Material: https://github.com/laura-dietz/neurosymbolic-representations-for-IR/

Entity Representations and Entity Ranking

DR. SHUBHAM CHATTERJEE

RESEARCH ASSOCIATE

DEPARTMENT OF COMPUTING SCIENCE

UNIVERSITY OF GLASGOW, UK

Shubham Chatterjee. Neuro-Symbolic Approaches for Information Retrieval. ECIR 2023
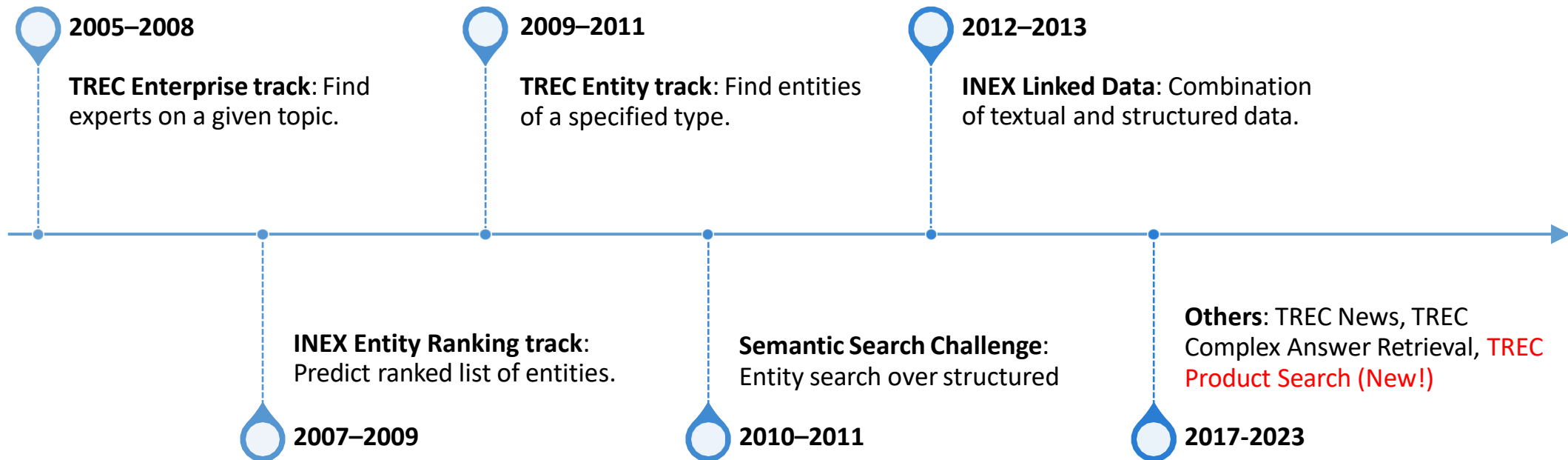
Task: Entity Ranking

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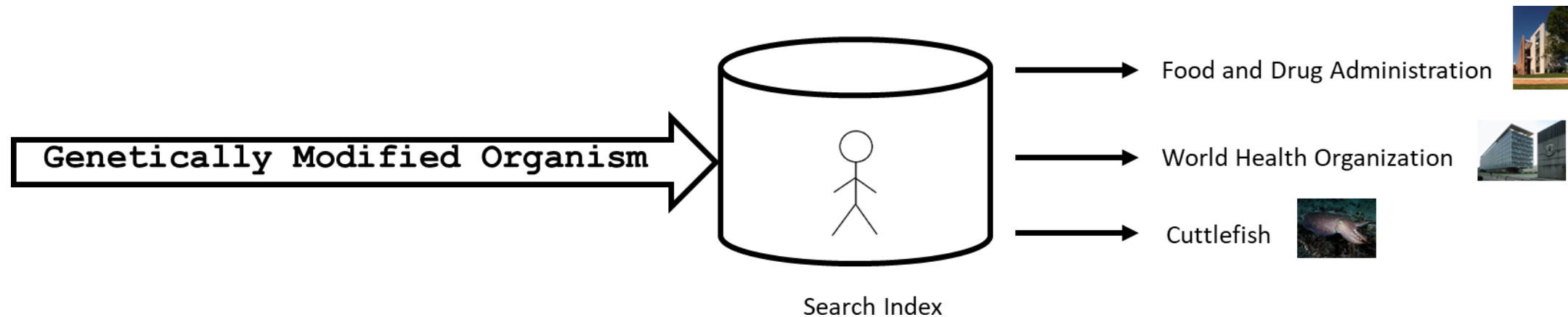
Entity Ranking: Over the Years

2005–2008

TREC Enterprise track: Find experts on a given topic.

2009–2011

TREC Entity track: Find entities of a specified type.

2012–2013

INEX Linked Data: Combination of textual and structured data.

INEX Entity Ranking track: Predict ranked list of entities.

2007–2009

Semantic Search Challenge: Entity search over structured

2010–2011

Others: TREC News, TREC Complex Answer Retrieval, TREC Product Search (New!)
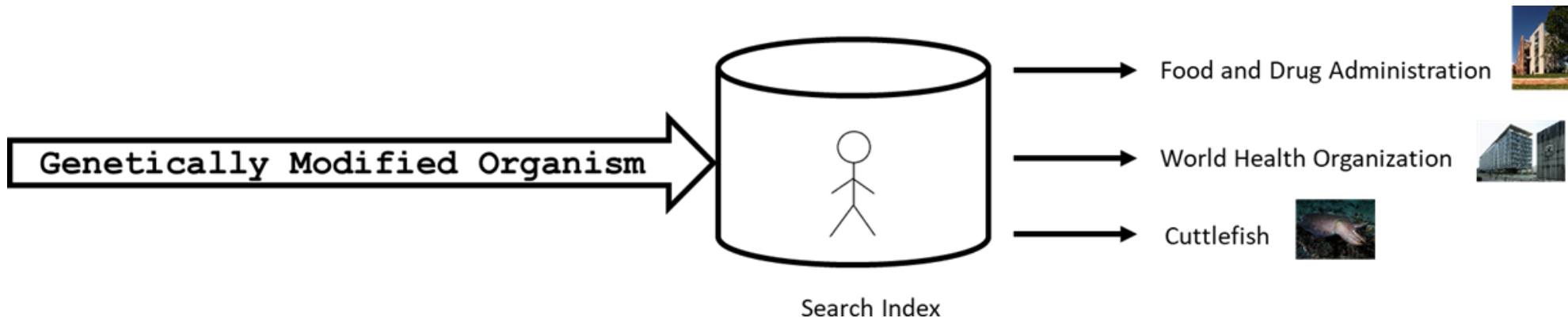
2017-2023

How Do Systems Usually Retrieve Entities?

Probabilistic Unstructured Models

Consider an entity as a single "document".

❑ Create a search index of all entities.

❑ Match the query against the entity representation.

One LM for whole entity

Genetically Modified Organism → Search Index

- Food and Drug Administration
- World Health Organization
- Cuttlefish

Probabilistic Unstructured Models

Consider an entity as a single "document".

❑ Create a search index of all entities.

❑ Match the query against the entity representation.
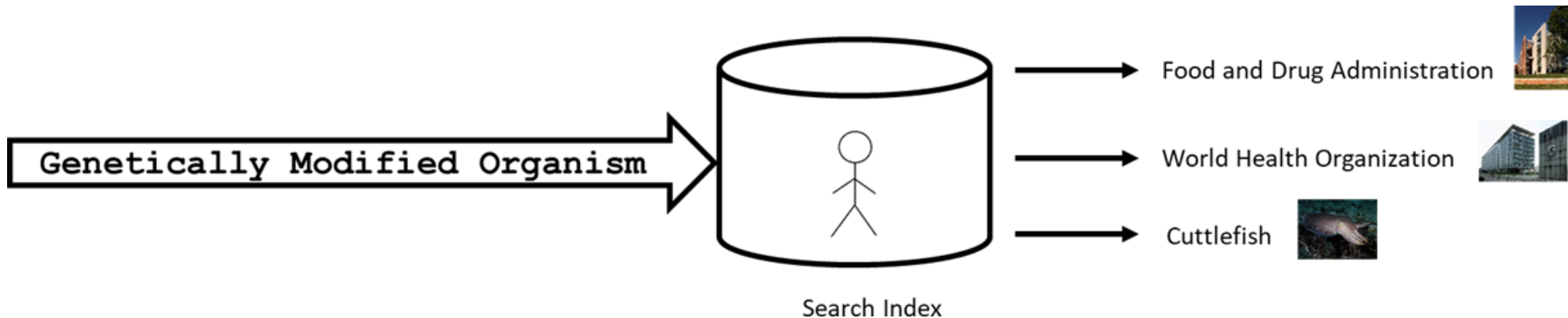
One LM for whole entity

❑ **Probability distribution over sequences of words.**

❑ **Generate probabilities by training on text corpora in one or many languages.**

Fielded Models

Consider an entity as a "document" with multiple fields.

❑ Create a search index of all entities.

❑ Match the query against the entity representation.
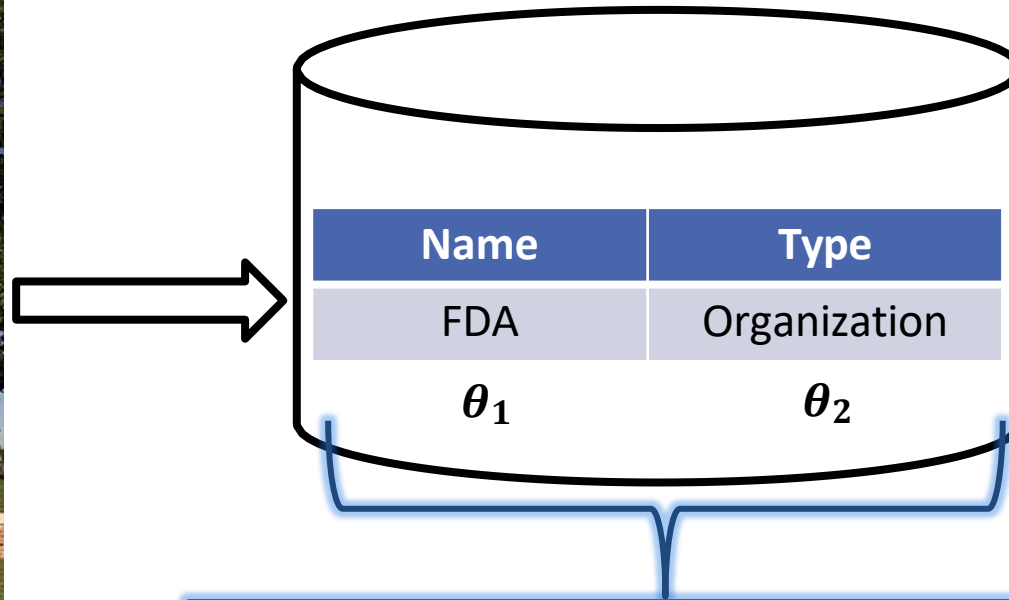
Multiple LMs for different fields

Genetically Modified Organism → Search Index

Food and Drug Administration

World Health Organization

Cuttlefish

Fielded Models: Multiple LM for Different Fields

Represent entities as "documents" with different fields.

Entity: FDA

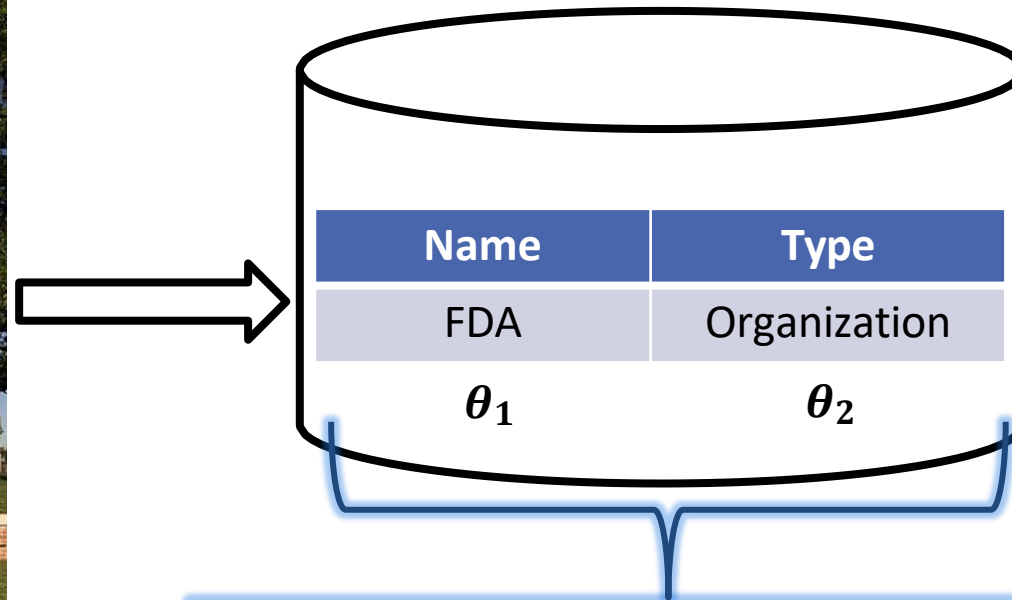| Name	Type
FDA	Organization
θ_1	θ_2

Different language models for different fields.

Fielded Models: Multiple LM for Different Fields

Represent entities as "documents" with different fields.

Name	Type
FDA	Organization
θ_1	θ_2

Use Coordinate Ascent

**Entity: FDA**

Combine using a linear mixture with field weights.

Learning-to-Rank Models

Use feature vectors of (query, entity) pairs to train a ML model.

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*
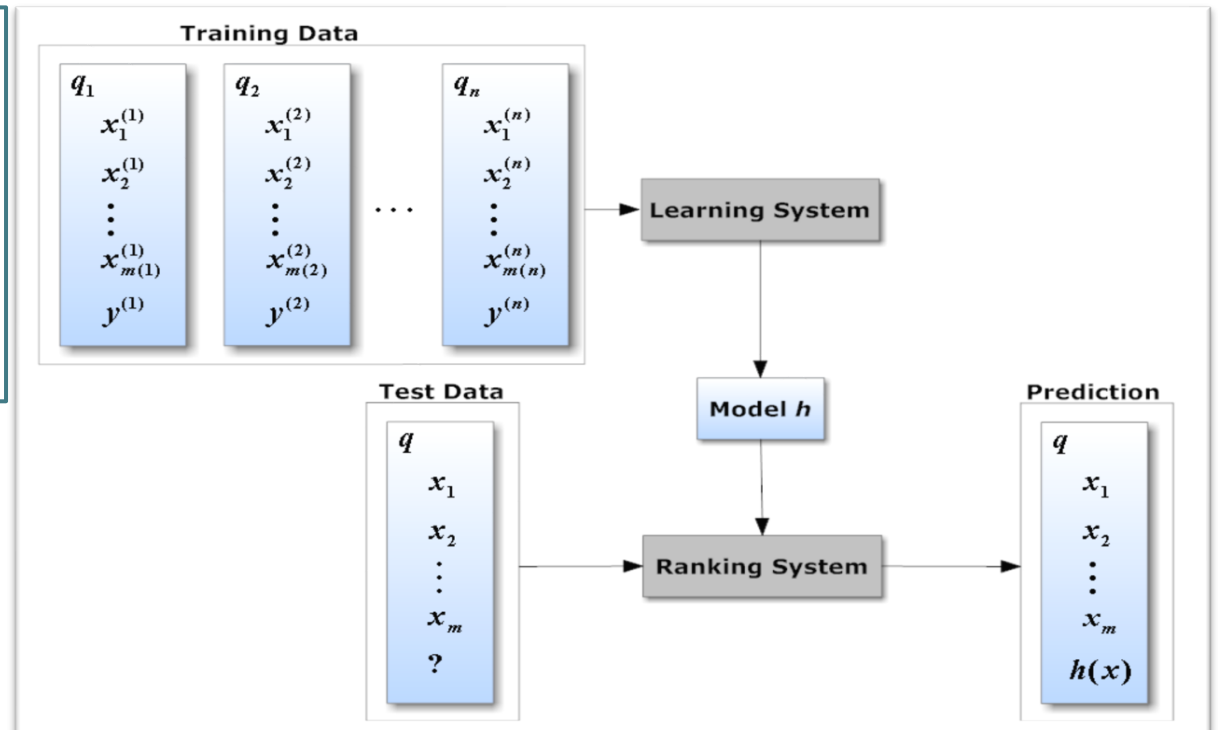
Type-Aware Entity Retrieval
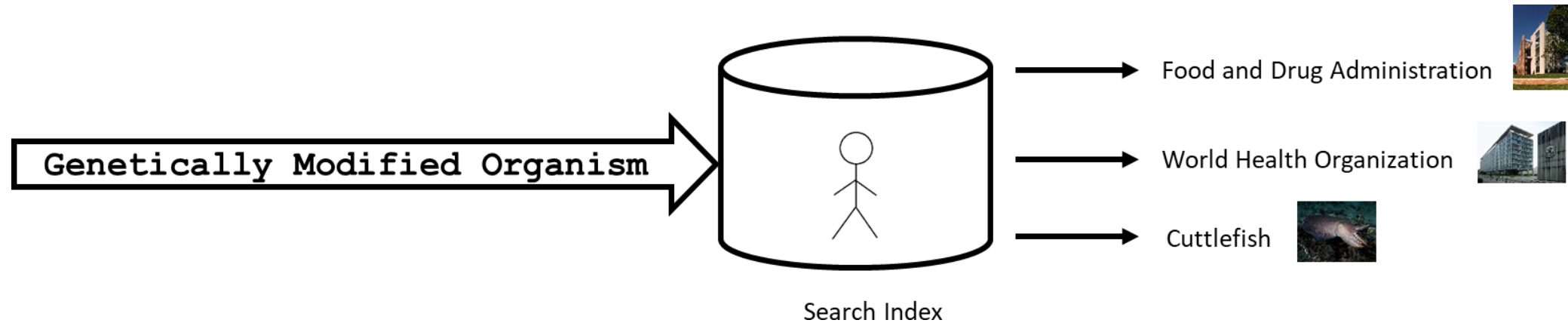
❑ Uses entity types from a type taxonomy (e.g., Wikipedia categories).

❑ 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 the query against the representation of the entity.

Non-neural methods: Sparse

Neural methods: Dense

Genetically Modified Organism → Search Index

Food and Drug Administration

World Health Organization

Cuttlefish

Search Index

Neural Entity Ranking

Graph Embeddings: Learning Dense Entity Representations

Sparse, high-dimensional non-Euclidean space

vertex

edge

Original graph $G(V, E)$

low-dimensional space (continuous dense vectors)

$f: v_i \rightarrow z_i \in \mathcal{R}^L$

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.

4/3/2023

Shubham Chatterjee. Neuro-Symbolic Approaches for Information Retrieval. ECIR 2023

15

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.

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

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:

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

Confidence score of entity linker.

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

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

Use LTR optimized for NDCG@100.

Shubham Chatterjee. Neuro-Symbolic Approaches for Information Retrieval. ECIR 2023

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

❑ **Entity Ranking**: Cosine similarity between query and entity embeddings.
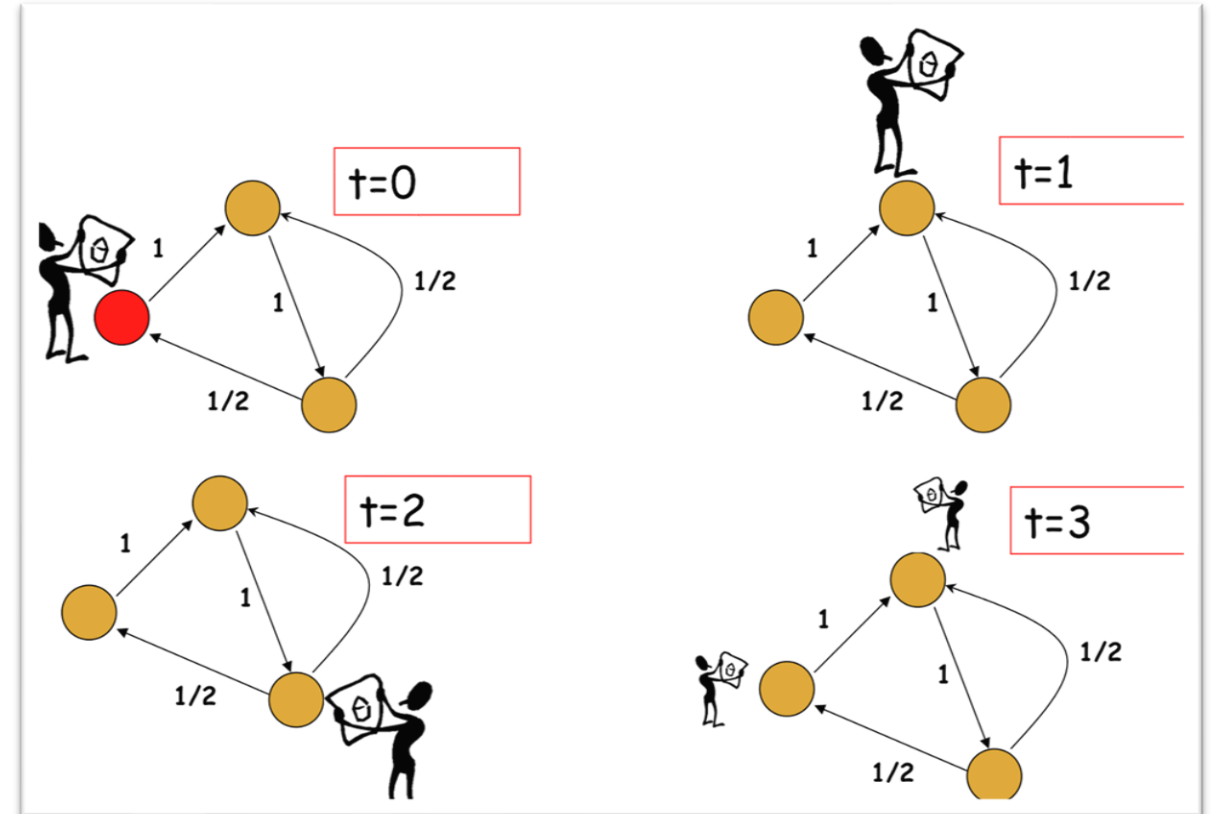
Graph Embeddings: Issues for IR

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!)**
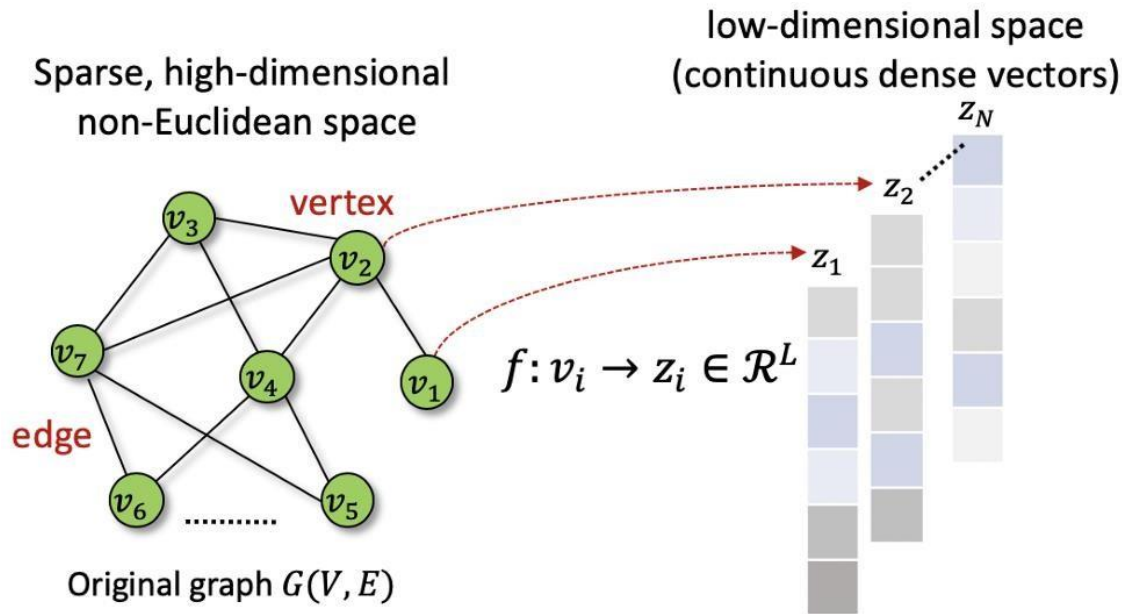
How About We Use BERT?

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

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.

❑ **Idea:** Wikipedia2Vec embeds words and entities into the same vector space → Alignment learnt using words can also be applied to entities.

How Well Does BERT Understand Entities?

MAP DBpedia-All - Pairwise
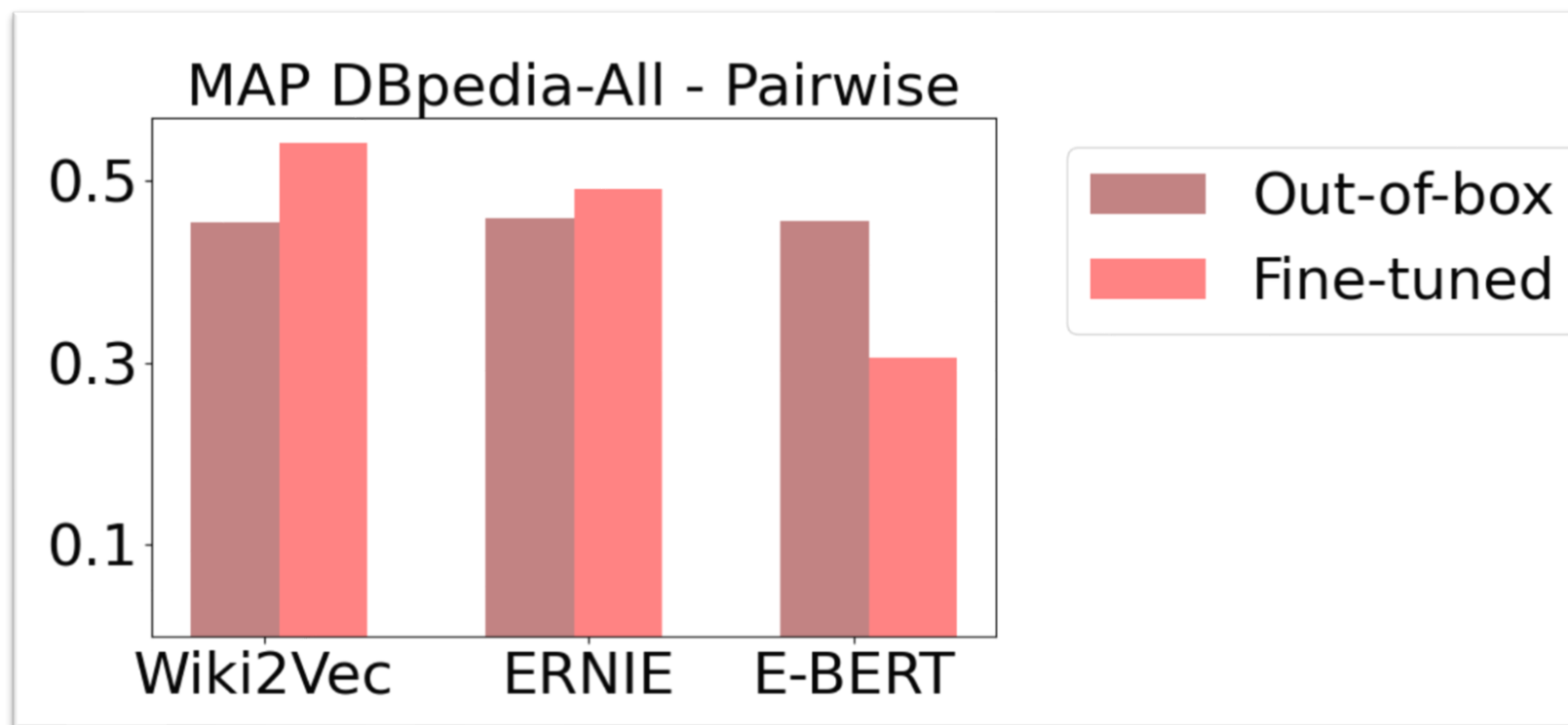
Results for GEEER with ERNIE and E-BERT. DBpedia-Entity v2.

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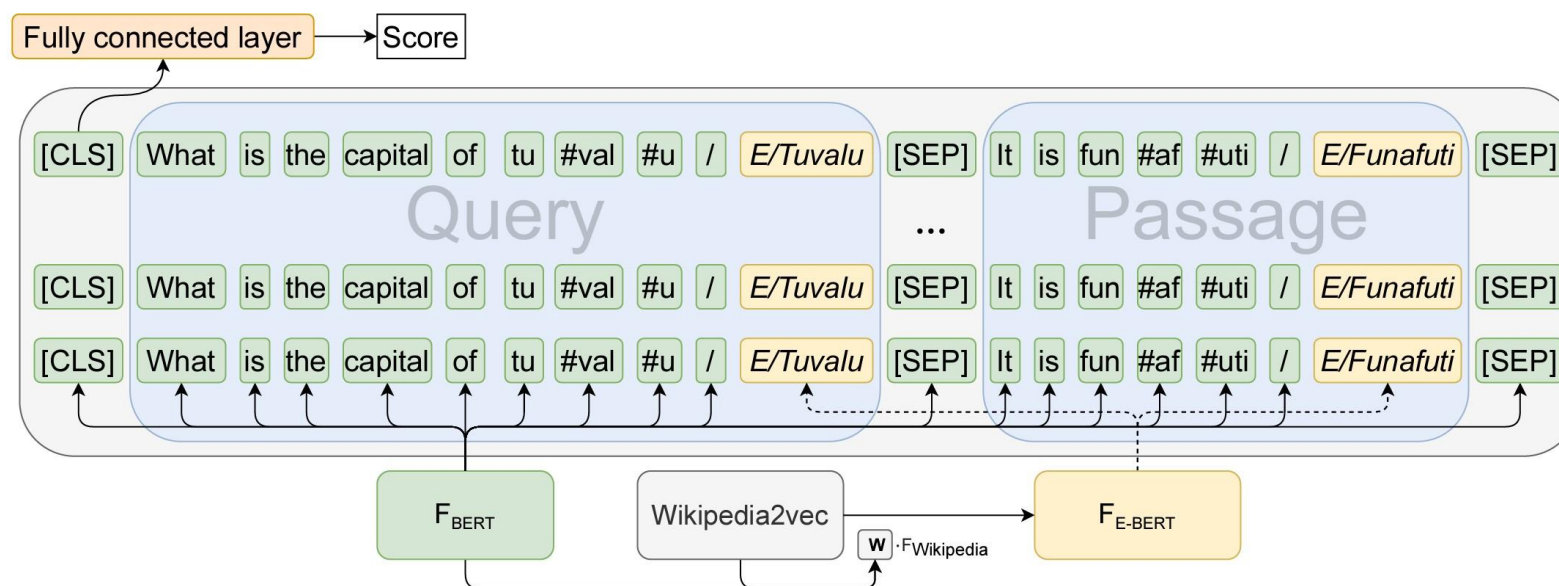
Picture Credit: *Gerritse et al. Entity-aware Transformers for Entity Search. 2022.*

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.

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.

······

BERT

1   2   3   ······   512

[CLS] FDA federal
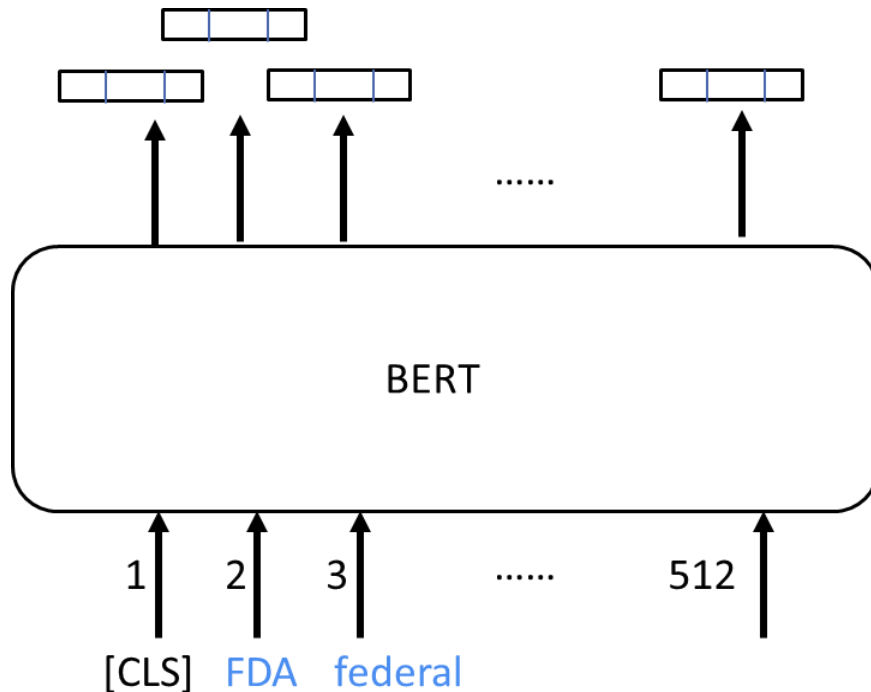
- Lead Text = Static entity description
- No knowledge of the query!
- Static entity embeddings.

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.
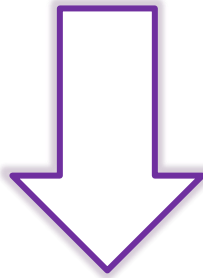
What if we had this text instead of lead text?

BERT-ER: Query-Specific Entity Descriptions

Query-specific entity descriptions→ descriptions that mention relevant connections between the query and the entity.
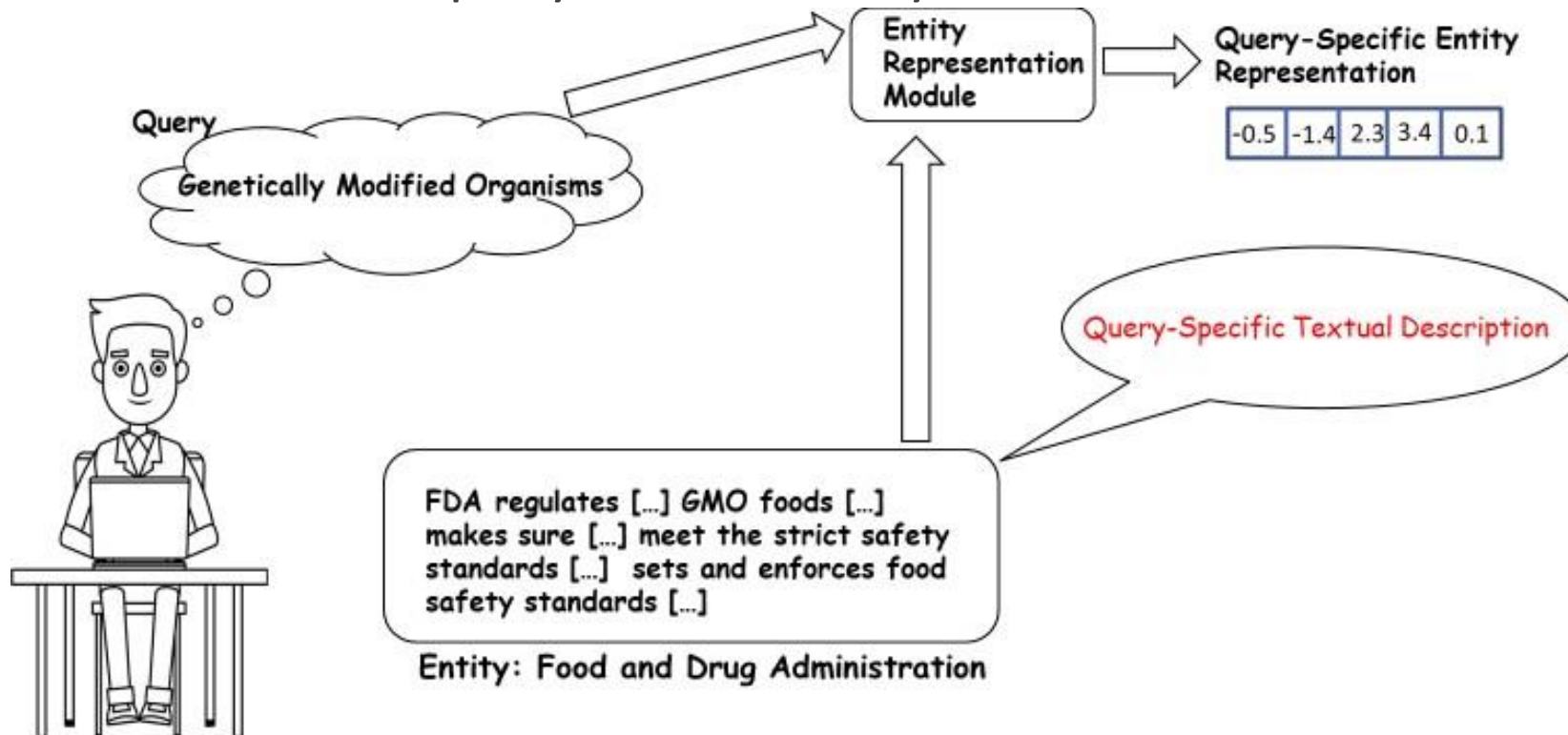
Shubham Chatterjee. Neuro-Symbolic Approaches for Information Retrieval. ECIR 2023

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

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**

Query: Genetically Modified Organism

Paragraph Index

FDA

Pseudo-Relevant Paragraphs
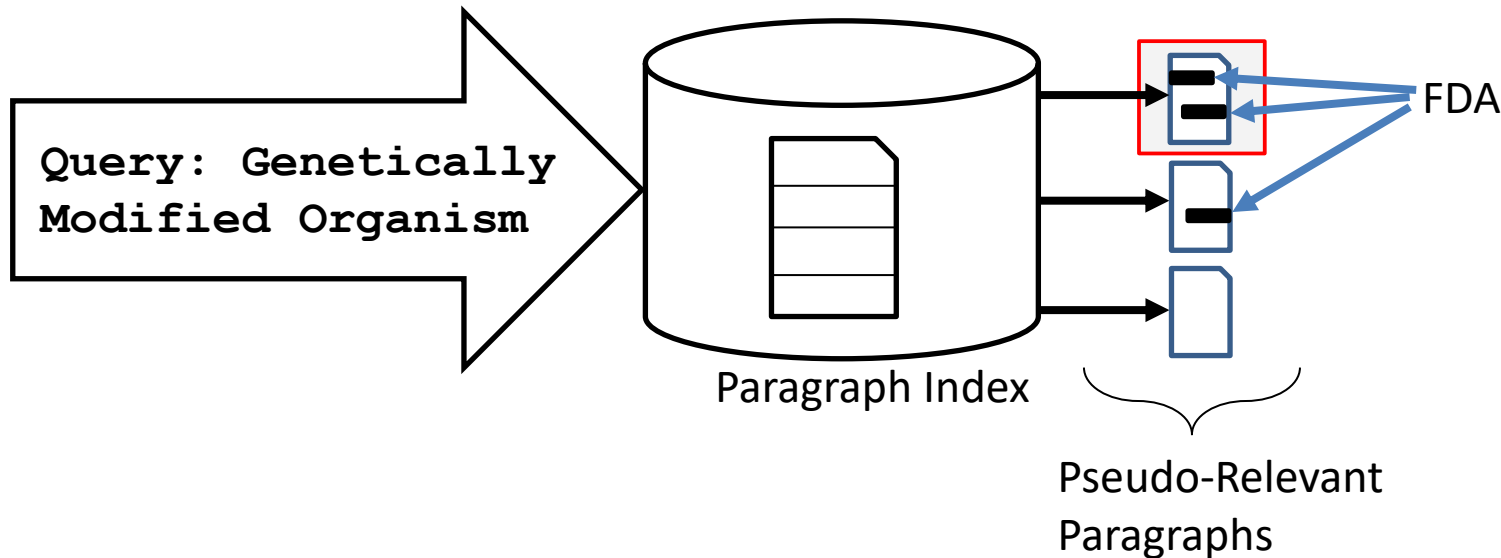
❑ **Downside**: *Entity may not be central to the discussion in the text!*

4/5/2023

Shubham Chatterjee. Neuro-Symbolic Approaches for Information Retrieval. ECIR 2023

31

Query-Specific Entity Descriptions: Alternative 3

Chatterjee and Dietz., 2022

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

Query: Genetically Modified Organism

Paragraph Index

FDA

Pseudo-Relevant Paragraphs
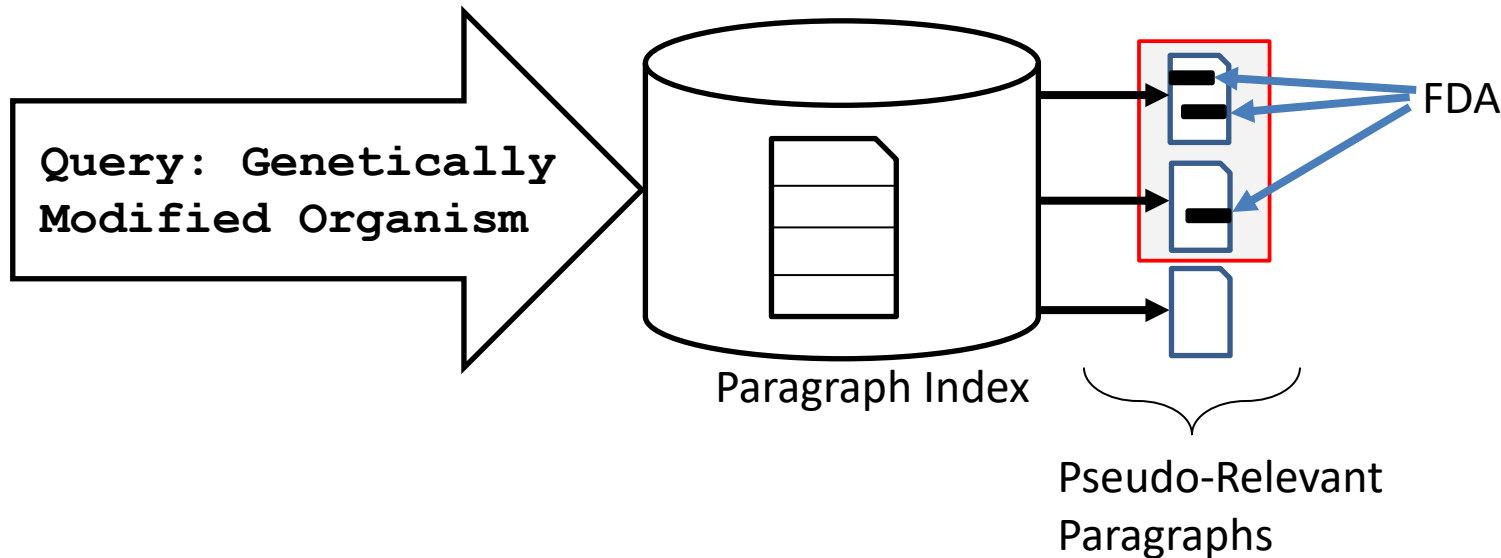
❑ Re-rank these documents.
❑ Two criteria:
1. How many relevant connections between query and entity?
2. Are the relevant connections central to the discussion in the text?

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

4/5/2023

Shubham Chatterjee. Neuro-Symbolic Approaches for Information Retrieval. ECIR 2023

32

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

Query: Genetically Modified Organism

Entity: Food and Drug Administration

Entity-Linked Corpus

Lead (Baseline)
The Food and Drug Administration (FDA) is a federal agency of the Department of Health and Human Services. ...

Aspect
Location
Regulatory programs
... The programs for safety regularion vary widely by the type of products, its potential risks, and the regulatory power ...

Lead

Aspect

Aspect

Aspect

PRF-Passage
A genetically modified organism (GMO) is [...] . **FDA** regulates most human and animal food, including GMO foods.[..]
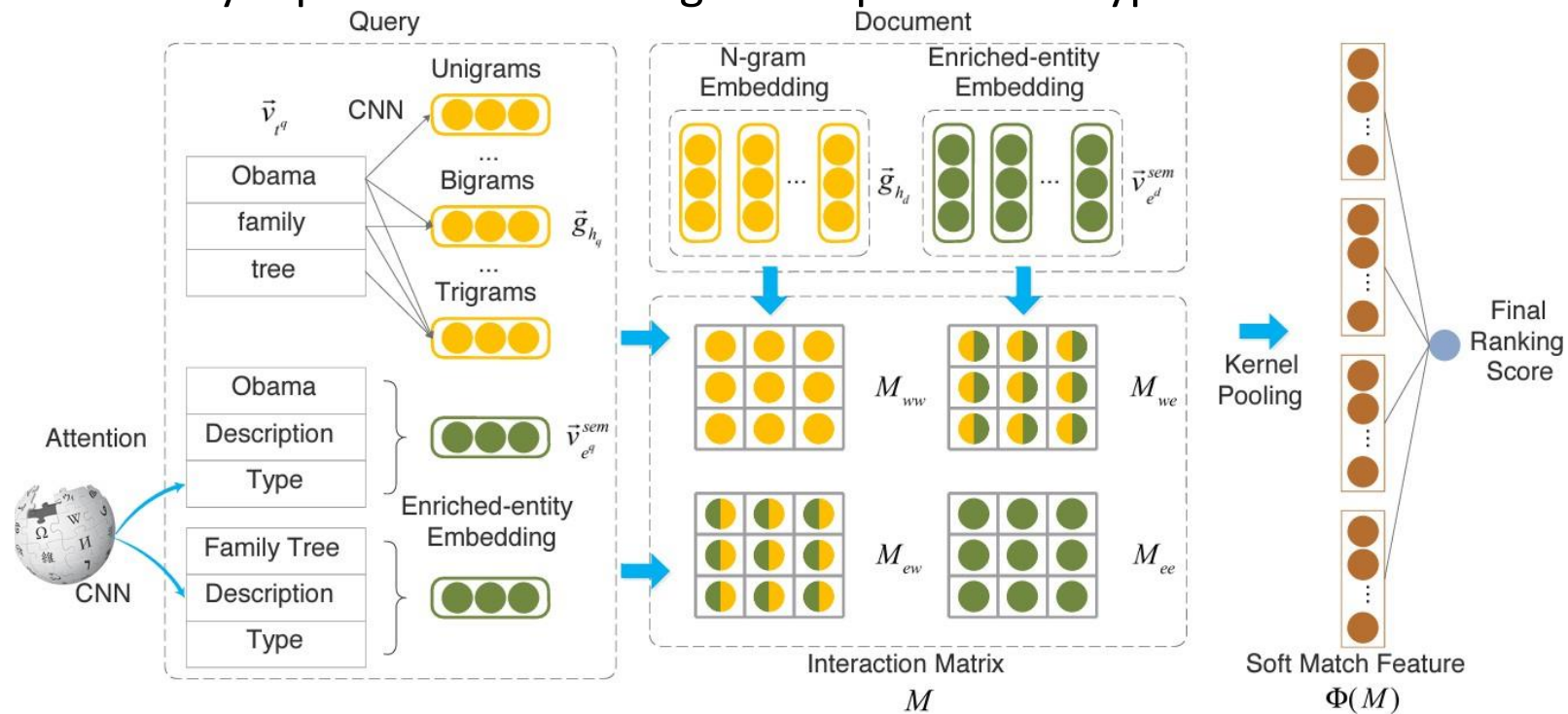
Entity-Support Passage
The U.S. **Food and Drug Administration** (FDA) ensures that GMOs are safe for human, plant, and animal health. ...

BERT

Query-specific Entity Representations

How Do We Use Entities?

Entity-Duet Neural Ranking Model

❑ 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

❑ Enrich query/document representation with entity representations.

❑ Cluster entities → Entity embeddings using clusters → Clusters act as "views" of the document.
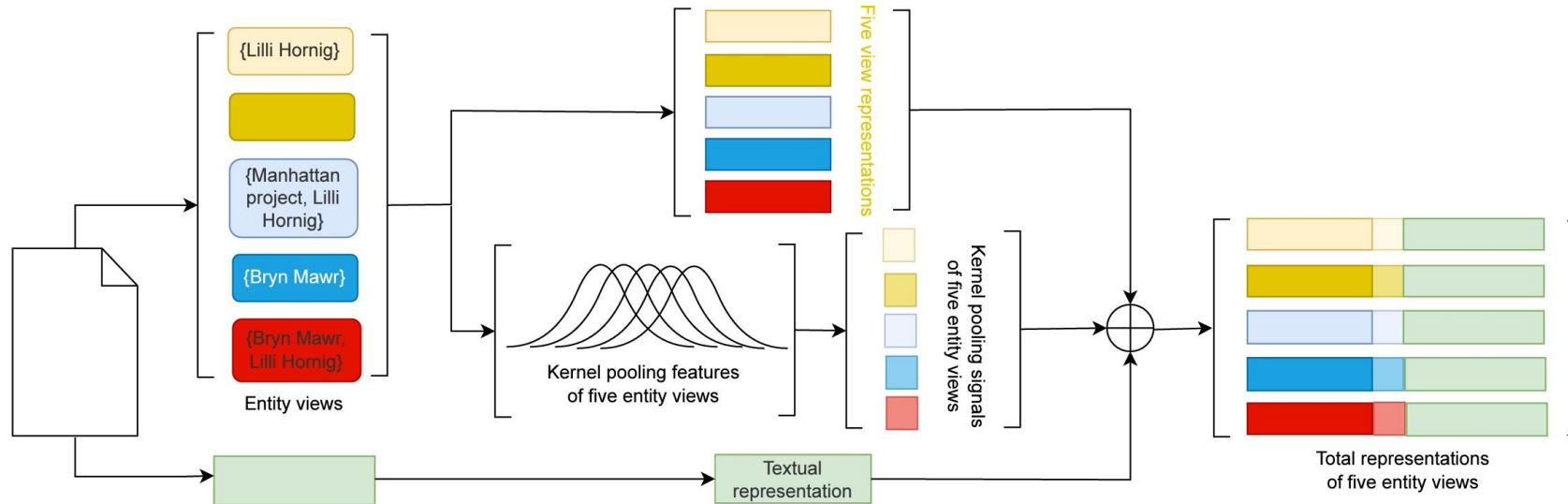

Figure 1: Overview of EVA with multiple representations. Entity clusters such as {Lilli Hornig}, {Manhattan project}, {Manhattan project, Lilli Hornig}, {Bryn Mawr} and {Bryn Mawr, Lilli Hornig} can be understood as different entity views of the passage. EVA generates one total representation for each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

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