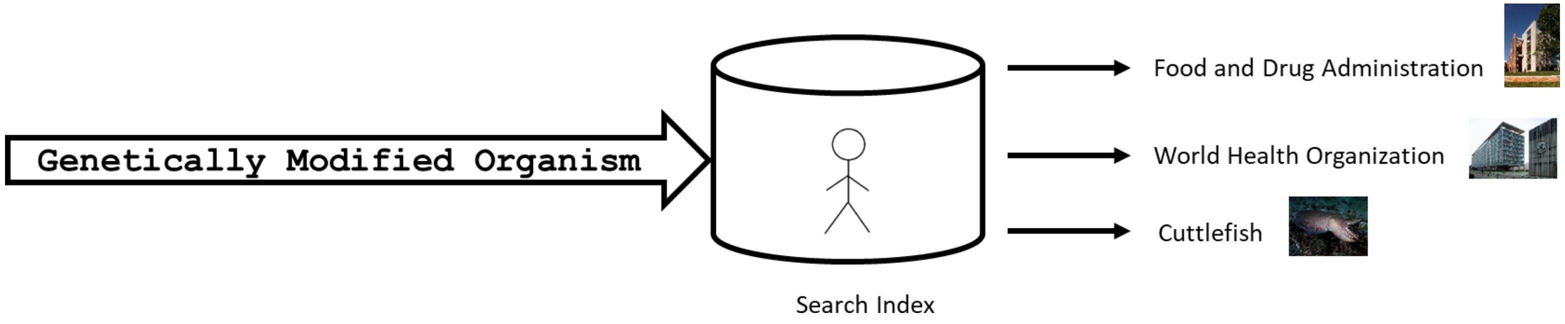


Picture credit: *http://web.ist.utl.pt/~catarina.p.moreira/coursera.html*

How Do Systems Usually Retrieve Entities?

- Create a search index of all entities.
- Match query against representation of the entity.

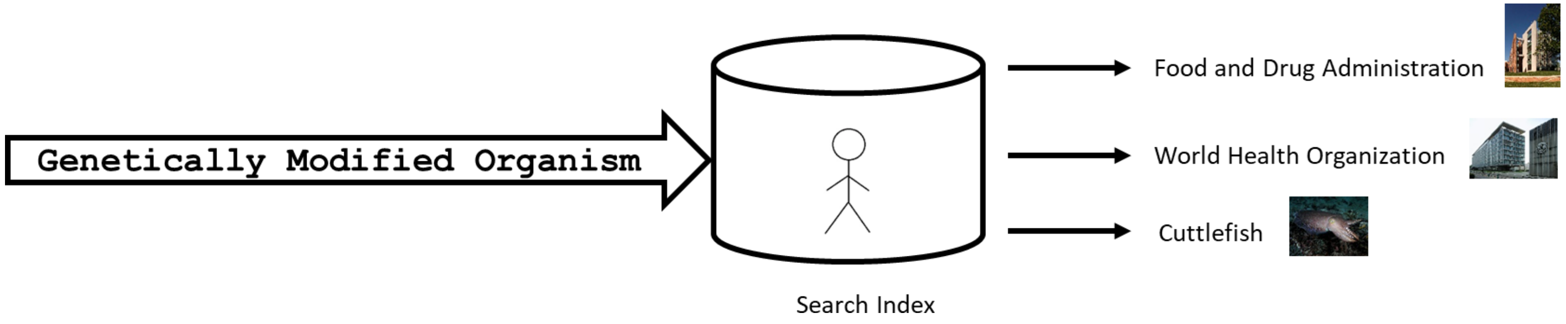
Use features to train ML model.



How Do Systems Usually Retrieve Entities?

- Create a search index of all entities.
- Match query against representation of the entity.

Use entity types from a KG.



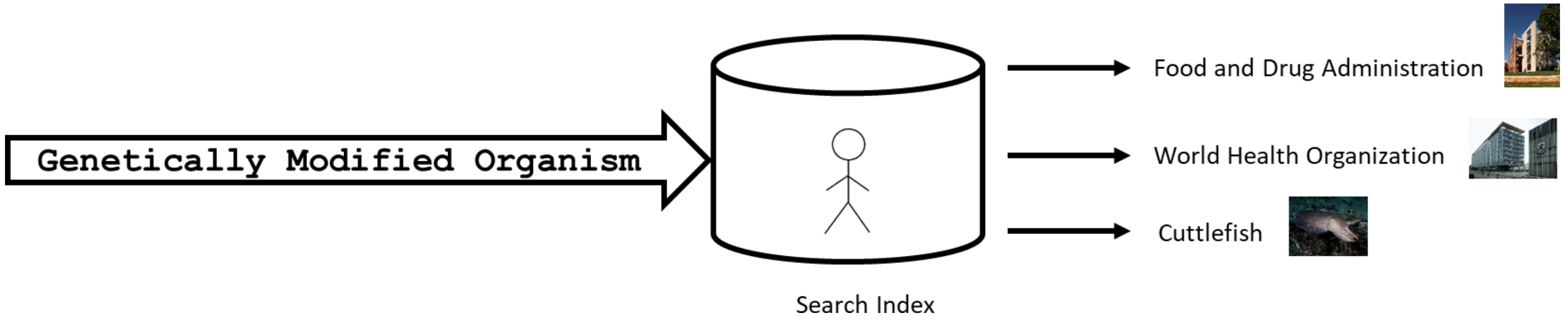
Use Entity Types From a KG: Type-Aware Entity Retrieval

- Uses entity types from a type taxonomy (e.g., Wikipedia categories).
- Query enriched with types of entities in the query (called target types)
- Example [Balog et al., 2011]
  - Learn a probability distribution over the query and entity types.
  - Similarity = KL divergence between two distributions

How Do Systems Usually Retrieve Entities?

- Create a search index of all entities.
- Match query against representation of the entity.

Use entity types from a KG.

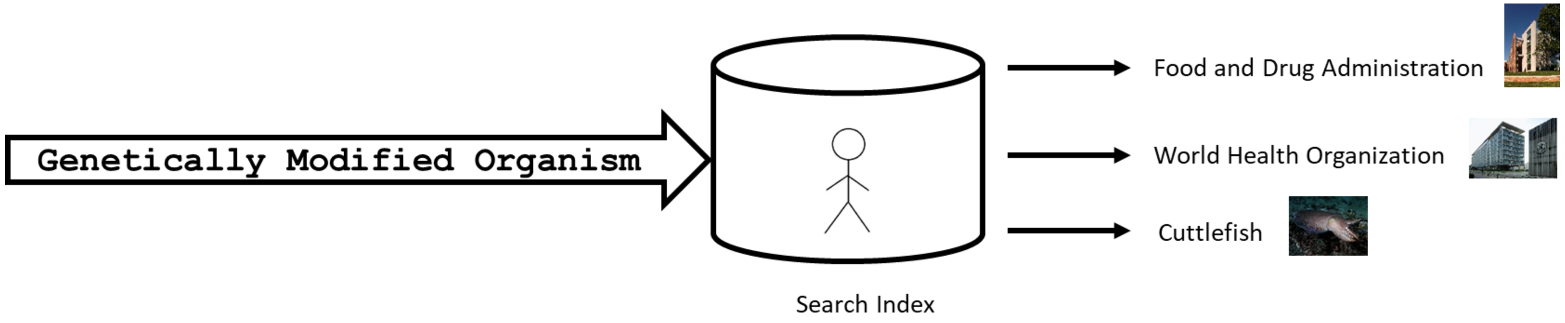


How Do Systems Usually Retrieve Entities?

- Create a search index of all entities.
- Match query against representation of the entity.

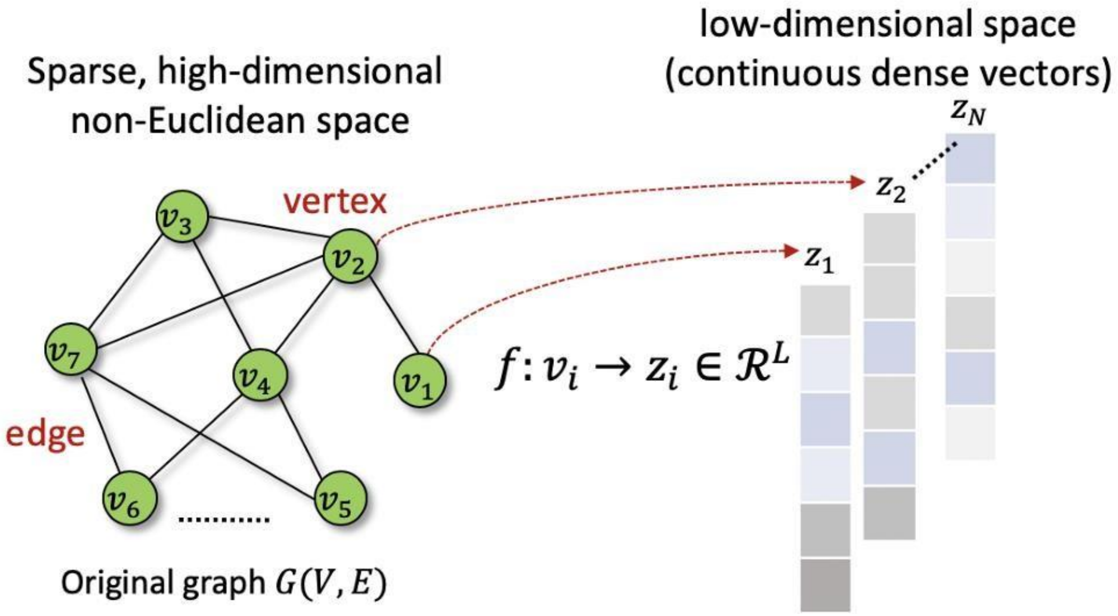
Non-neural methods: Sparse

Neural methods: Dense



Neural Entity Ranking

Graph Embeddings: Learning Dense Entity Representations



Picture Credit: *Mengjia Xu. Understanding Graph Embedding Methods and their Applications. 2020.*

- ❑Treat the knowledge repository as a KG.
- ❑Convert high-dimensional KG into low dimensional, dense and continuous vector spaces.
- ❑Graph structure properties are maximally preserved.
- ❑Examples: Wikipedia2Vec, TransE, TransR, etc.

Gerritse et al., 2020

GEEER: Using Wikipedia2Vec for Entity Ranking

- Re-ranks entities using Wikipedia2Vec.
- Shows that Wikipedia2Vec is useful for entity ranking.
- **General Idea**: Relevant entities for a given query are situated close (in graph embedding space) to the query entities identified by the entity linker.
- **Method**:
  - Compute the embedding-based score for an entity:

$$Score_{emb}(E,Q) = \sum_{e \in Q} C(e) \cdot \cos(\vec{E}, \vec{e})$$

  - Final Score = interpolation of the embedding-based and retrieval scores.

$$Score_{final}(E,Q) = \lambda \cdot Score_{emb}(E,Q) + (1-\lambda) \cdot Score_{ret}(E,Q)$$

Gerritse et al., 2020

GEEER: Using Wikipedia2Vec for Entity Ranking

- Re-ranks entities using Wikipedia2Vec.
- Shows that Wikipedia2Vec is useful for entity ranking.
- **General Idea**: Relevant entities for a given query are situated close (in graph embedding space) to the query entities identified by the entity linker.
- **Method**:
  - Compute the embedding-based score for an entity:

$$Score_{emb}(E,Q) = \sum_{e \in Q} \boxed{C(e)} \cdot \cos(\vec{E}, \vec{e})$$

Confidence score of entity linker.

  - Final Score = interpolation of the embedding-based and retrieval scores.

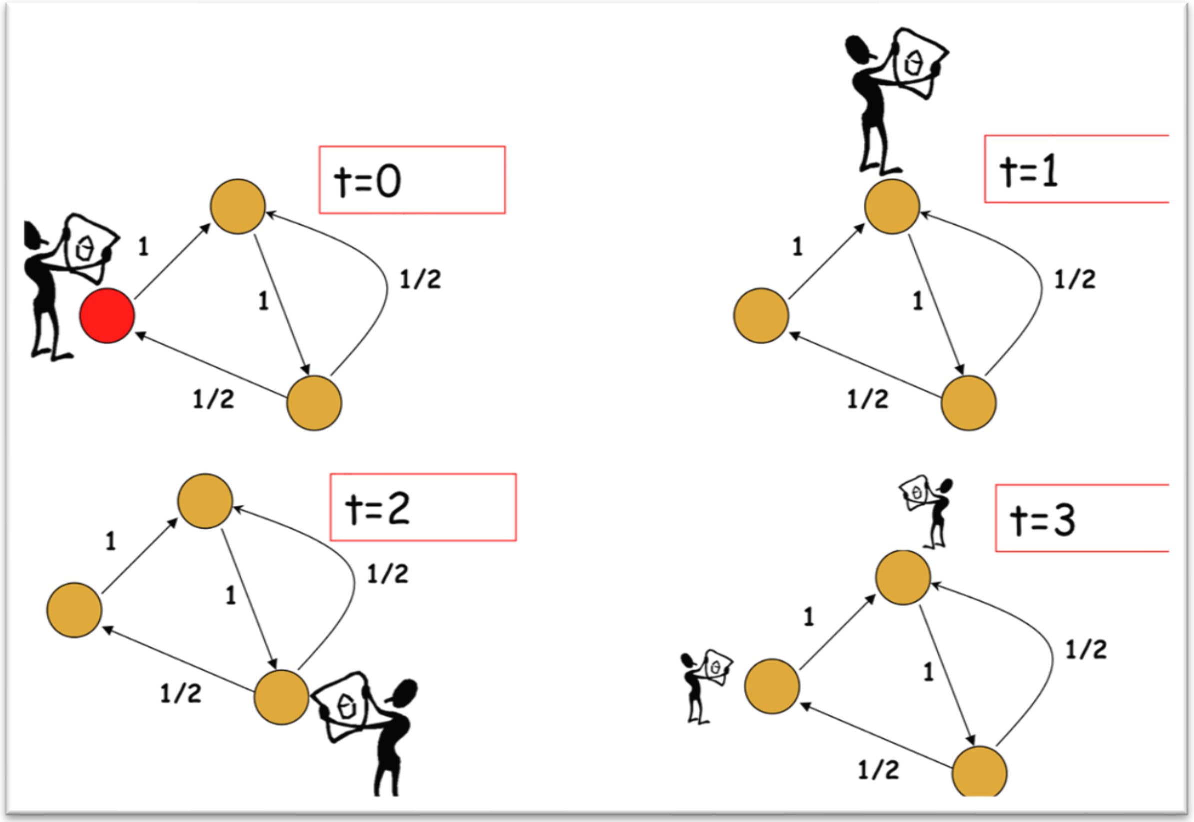
$$Score_{final}(E,Q) = \boxed{\lambda} \cdot Score_{emb}(E,Q) + (1-\lambda) \cdot Score_{ret}(E,Q)$$

Use LTR optimized for NDCG@100.

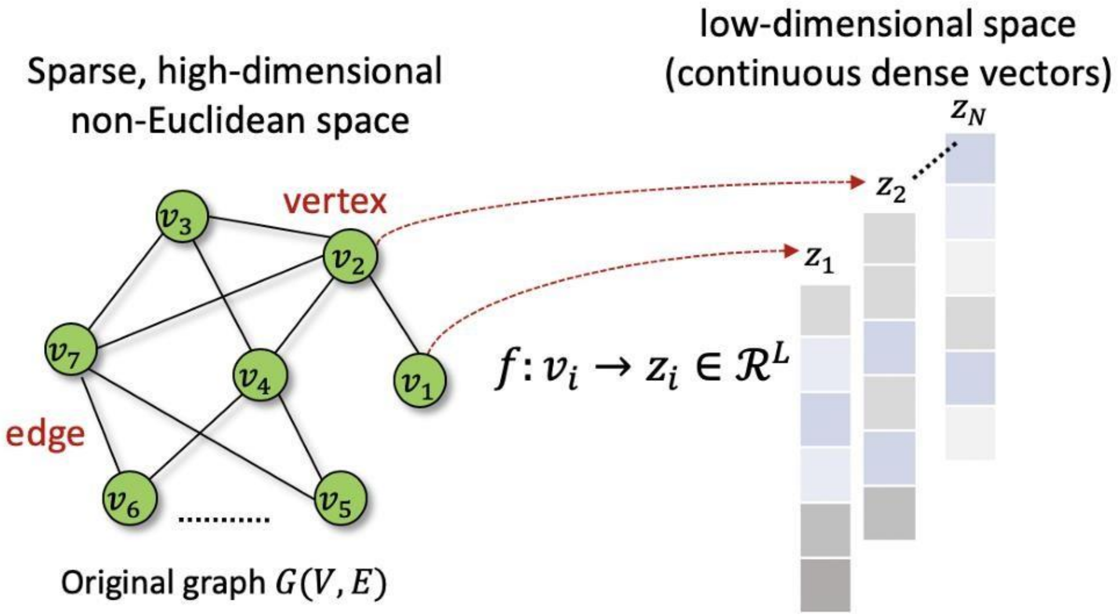
Nikolaev and Kotov., 2020

KEWER: Knowledge graph Entity and Word Embedding for Retrieval

- Joint embedding of entities and words (like Wikipedia2Vec).
- **Uses**: Entities, attributes, categories, KG structure.
- **Method**: Random walks over a KG.
  - Start from each entity.
  - Repeatedly follow directed edges
- **Training**: Skip-Gram-based model using generated random walks.
- **Entity Ranking**: Cosine similarity between query and entity embedding.



Graph Embeddings: Issues for IR



Picture Credit: *Mengjia Xu. Understanding Graph Embedding Methods and their Applications. 2020.*

- ❑ **IR: Considers explicit query.**
- ❑ **Current graph embedding methods: Not seen the query during training (query-agnostic!)**

How About We Use BERT?

- BERT shown to be useful for document retrieval.
- **Question:** Can we use BERT for entity ranking?

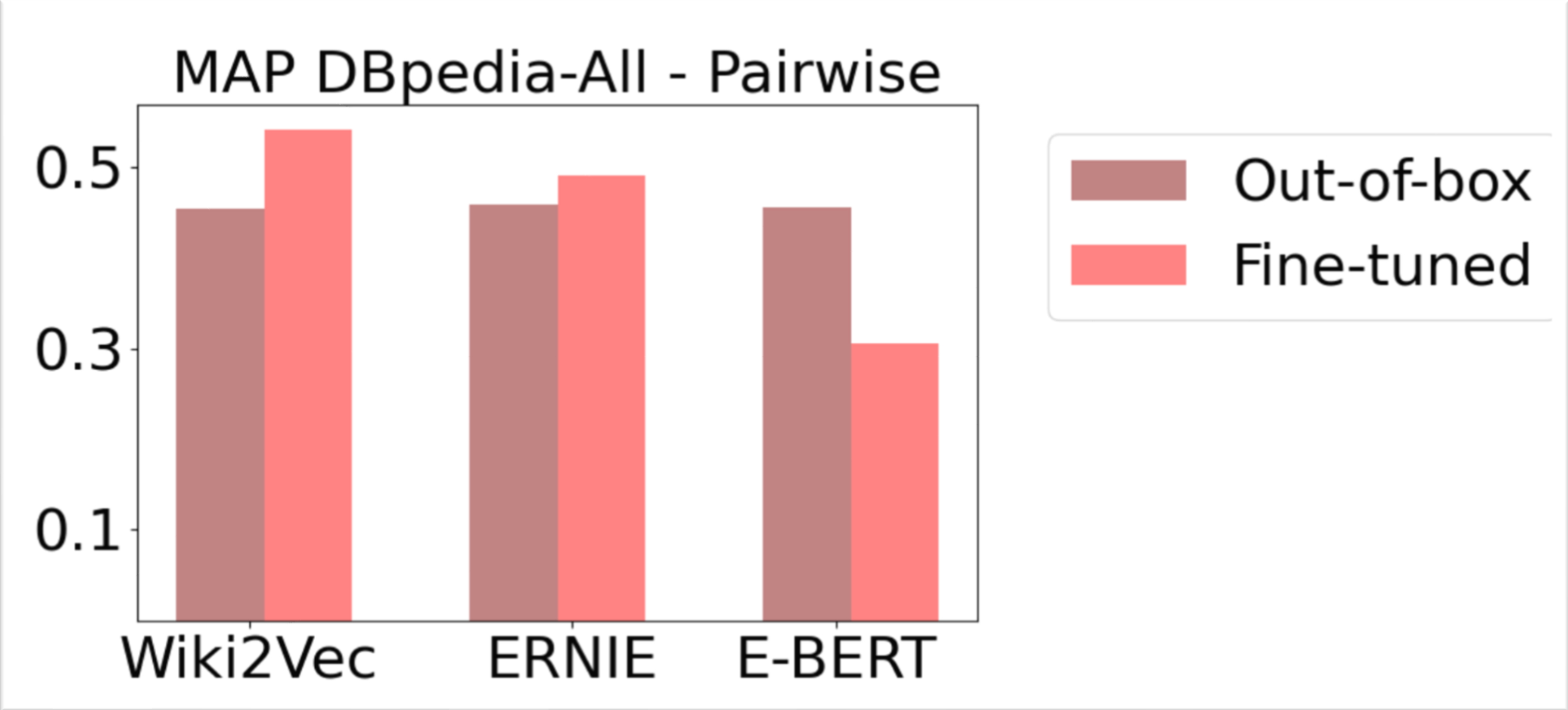
Poerner et al., 2020

One Idea: E-BERT: Injecting Entity information into BERT

- Aligns Wikipedia2Vec entity vectors with BERT's native word piece vectors.
- **Issue:** BERT's word piece dictionary does not contain any entities!
- **Solution:** Use common words in the vocabulary of BERT and Wikipedia2Vec.
- Wikipedia2Vec embeds words and entities into the same vector space → Alignment learnt using words can also be applied to entities.

Oza and Dietz., 2023

How Well Does BERT understand Entities?



Results for GEEER with ERNIE and E-BERT. Dbpedia-Entity v2.

Gerritse et al., 2022

EM-BERT: Entity-enriched BERT for Entity Ranking

- Aligns Wikipedia2Vec entity vectors with BERT's native word piece vectors (as in E-BERT).
- Entity vectors used with entity mentions.
- Fine-tuning: First MS MARCO passages, then DBpedia-Entity v2.

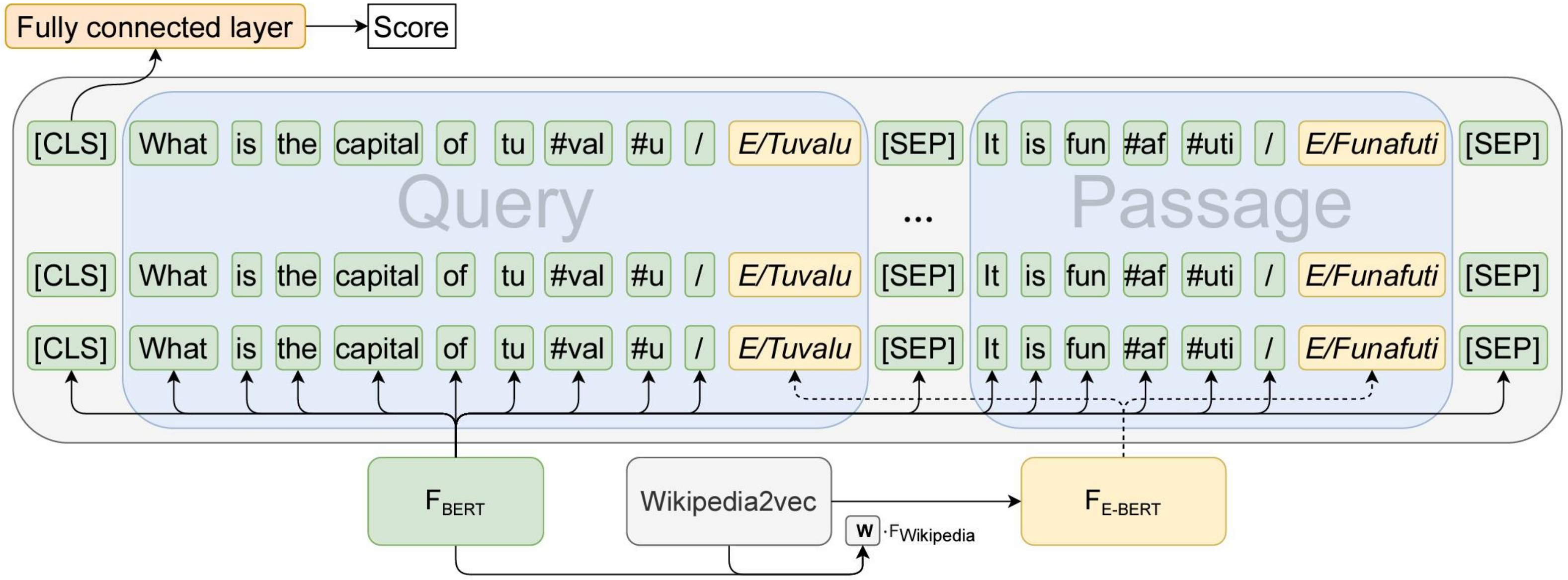


Figure 1: Illustration of the EM-BERT model. Entity annotated query and documents are tokenized and mapped to their corresponding vector representations using F_{BERT} and F_{E-BERT} functions.

Picture Credit: *Gerritse et al. Entity-aware Transformers for Entity Search. 2022.*

Gerritse et al., 2022

EM-BERT: What do we learn?

- Substantial improvements over SOTA for entity ranking!
- Helps:
  - Complex natural language queries,
  - List search queries, and
  - Queries containing tail entities

One Idea: Injecting Entity information into BERT

Another Idea: Can we utilize existing knowledge in BERT?

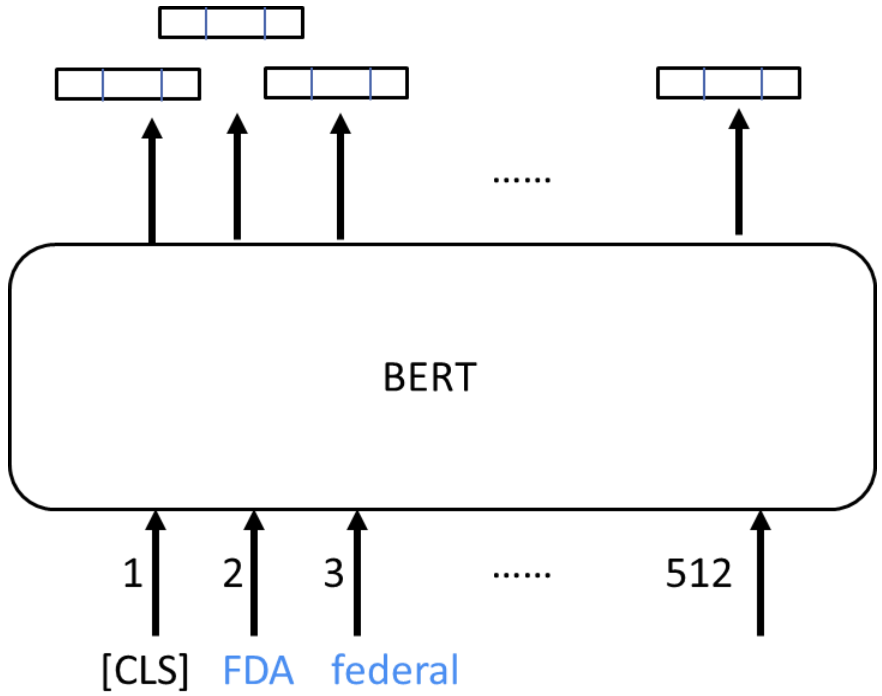
- BERT has already seen a lot of the world (from books and Wikipedia).
- BERT can probably infer the connection between the query and entity from a term-based entity description.
- Term-based entity description = Introductory Wikipedia paragraph (most often).

Query: Genetically Modified Organism
Relevant Entity: Food and Drug Administration

Lead Text

Food and Drug Administration

The United States **Food and Drug Administration** (**FDA** or **USFDA**) is a federal agency of the Department of Health and Human Services. The FDA is responsible for protecting and promoting public health through the control and supervision of food safety, tobacco products, dietary supplements, prescription and over-the-counter pharmaceutical drugs (medications), vaccines, biopharmaceuticals, blood transfusions, medical devices, electromagnetic radiation emitting devices (ERED), cosmetics, animal foods & feed[3] and veterinary products.



- Lead Text = Static entity description
- No knowledge of the query!
- Static entity embeddings.

Query: Genetically Modified Organism
Relevant Entity: Food and Drug Administration

FDA regulates most human and animal food, including GMO foods. In doing so, FDA makes sure that foods that are GMOs or have GMO ingredients meet the same strict safety standards as all other foods. FDA sets and enforces food safety standards [...]

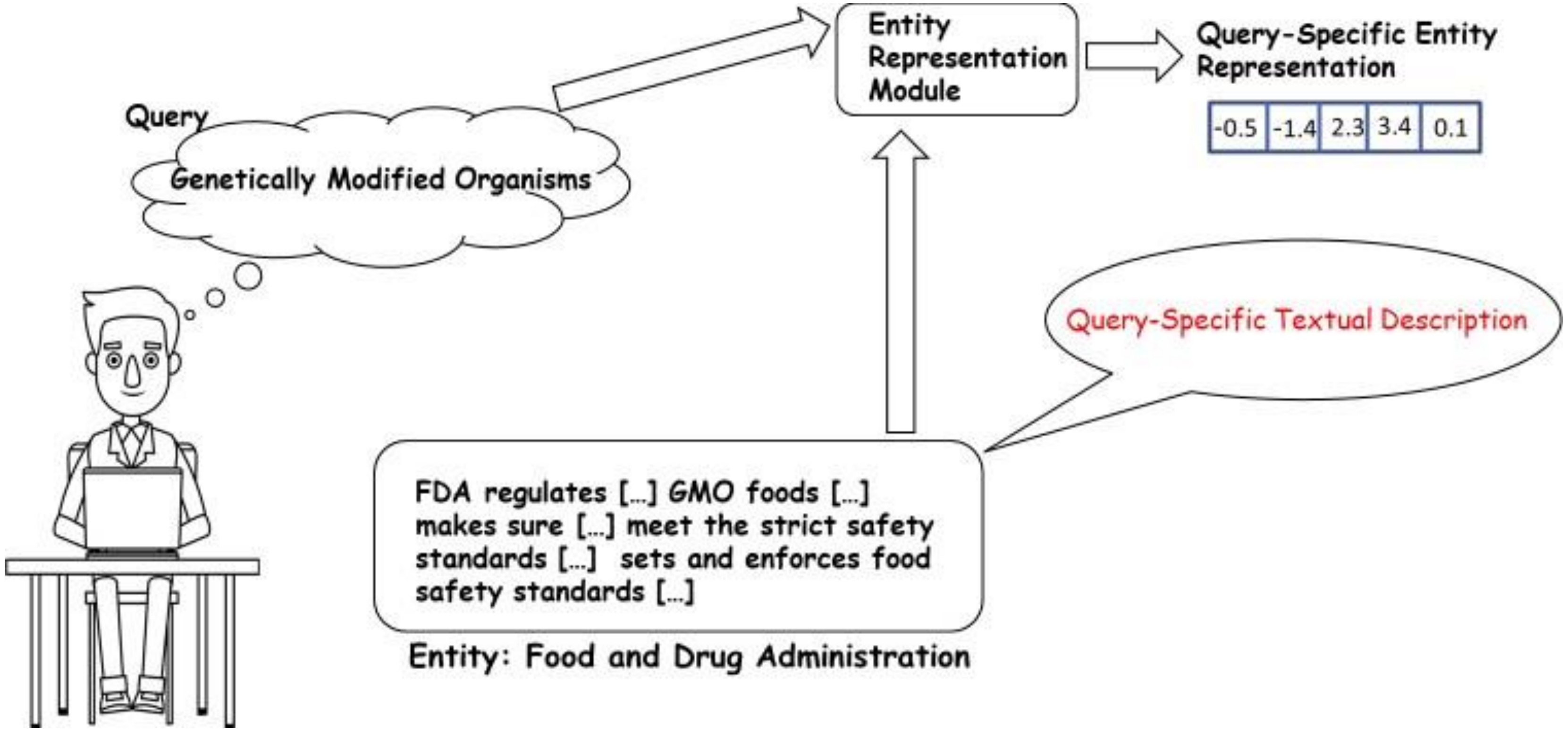
Clarifies the connection between the query and entity.

What if we had this text instead of lead text?

Chatterjee and Dietz., 2022

BERT-ER: Query-Specific Entity Descriptions

Query-specific entity descriptions→ descriptions that mention relevant connections between the query and the entity.



Food and Drug Administration

The United States **Food and Drug Administration** (**FDA** or **USFDA**) is a federal agency of the Department of Health and Human Services. The FDA is responsible for protecting and promoting public health through the control and supervision of food safety, tobacco products, dietary supplements, prescription and over-the-counter pharmaceutical drugs (medications), vaccines, biopharmaceuticals, blood transfusions, medical devices, electromagnetic radiation emitting devices (ERED), cosmetics, animal foods & feed[3] and veterinary products.

1 Organizational structure
2 Location
3 Scope and funding
4 Regulatory programs
5 Science and research programs
6 Data management
7 History

Aspects = Top-Level Sections

Chatterjee and Dietz., 2022

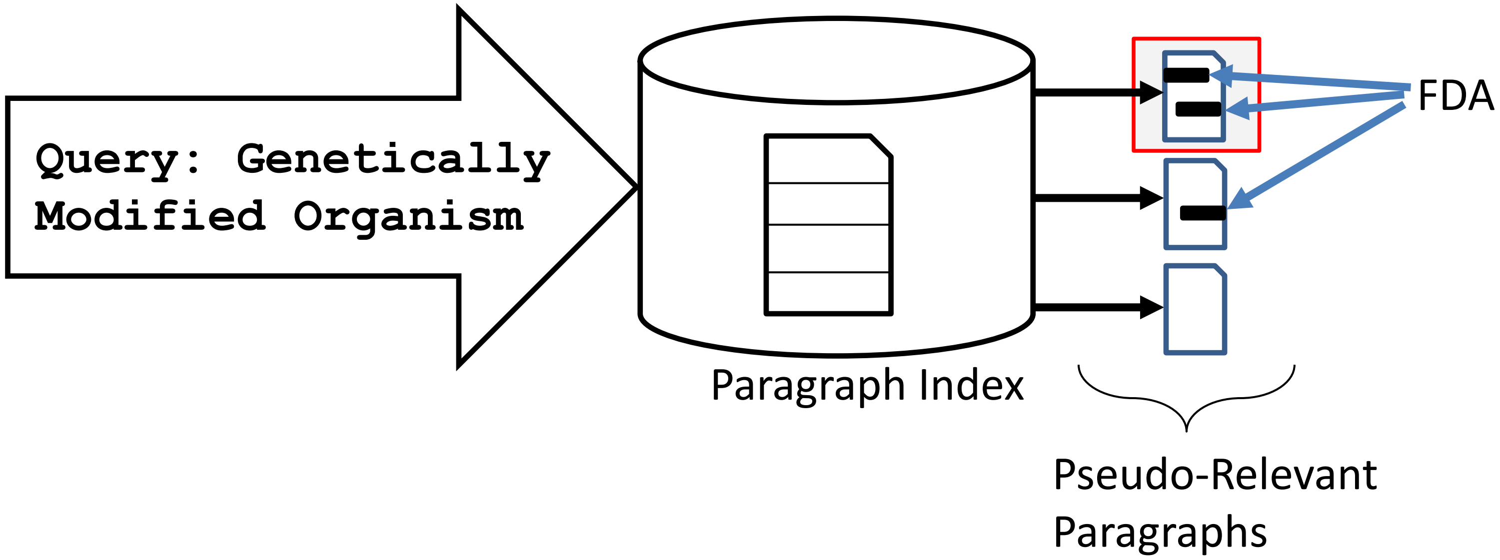
Query-Specific Entity Descriptions: Alternative 1

- ❑**Using Wikipedia: Top-Level Sections**
- ❑Identify relevant top-level sections from the Wikipedia page. ***(Why? —because the lead text is does not elaborate the relevance!)***
- ❑Use catalog of top-level sections (aspects) from Ramsdell et al., 2020.
- ❑ **Downside: *Wikipedia articles often do not contain all relevant information!***

Query-Specific Entity Descriptions: Alternative 2

Chatterjee and Dietz., 2022

- ❑ **Using Paragraph Collection: PRF Passages**

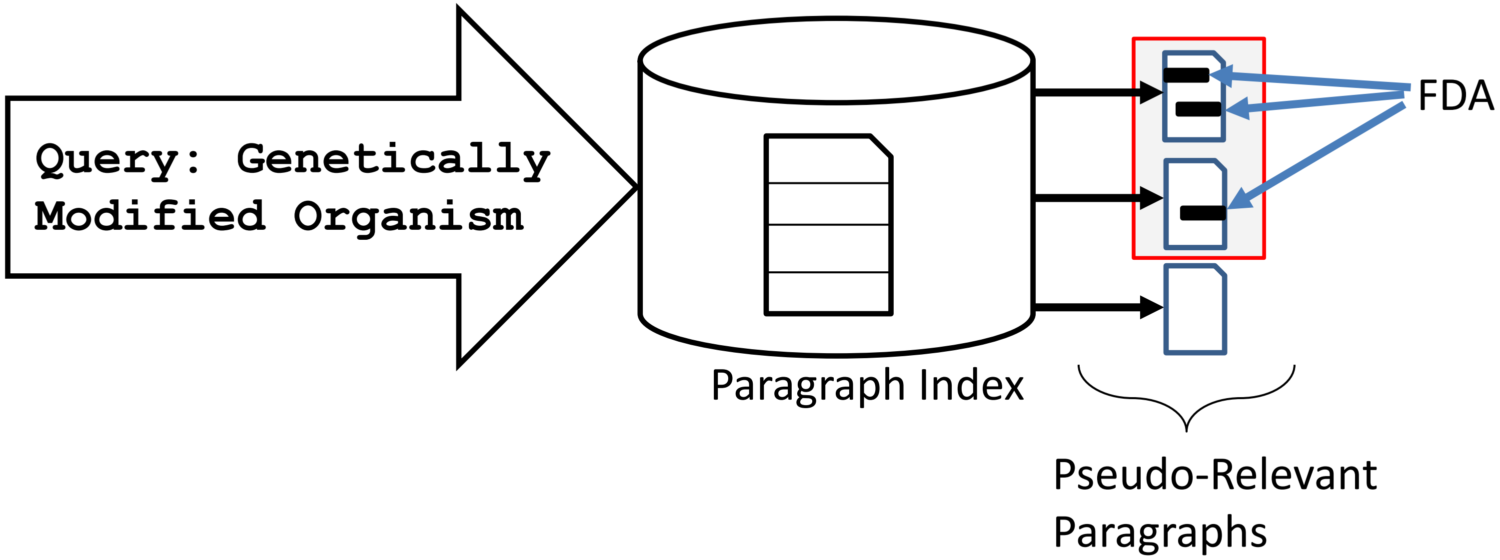


❑ **Downside:** ***Entity may not be central to the discussion in the text!***

Query-Specific Entity Descriptions: Alternative 3

Chatterjee and Dietz., 2022

- ❑ **Using Paragraph Collection: Entity-Support Passages**



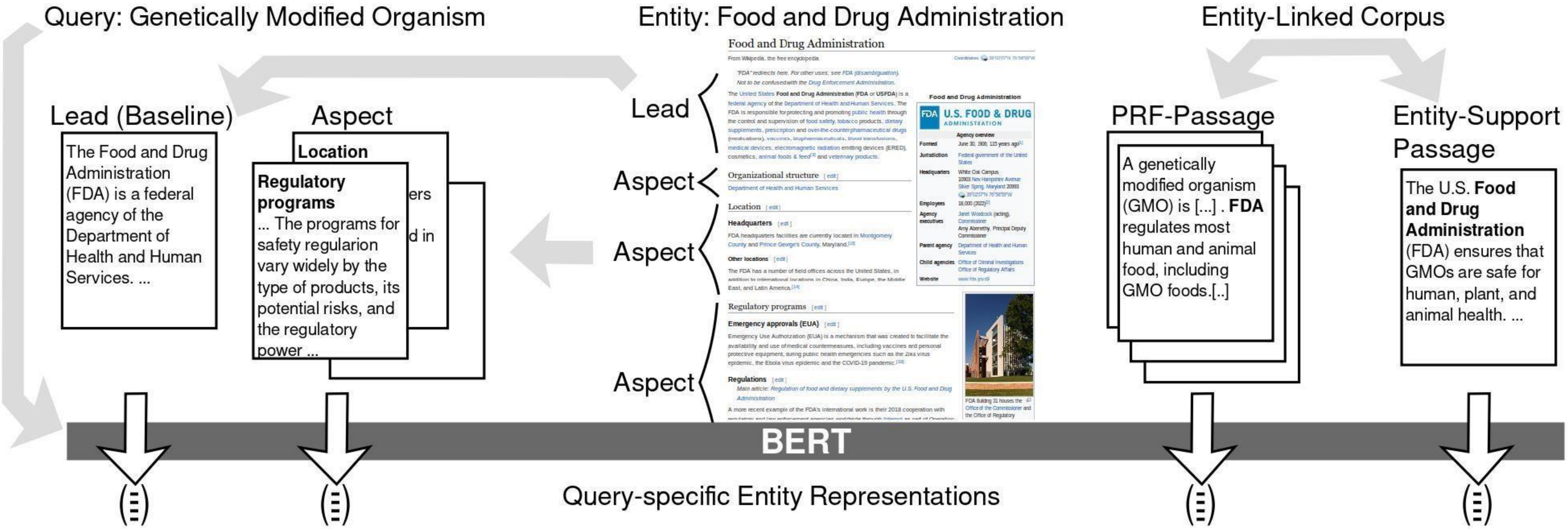
- ❑ Re-rank these documents.
- ❑ Two criteria:

1. How many relevant connections between query and entity?
2. Are the relevant connections central to the discussion in the text?

Entity-Support Passages: ***Chatterjee and Dietz, ICTIR 2019.***

BERT-ER: Alternative Approaches for Query-Specific Entity Description s

Chatterjee and Dietz, SIGIR 2022



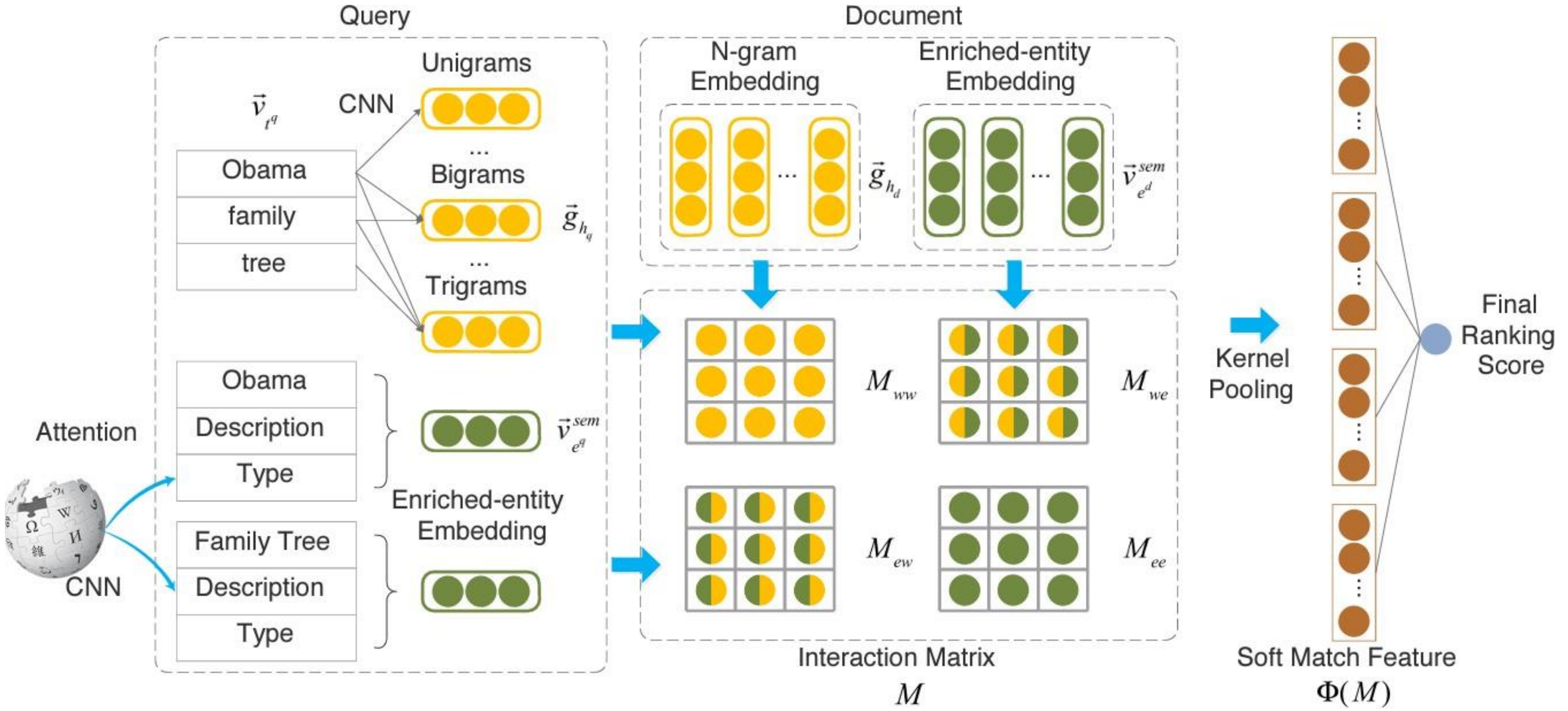
Shubham Chatterjee and Laura Dietz. 2022. ***BERT-ER: Query-specific BERT Entity Representations for Entity Ranking.*** In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (**SIGIR '22**).

How Do We Use Entities?

Entity-Duet Neural Ranking Model

Liu et al., 2018

- ❑ Incorporates entities in interaction-based neural ranking models.
- ❑ Learns entity representations using: descriptions and types.



***Picture Credits**: Liu et al. Entity-Duet Neural Ranking. 2018.*

Dense Retrieval With Entity Views

Tran and Yates, 2022

- Enrich query/document representation with entity representations.
- Cluster entities → Entity embeddings using clusters → Clusters act as "views" of the document.

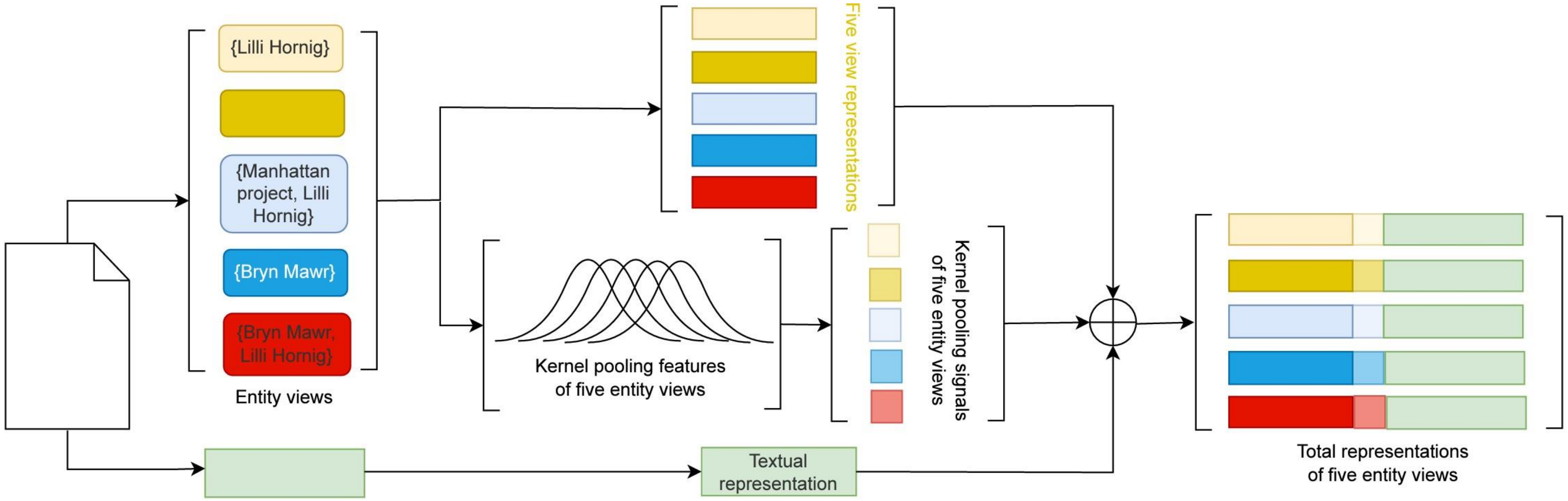


Figure 1: Overview of EVA with multiple representations. Entity clusters such as {Lilli Hornig}, {Manhattan project}, {Manhattan project, Lilli Hornig}, {Bryn Mawr} and {Bryn Mawr, Lilli Hornig} can be understood as different entity views of the passage. EVA generates one total representation for each view, which enriches a textual representation with the entities present.

***Picture Credits**: Tran and Yates. Dense Retrieval with Entity Views. 2022.*