Neuro-Symbolic Representations for IR
1.2 –Welcome and Purpose of this Tutorial

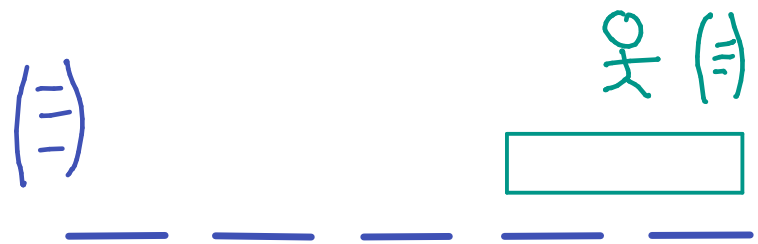
Laura Dietz

ECIR

2023

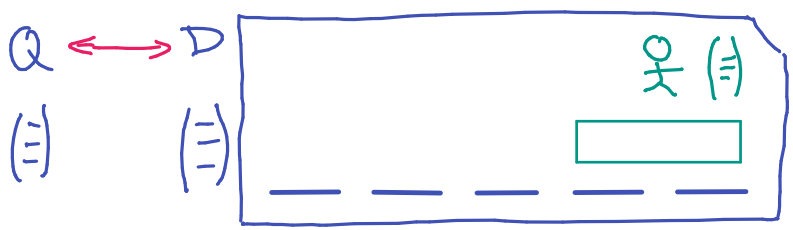
Neuro-Symbolic Representations...

- Combine strengths of neural networks and symbolic methods
- Neural: continuous representations to capture noisy concepts
- Symbol: capture discrete aspects of the task
- Symbols to help
  - resolve ambiguity in text, such as entity annotations
  - draw on structured external knowledge, such as knowledge graphs
  - derive rich entity representatinos and topically relevant connections
  - reason with compositional logic, such as First-order logic of predicates
- Goal: Combining different types of knowledge in a unified framework



.. Representations for Information Retrieval

- Symbolic entity representations can be incorporated into IR models
- Capture semantic and symbolic information associated with entities
- text amd symbolic entity representations can enhance:
  - Query rewriting: semantically related terms
  - Semantic matching: understand underlying meaning
  - Document ranking: what is more relevant
  - Knowledge graphs: structured background info
  - Complex information: connect pieces of relevant information



chatGPT criticism: Faithfulness

ChatGPT hallucinates facts. You don't know which are correct.

LA What is the usual time difference between London and Boston?

The usual time difference between London and Boston depends on the time of year, as both cities observe daylight saving time, but they begin and end at different dates.

During standard time:

- London is in GMT (Greenwich Mean Time)
- Boston is in EST (Eastern Standard Time)

  There is a 5-hour difference, with Boston being 5 hours behind London.

During daylight saving time:

- London is in BST (British Summer Time)
- Boston is in EDT (Eastern Daylight Time)

  There is a 4-hour difference, with Boston being 4 hours behind London.

Keep in mind that the dates for daylight saving time may change, so it's always good to double-check the current time difference between the two cities.

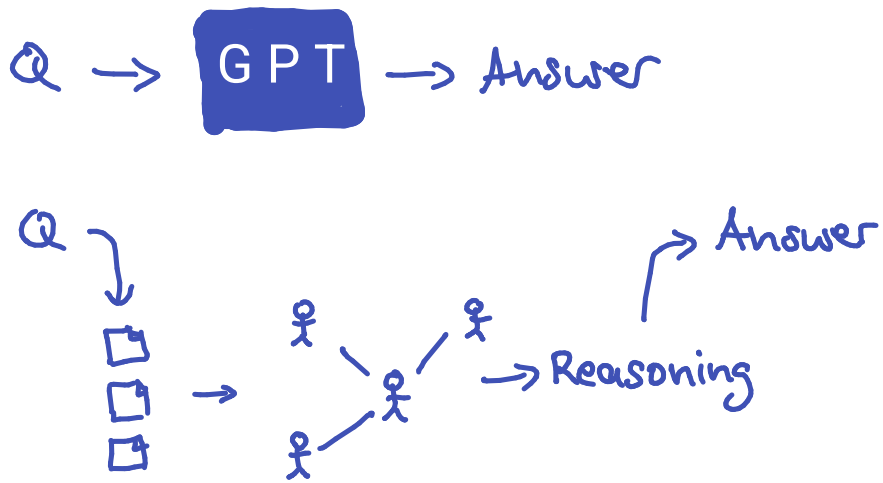
chatGPT criticism: Faithfulness

ChatGPT hallucinates facts. You don't know which are correct.

Approaches

- Retrieval-augmented Generation: Retrieve provenance first
- Cross-check with complementary information sources
- Verify with manually-checked Knowledge Graph
- Ask for its Chain-of-thoughts, to check each.

Verification: break down information into hard facts
→symbols



Current approach: More training data (Lambda)

Mix Generative and Discriminative LLM

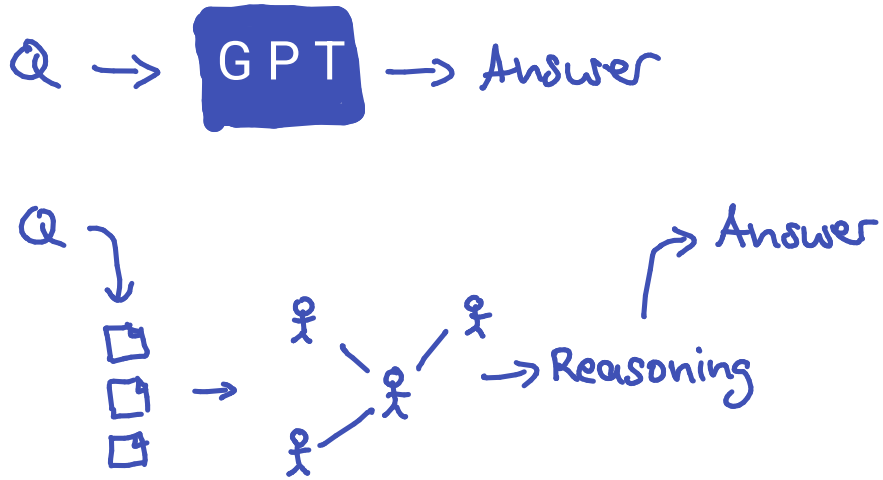
- Key goals: quality, safety, and groundedness
- Grounded in an external (verified) knowledge resource

1. Ask crowd workers for information seeking queries
2. Generate search query for retrieval
3. Generate answer from retrieved documents
4. Assessor judges how close response follows the source

Thoppilan 2022. "Lamda: Language models for dialog applications"

Smarter Approach?

Maybe we have ideas for a smarter approach?



.oO(Maybe we figure it out today?)

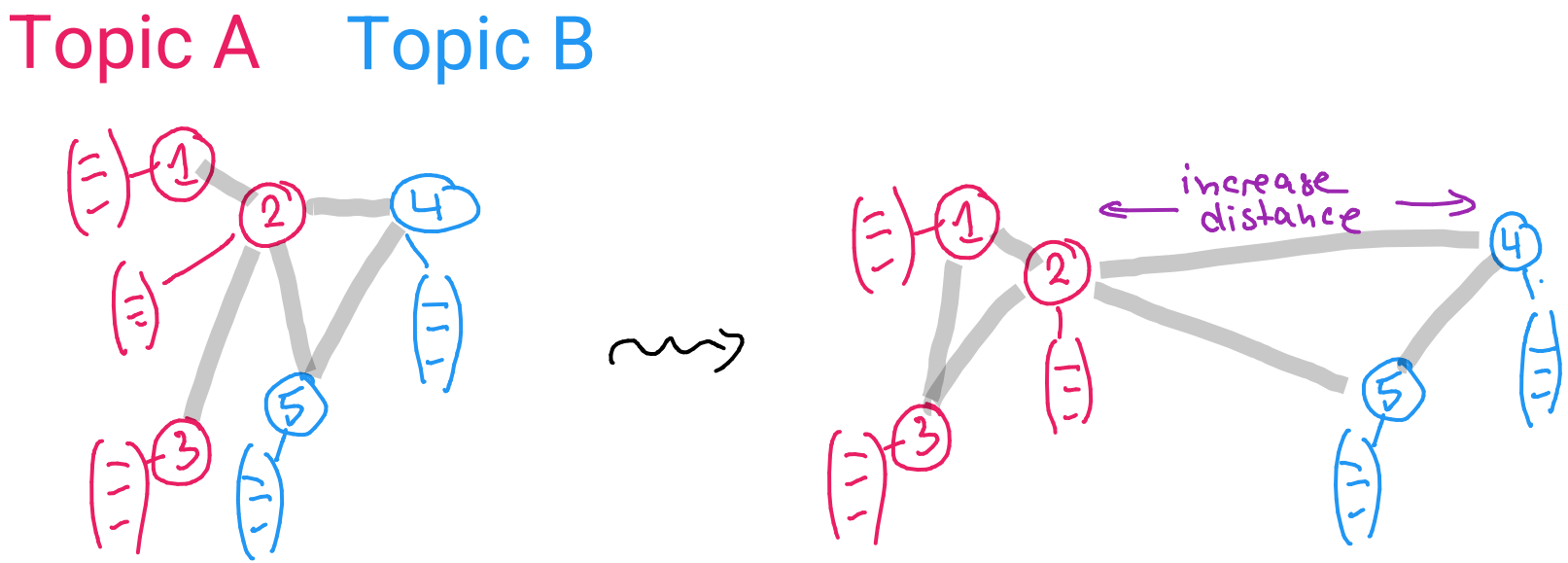
Outline

1. Neural Representation with LLMs
2. Bi-encoders
3. Cross-Encoders
4. Why only Neuro—and not Symbolic?

Metric Learning

Task:

- Given data points (vectors), with true topic labels
- Project into vector space where points with
  - same labels are close
  - different labels are far apart

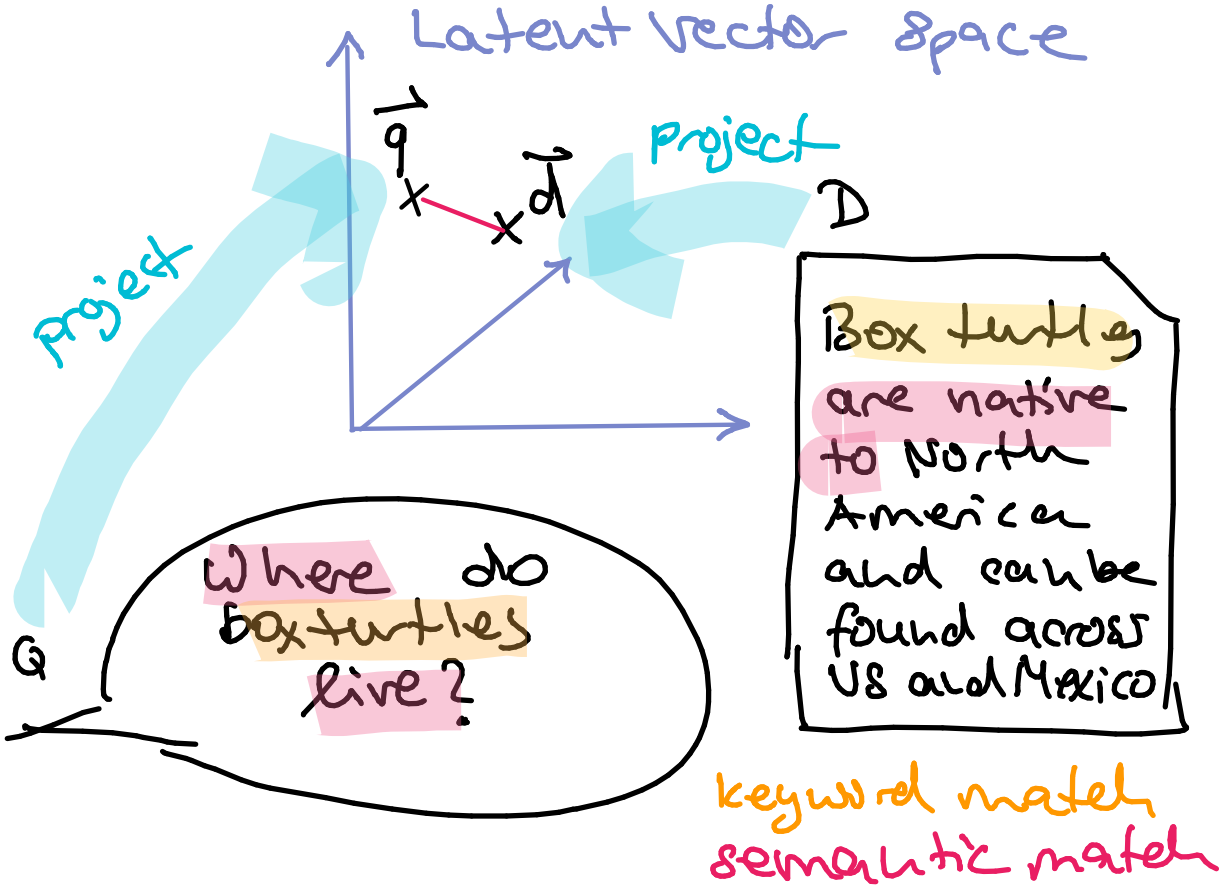


Train: latent space has "useful" representation of proximity.

→ Can use Standard clustering algorithm to obtain topical clusters.

SOTA: ad hoc Text Ranking

Dense Retrievel Approach:

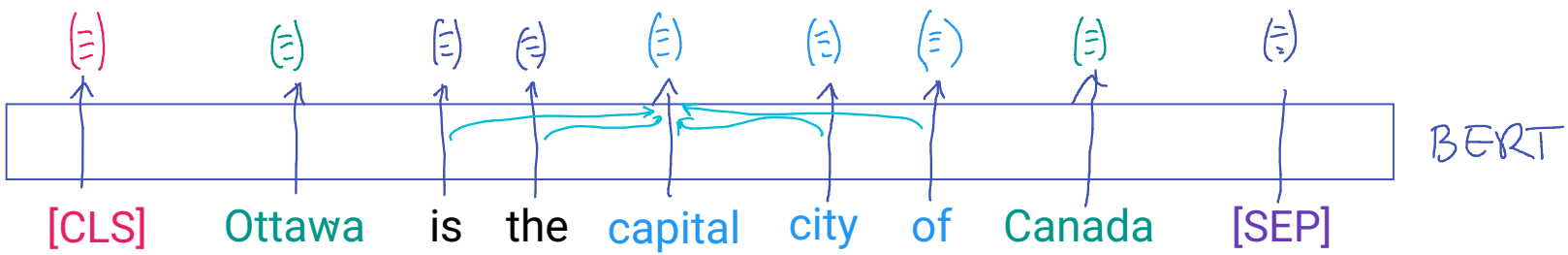


Latent space and projections are trained, so that **q** and **d** are close whenever documents are relevant for the query.

Text Representation with Large Language Models (LLM)

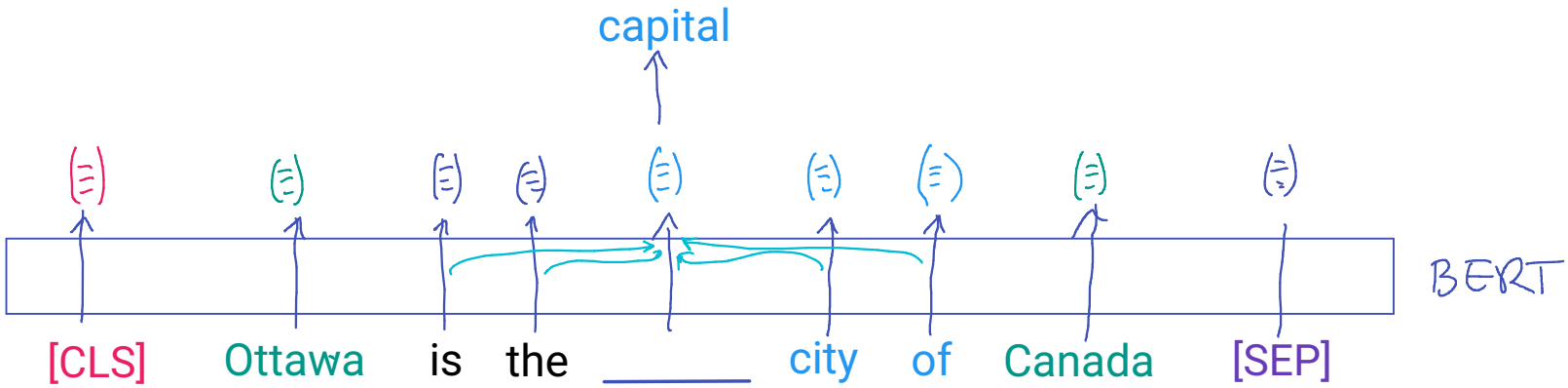
Large Language Models (BERT, T5, GPT, Word2Vec)

- Capture rich contextual information
- Can be fine-tuned for tasks such as entity linking, recognition, or relation extraction
- Represent each word as a vector



Each word vector is aware of neighboring words!
Also: special `[CLS]` and `[SEP]` tokens have vectors

Trained with CLOZE "fill the blanks"

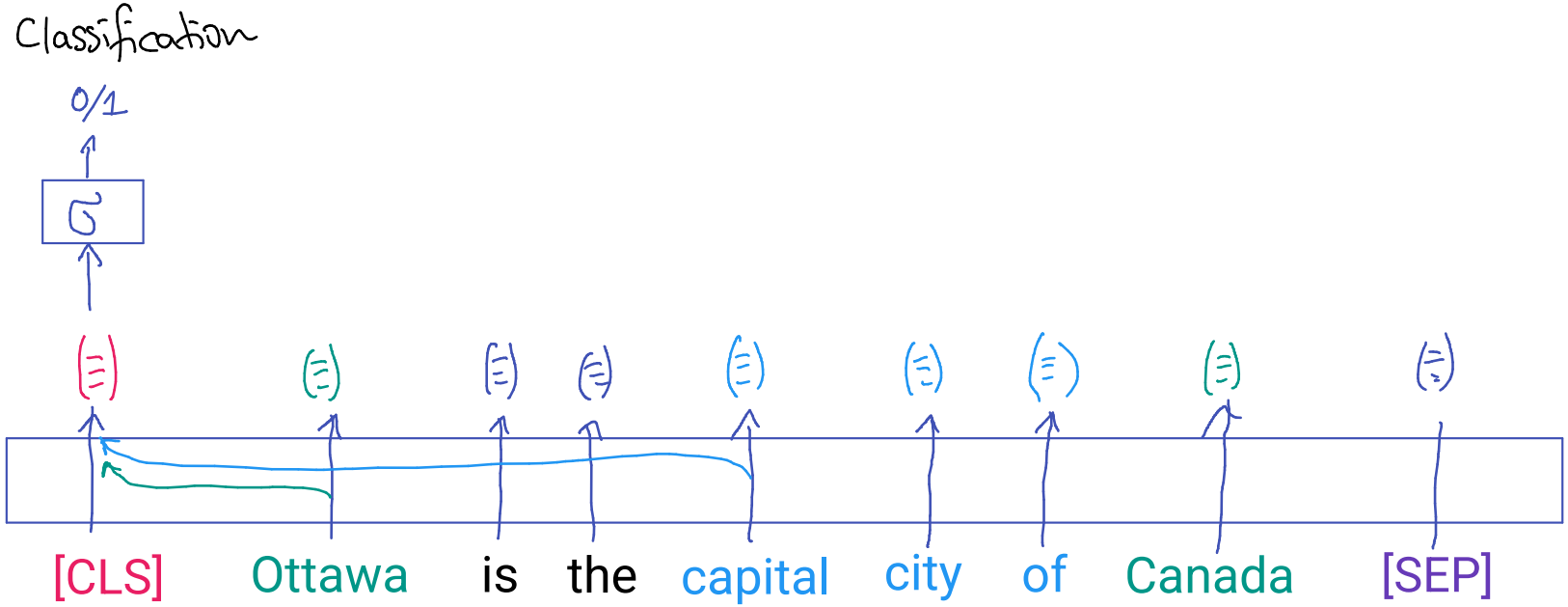


Each word vector is aware of neighboring words!

...more on that later!

LLMs for Classification

- Connect ground truth to vector of the [CLS] token
- Connect to a logistic/softmax layer to predict yes/no



The [CLS] token is also informed by neighboring words.

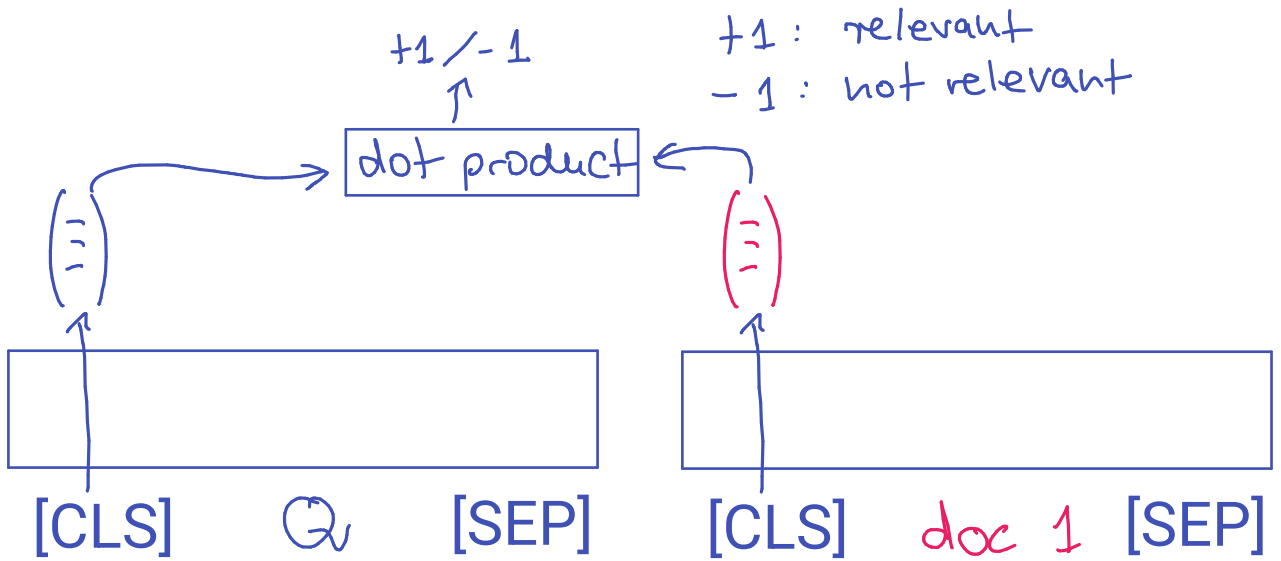
Outline

1. Neural Representation with LLMs
2. Bi-encoders
3. Cross-Encoders
4. Why only Neuro—and not Symbolic?

Dense Retrieval: Bi-encoders

≈ Dense Passage Retriever (DPR) ≈ "Late Interaction" ColBERT

- Bi-Encoder model
  - A query encoder, which encodes the query **q**
  - A document encoder, which encodes text **d**
- Rank score as dot product: $score = \mathbf{q} \cdot \mathbf{d}$

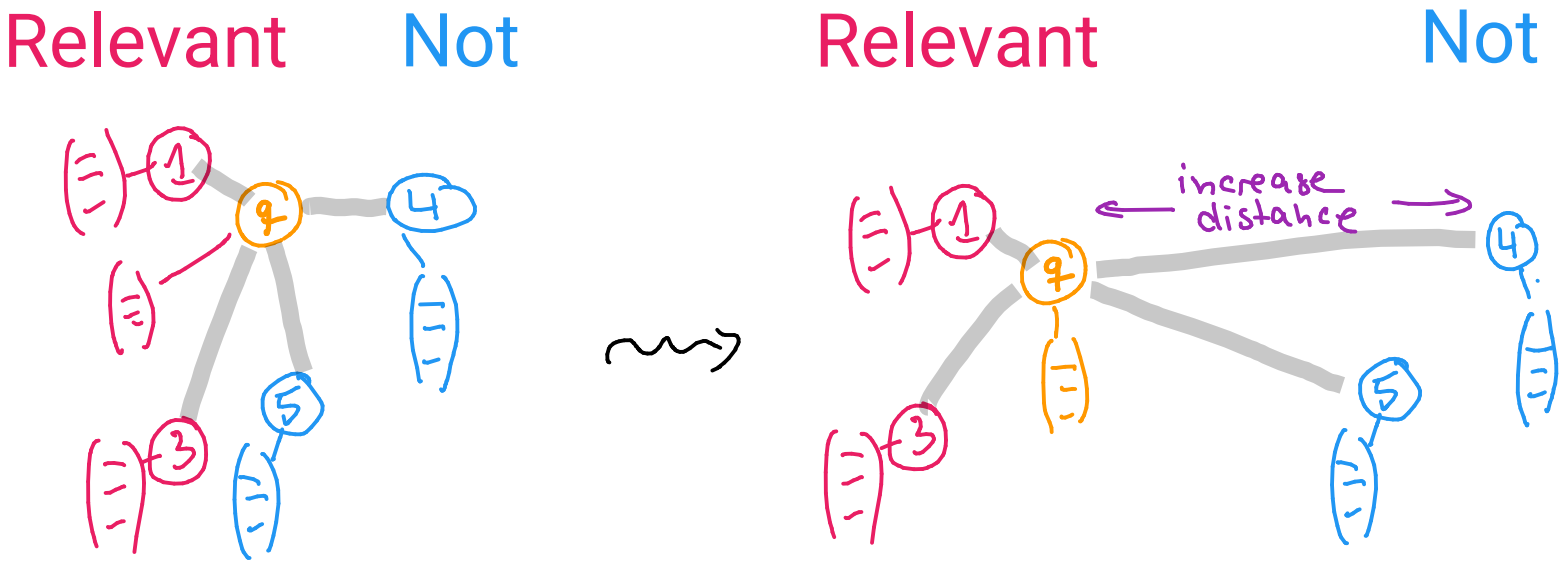


Q: Query do box turtles live

Ranking as Metric Learning

Task:

- Given query and document points, with relevance labels
- Project into vector space where
  - relevant points close to query
  - not relevant points are far away

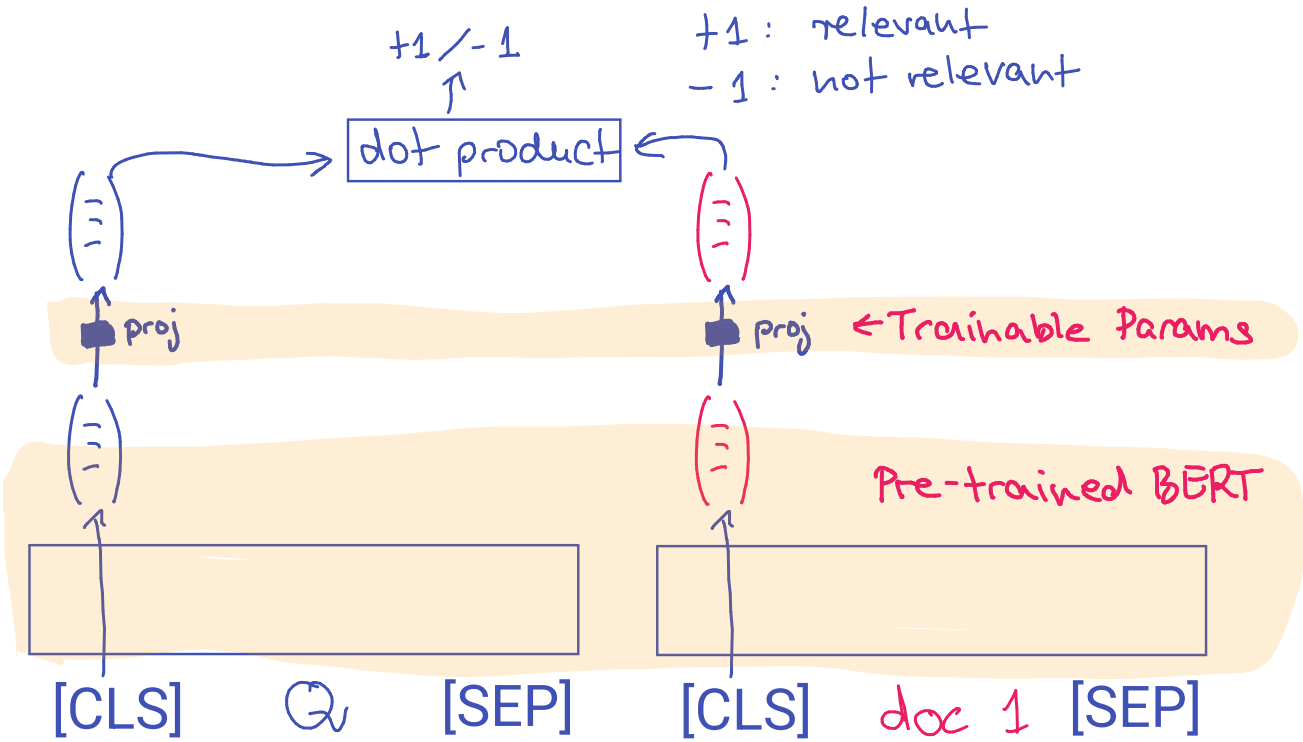


Train: latent space trained so that proximity = relevance.

→ Can use Standard K-NN for ranking

Fine-tuning Bi-encoders

- Projection layer for training (on fixed pre-trained LLM)
- (Non-) Linear projection with trainable parameters
- Alternative: Single trainable Transformer layer



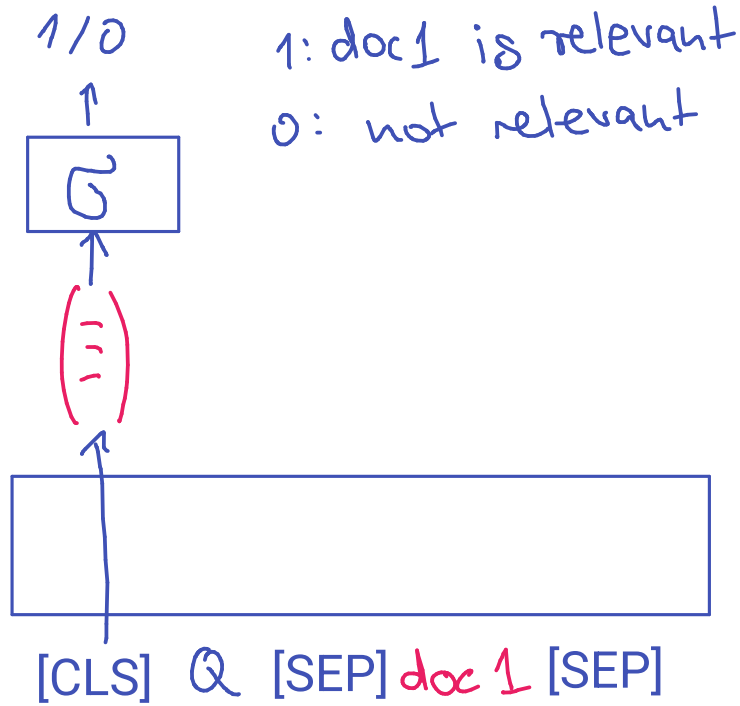
Outline

1. Neural Representation with LLMs
2. Bi-encoders
3. Cross-Encoders
4. Why only Neuro—and not Symbolic?

Ranking: Mono-BERT

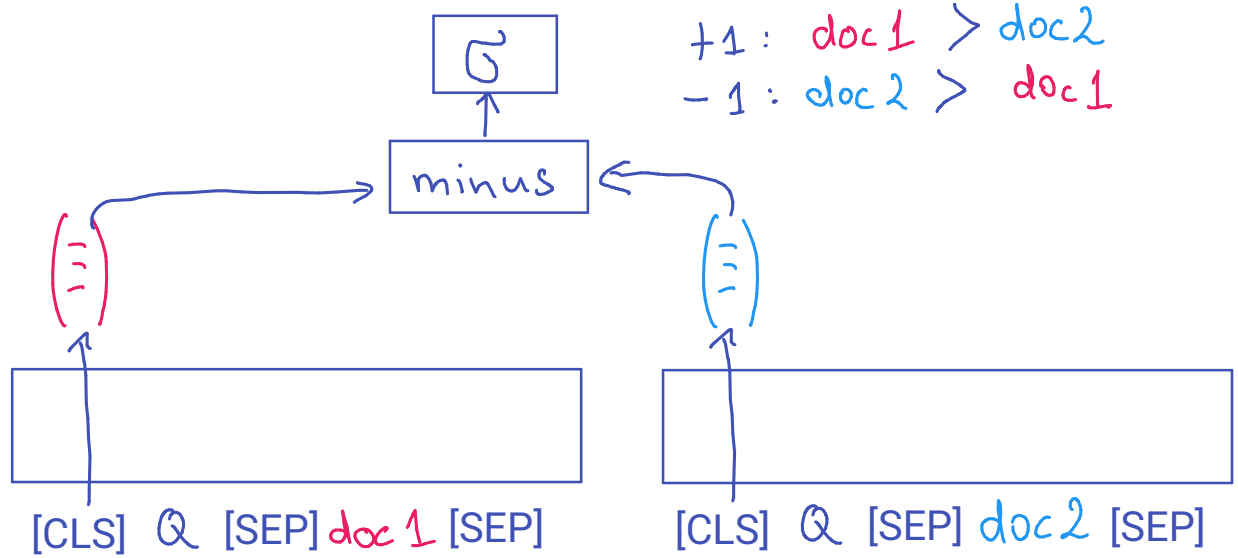
Task:

- Given query - document pair
- Classify [CLS] into relevant / not
- Loss function: Binary Cross-entropy



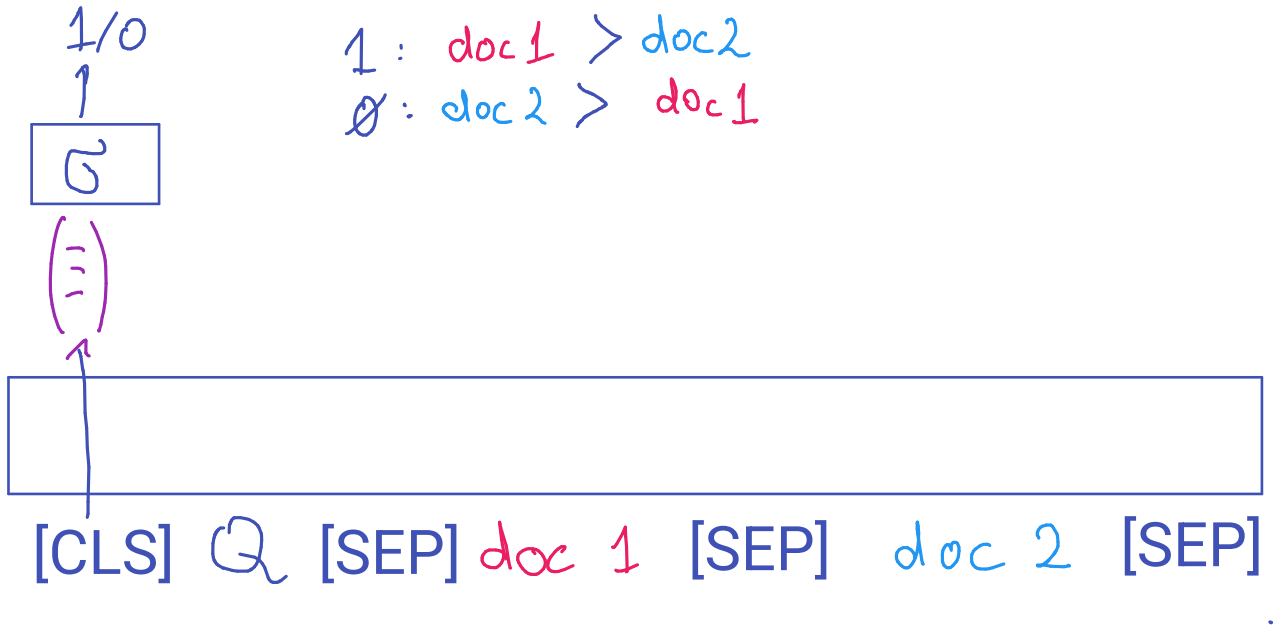
Mono-BERT optimized with Pair-wise Ranking Loss

- Classify which document is more relevant for query.
  doc 1 or doc 2?



Duo-BERT: Classify Query and Document Pair

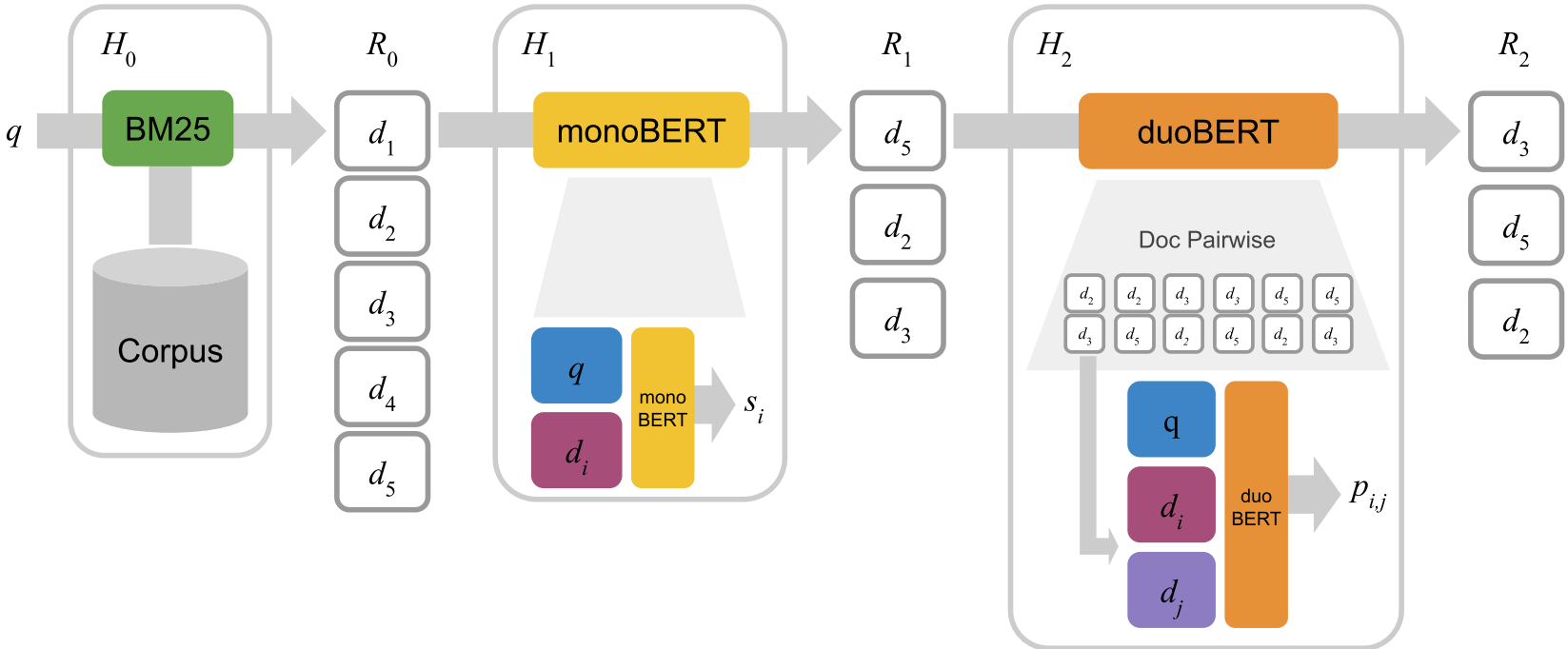
- Classify which document is more relevant for query. doc 1 or doc 2?



Downsides of Cross-encoders:

- Vectors can't be indexed.
- Slower model at query-time.

Mono-Duo BERT for Web Search



Multi-stage candidate sets for speed/accuracy tradeoffs.

DuoBERT implementation Nogueira et al. "Multi-stage document ranking with BERT"

Outline

1. Neural Representation with LLMs
2. Bi-encoders
3. Cross-Encoders
4. Why only Neuro—and not Symbolic?

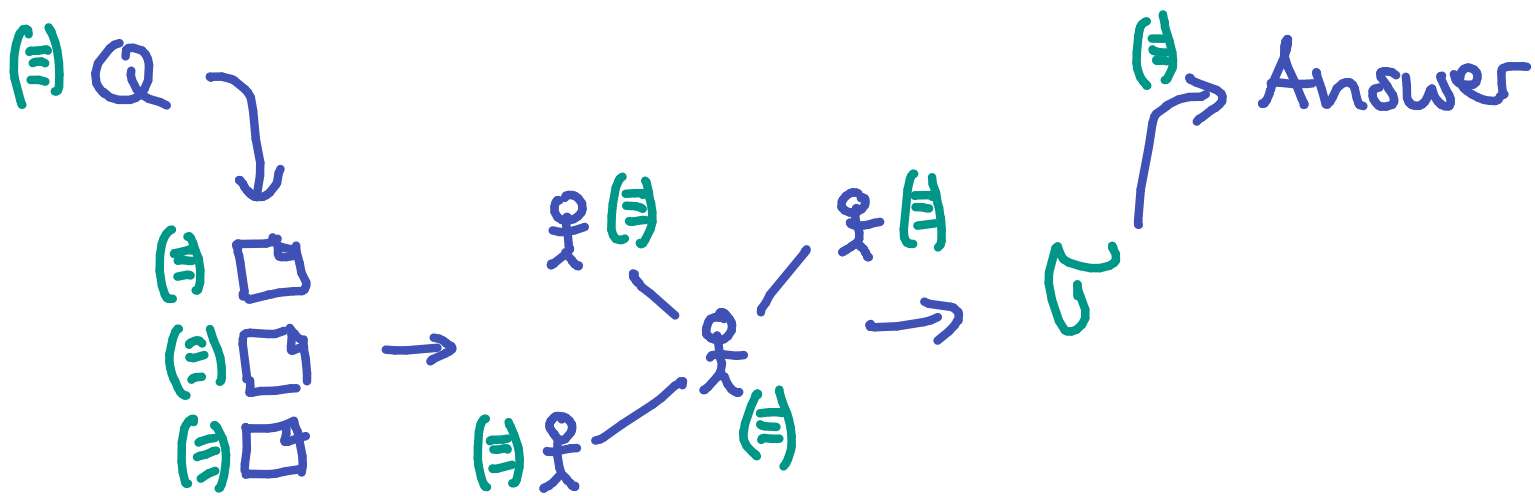
Why not symbolic?

Current state-of-the-art in ad hoc text retrieval
"Only" neural text representations

Why not neuro-symbolic?

- Is it not helping?
- Is it too complicated to implement?
- Did we not yet figure out how to reap benefits?

.o0(Maybe we figure it out today!)



BERT struggles to understand Entities

Common thinking: BERT can do everything!

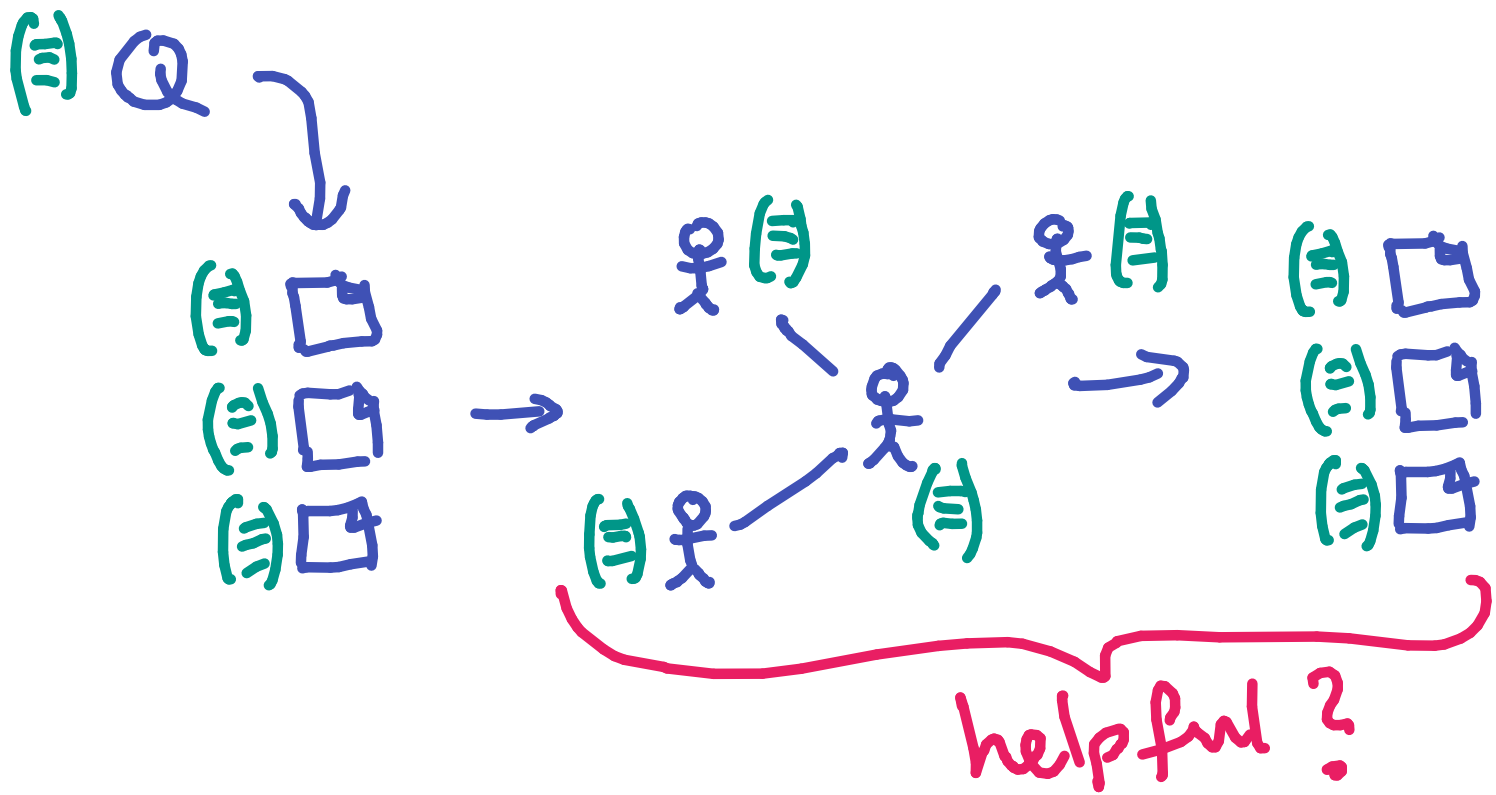
– including understanding entities

But: Empirical evidence suggesting that BERT (Transformer-based) embeddings struggle to represent entities. Otherwise we would not need to put as much work in Entity Ranking.

... more on that later.

Do Entities help?

- In 2014–2019: Entity L2R features improve text ranking
- In 2019–2023: Entity features are not helping neural rankers
- 2022: Significant improvements on LLM-based entity ranking
- Today: Need to reconsider ad hoc text ranking with new "prescription-strength" entity representations



Conversational Search & Assistance

To drive a conversation,

- need to track important concepts (=entities) in the conversation
- reason which concepts are relevantly related
- integrate textual provenance

Prime application for Neuro-Symbolic IR!
...more on that later.

This tutorial

- Overview of progress made on Neuro-Symbolic Representations
- Cover entities, KGs, graph approaches
- Spectrum between purely neural and purely symbolic

Goals:

- Trigger new ideas for new papers
- Spur exciting discussions
- "Stick our heads together" to figure out better approaches

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

Tutorial Timetable

Part 1: Symbolic AI representations and tasks

- Welcome/Purpose of this tutorial ←You Are Here
- (Sub)symbolic AI, and representations
- Question Answering on Knowledge Graphs

Part 2: Text-to-symbols and Ranking

- Neural Text Representations
- Text-Symbol Alignment and Semantic Annotations
- Entity Representations and Entity Ranking

Part 3: Neuro-symbolic representations for Reasoning

- Reasoning about Relevance
- Neuro Pseudo-Relevance Feedback with Explainability

Part 4: Applications for Neuro-symbolic approaches

- Use Case: Knowledge Discovery
- Use Case: Task-based Assistance
- Use Case: Generating relevant (long-form) Articles
- Panel & Discussion

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