Neuro-Symbolic Representations for IR 3.3 – From PRF to Retrieval Enhanced Generation

Laura Dietz

SIGIR

2023

Tutorial Timetable

Part 1: Knowledge Graphs and Entities

- 1.1 Welcome & Motivation
- 1.2 Knowledge Graphs and GPT
- 1.3 Entity Linking

Part 2: Neuro-Symbolic Foundations

- 2.1 Ranking Wikipedia Entities / Aspects
- 2.2 Neural Text Representations and Semantic Annotations
- 2.3 Infusion of Symbolic Knowledge into Text Representation

Part 3: Reasoning, Robustness, and Relevance

- 3.1 Denoising Dense Representations with Symbols
- 3.2 Reasoning about Relevance
- 3.3 From PRF to Retrieval Enhanced Generation ←

Part 4: Emerging Topics

- 4.1 Conclusion and Outlook
- 4.2 Panel Discussion

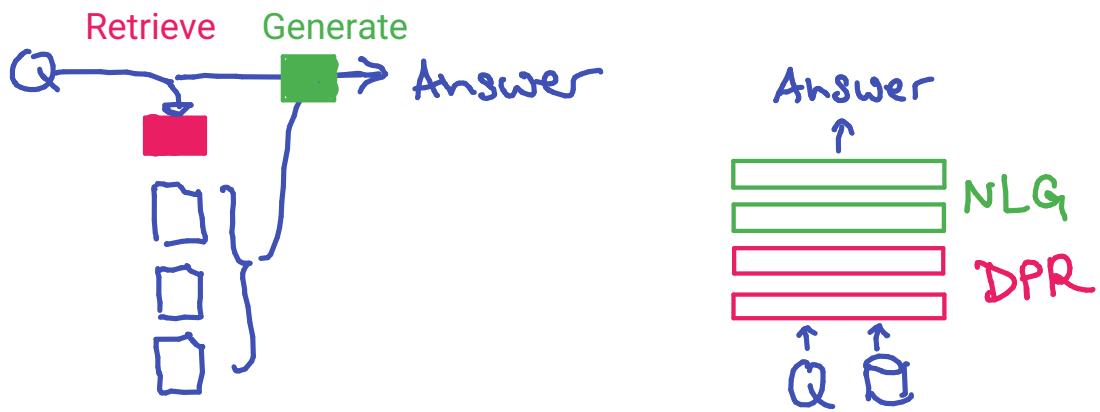
https://github.com/laura-dietz/neurosymbolic-representations-for-IR/SIGIR23/

Retrieval-Augmented Generation Models

- Combines retrieval and generation to produce relevant summaries or responses to a given query with inter-dependent steps
- The typical retrieval-augmented generation model would:

  Retrieve top-k passages with Dense Retrieval Model

  Generate natural language answer from retrieval results
- Optionally: add few-shot learning



Both components are trained end-to-end.

Lewis, 2020. "Retrieval-augmented generation for knowledge-intensive nlp tasks."

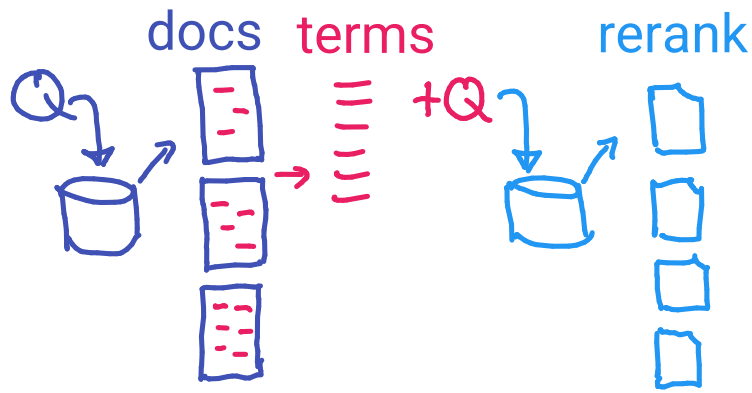
Reminds me of Pseudo-Relevance Feedback / RM3

Standard technique for Query Expansion Proceeds in three phases:

Given query,

- **Retrieve** documents, pretend they are relevant
- **Analyze** documents for frequently associated terms
- **Exploit** frequent terms to expand query (or to re-rank)

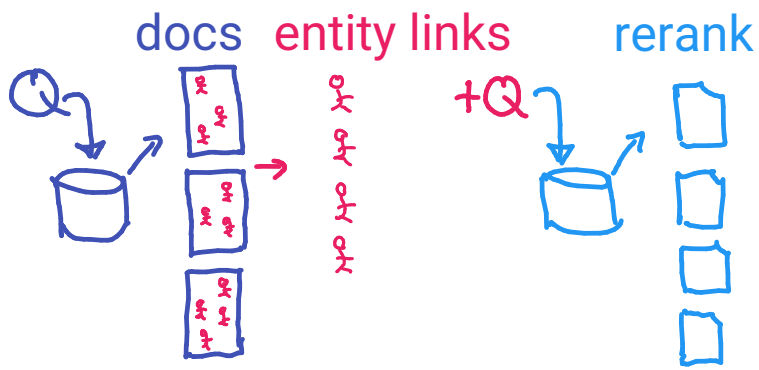
This part: closer look at these three phases in connection to Neuro-Symbolic approaches



Symbolify: Entity Link PRF

Best performing feature is "entity context model" Given query,

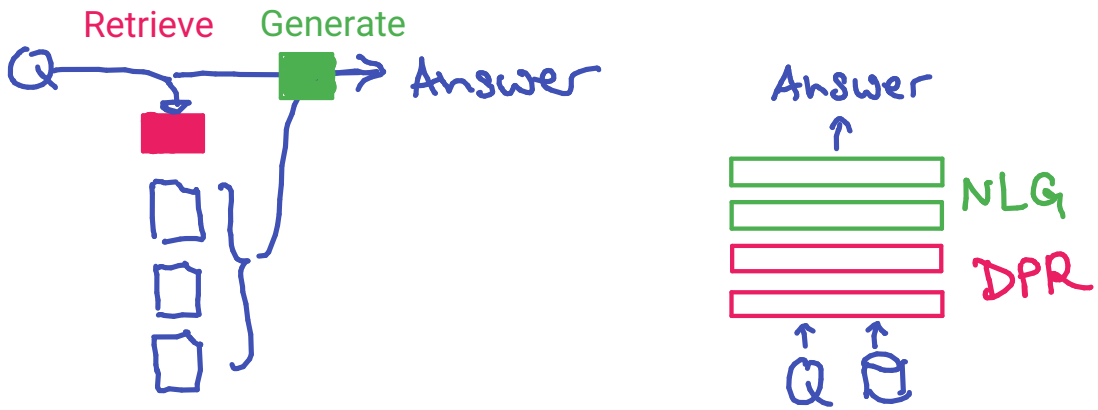
- **Retrieve** passages, which are annotated with entity links
- **Analyze** passages for frequently mentioned entities
- **Exploit** frequent entities, prefer documents which link to them



Exploit that symbols are less ambiguous and more meaningful

Dalton 2014, "Entity Query Feature Expansion "

Retrieval-augmented Generation



Given query,

- **Retrieve** passages (dense retrieval)
- **Analyze** —
- **Exploit** generate answers from passages

- Both components are trained end-to-end.
- Backpropagation will train retrieval model implicitly.
- No explicit retrieval benchmark needed!

Zamani, 2022. "Retrieval-Enhanced Machine Learning."

Comparison

	RAG	RM3	Entity Link PRF
Retrieve	trained end-to-end	heuristic (BM25)	heuristic (BM25)
Analyze		heuristic (RM)	heuristic (RM)
Exploit	trained	heuristic	trained (L2R)

Outline

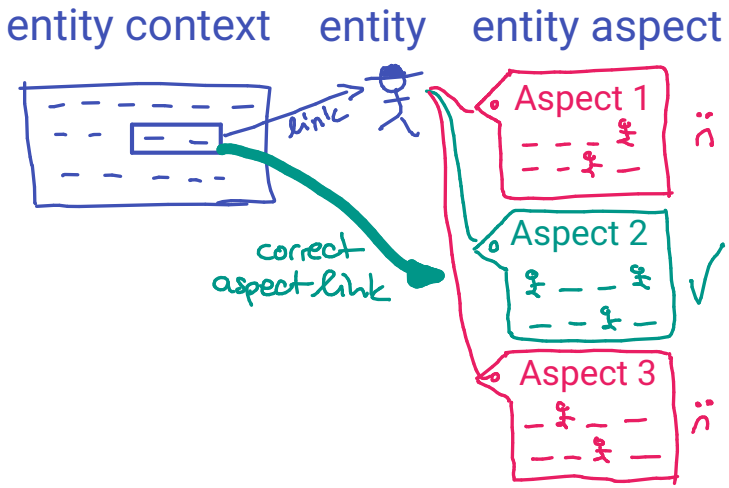
1. Guiding Entities
2. Coordinated Benchmarks
3. Exploiting Explainability in PRF
4. Retrieval-augmented Generation
5. Conclusion

Example Task: Entity Aspect Linking

Task:

- Given context passage with entity link
- given catalog of different aspect for this entity
- predict the most relevant aspect for the context

Ground truth harvested from hyperlinks to a Wikipedia section.



[Oysters] influence ecosystems through nutrient cycling

- Anatomy
- Ecosystem Services
- As Food

Aspect catalog = sections of entity's Wiki article

Each section content has entity links!

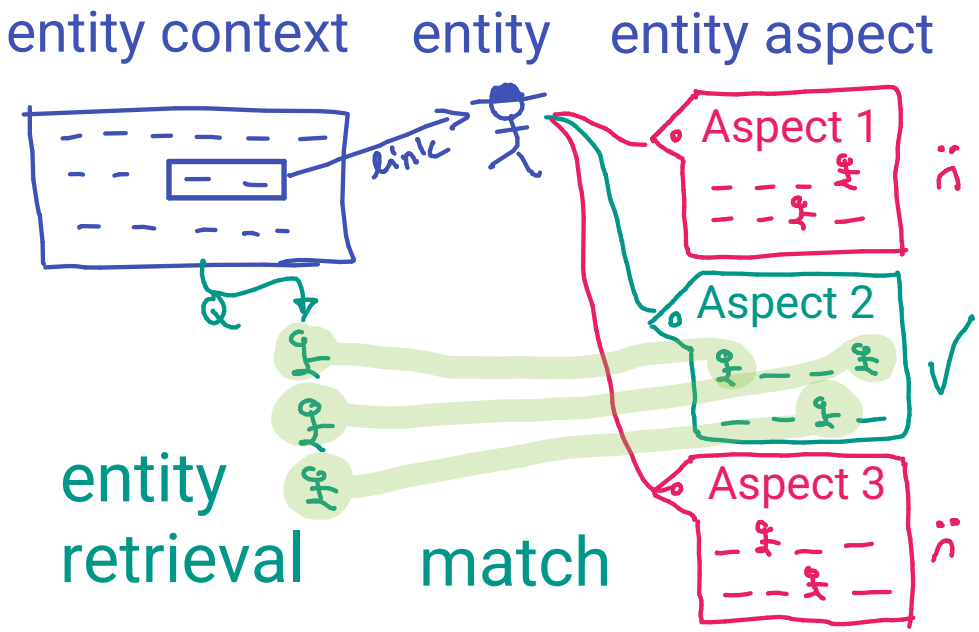
Nanni 2018. "Entity-aspect linking" - Ramsdell 2020. "A large test collection for entity asp

Entity Guides for Semantic Annotations

Given context (=query),

- **Retrieve** passages (heuristic)
- **Analyze** passages to predict relevant entities (trained)
- **Exploit** entities for semantic annotation task (trained)

Entity Ranking and Matching task are both trained individually, but on a **coordinated benchmark**.

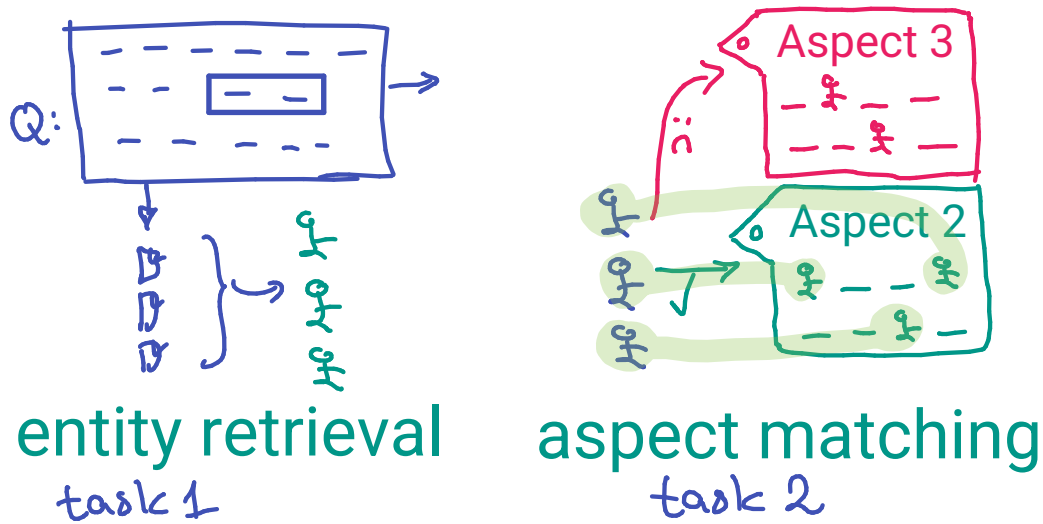


Chatterjee 2022, "Predicting Guiding Entities for Entity Aspect Linking."

Coordinated Benchmarks for Joint Tasks

1. Task: Entity Retrieval
Query $\rightarrow$ Entity that is mentioned in the Section's content

2. Task: Aspect Matching
Query, Entities $\rightarrow$ Aspect (= Wikipedia Section)



Predicting with Entity Guides for Entity Aspect Linking

Prediction:

Given context passage with entity link

- **Retrieve** passages
- **Analyze** passages to rank entities (coordinated training)
- **Exploit** top entities to find right aspect (trained)

	Entity Ranking			Derived Aspect Ranking			LTR (Derived + Lexical)		
Entity Ranking Method	MAP@100	P@R	NDCG@100	P@1	MAP	NDCG@20	P@1	MAP	NDCG@20
BERT (PRF-Psg)*	**0.51***	**0.51***	**0.51***	**0.78***	**0.83***	**0.84***	**0.89***	**0.94***	**0.95***
BERT (LeadText)	0.33▾	0.31▾	0.57▴	0.35▾	0.56▾	0.66▾	0.64▾	0.78▾	0.83▾
Relatedness (Wikipedia2Vec [51])	0.30▾	0.28▾	0.53▴	0.35▾	0.55▾	0.65▾	0.64▾	0.78▾	0.83▾
Relatedness (E-BERT [37])	0.30▾	0.28▾	0.53▴	0.35▾	0.55▾	0.65▾	0.64▾	0.78▾	0.83▾

Chatterjee 2022, "Predicting Guiding Entities for Entity Aspect Linking."

Outline

1. Guiding Entities
2. Coordinated Benchmarks
3. Exploiting Explainability in PRF
4. Retrieval-augmented Generation
5. Conclusion

Definition Coordinated Benchmark

- Multiple relevance-oriented tasks
- Same set of queries
- Benchmarks agree with each other across tasks

Wide range of relevance-oriented tasks obtained via Wikimarks:

Passage Retrieval
Entity Retrieval
Search result clustering
Query-specific Entity Linking
Context-based Entity-Aspect linking
Outline Prediction
Information Ordering
Query-focused Summarization

Build a system by integrating sub-task components!

Article Generation

Task:

- Given query/topic
- generate a long-form article that informs about various aspects of the query

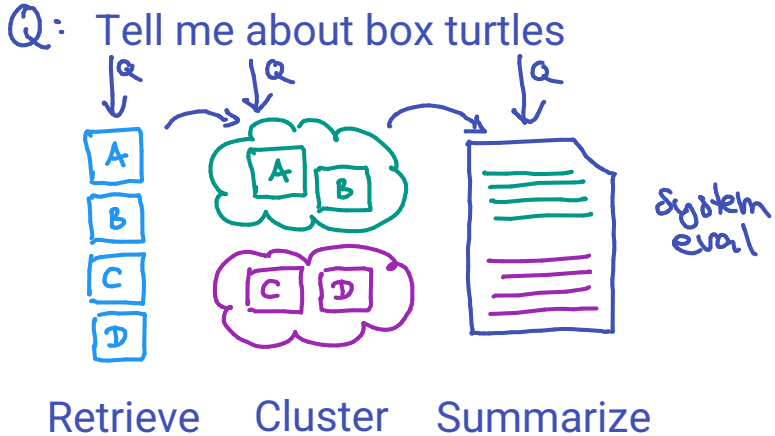
Idea:

1. Train to write Wikipedia articles
2. Read the web to produce articles for new topics.

Components for Article Generation

Given query/topic

- **Retrieve** relevant passages ← train/predict ↓
- **Cluster** passages into relevant sub-topics ← train/predict ↓
- **Summarize** text and remove redundancy ← train/predict ↓
- **Stitch** all text parts together ← System eval

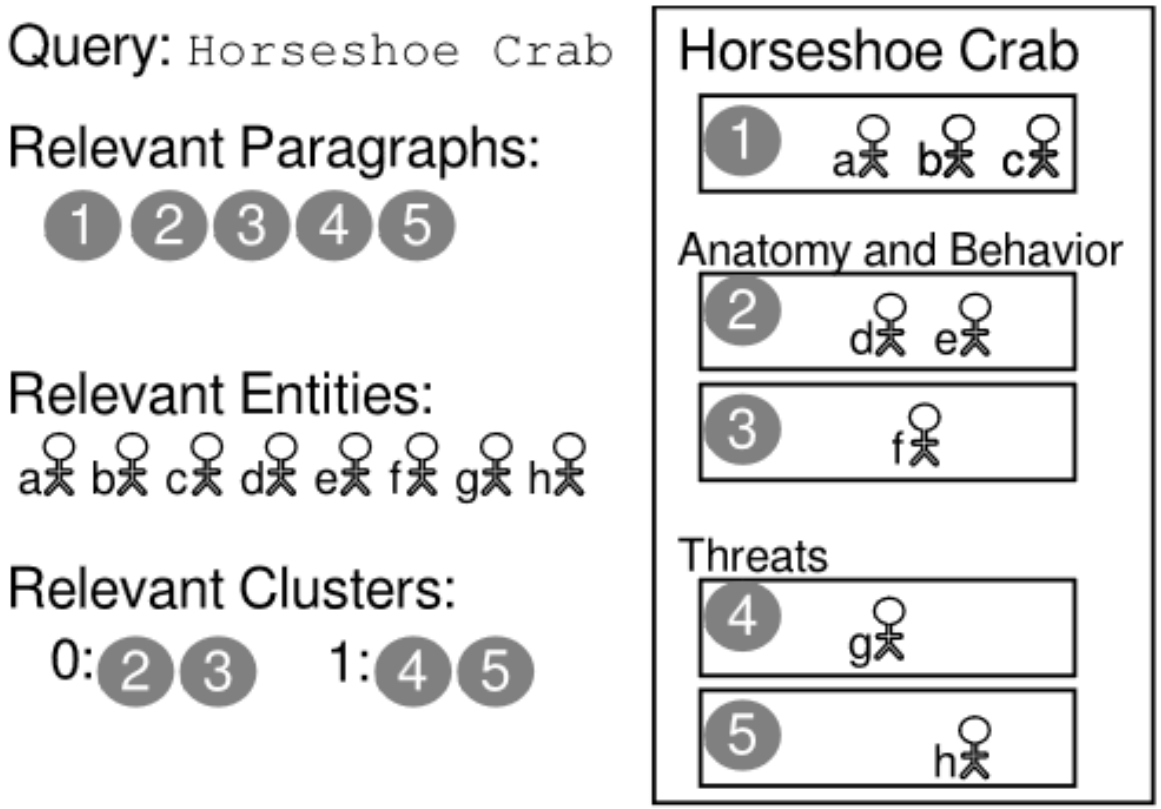


Train with Wikimarks Benchmarks

Dietz 2022 "Wikimarks: Harvesting Relevance Benchmarks from Wikipedia."

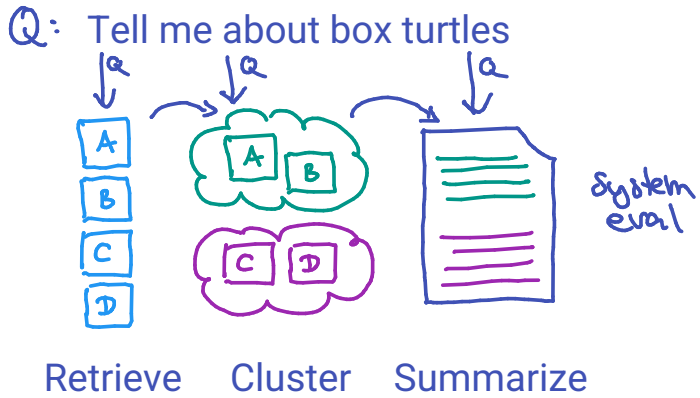
Trained on Wikimarks: Coordinated Benchmark

"Puzzle": (1) Take articles apart, (2) train to re-assemble



Dietz 2022 "Wikimarks: Harvesting Relevance Benchmarks from Wikipedia."

Components for Article Generation



Method	MAP
BM25-Section	**0.1566**
BM25-All	0.1429
BM25-Title	0.1010

Method	ARI
QS3M Mean	**0.3002**
QS3M Lead	0.2983
QS3M Title	0.2891
SBERT Euclid	0.2631
SBERT Cosine	0.2585

Coordinated benchmarks,
For each component:

- train each method
- select best method

+ System-level evaluation

Retrieval	Clustering	Precision
BM25-All	QS3M Lead	▲0.533
	QS3M Mean	0.537
	QS3M Title	▲**0.548**
	SBERT Cosine	0.530
	SBERT Euclid	▲0.541
BM25-Section	QS3M Lead	▲0.535
	*QS3M Mean	*0.525
	QS3M Title	0.530
	SBERT Cosine	0.516
	SBERT Euclid	0.529
BM25-Title	QS3M Lead	0.519
	QS3M Mean	0.512
	QS3M Title	0.523
	SBERT Cosine	▼0.494
	SBERT Euclid	▼0.507
Manual	Manual	▼0.409

Comparison (2)

	RAG	RM3	Entity Link PRF
Retrieve	trained end-to-end	heuristic (BM25)	heuristic (BM25)
Analyze	—	heuristic (RM)	heuristic (RM)
Exploit	trained (backprop)	heuristic	trained (L2R)

	Entity Guides	Article Generation
Retrieve	heuristic	heuristic
Analyze	coord training	coordinated training
Exploit	training	coordinated training

End-to-end $\rightarrow$ Compromise? $\leftarrow$ Coordinated Training

Multi-objective Coordinated Training

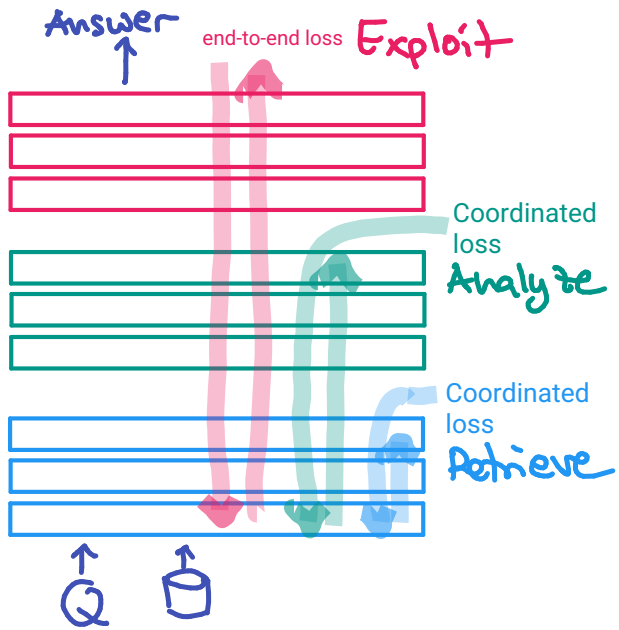
Standard approach: end-to-end training for task objective

- Costly (time + data)
- Issues propagrating deep

Coordinated Training Objectives

- Directly train lower layers
- Also train to serve the task

Variant: initialization with layer-block pre-training

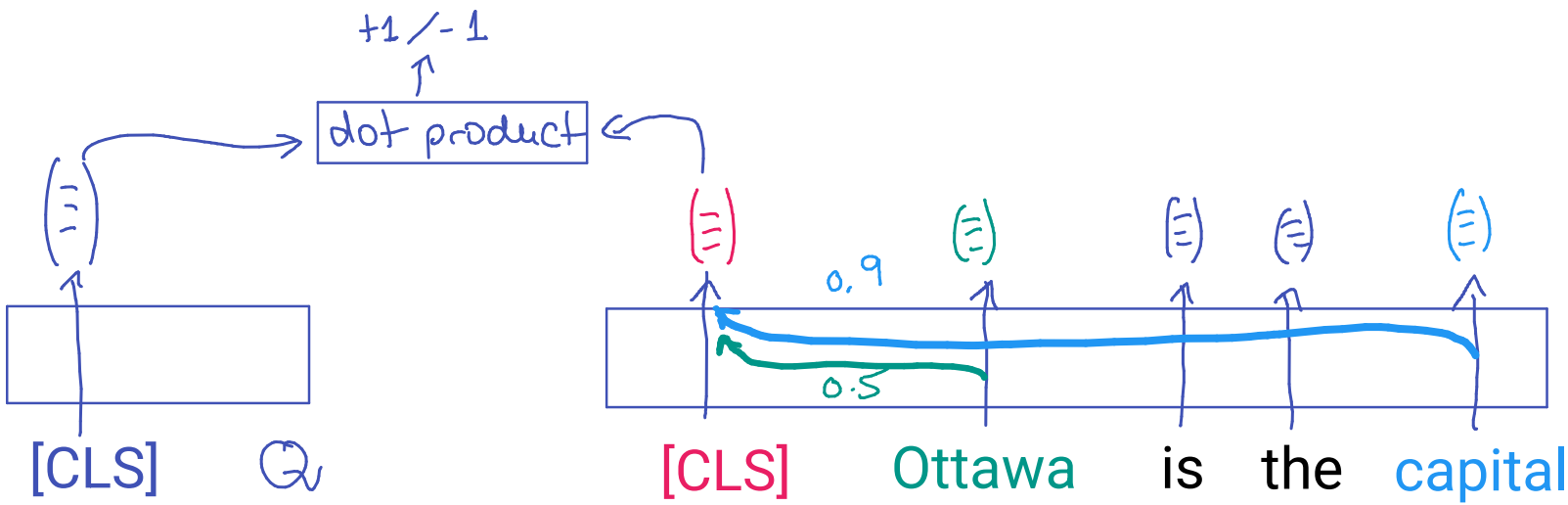


Outline

1. Guiding Entities
2. Coordinated Benchmarks
3. Exploiting Explainability in PRF
4. Retrieval-augmented Generation
5. Conclusion

Attention ≠ Relevance

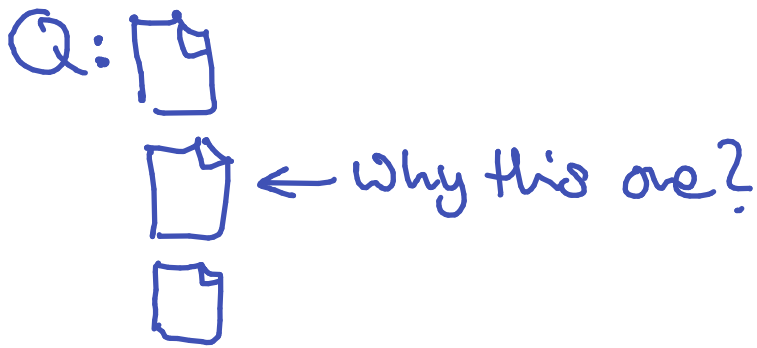
- Common idea: Parts with high attention are more relevant
- Sadly, this is not confirmed by empirical evaluations
  - Overparametrization = Different weights lead to same results.
  - Attention weigths can be absorbed in representation.



$0.5 \cdot$ (Ottawa) $= 0.1 \cdot$ (5 5 5)

Explainability, Saliency, or Rationale Models

- Explainability method for neural networks
- Which input features are most important for the model's predictions?



WHY is based on gradients of prediction w.r.t input features.

Anand 2022 "Explainable Information Retrieval: A Survey"

Exploiting Explainability in Retrieve-Analyze-Exploit

Retrieve Use Dense Retrieval model (DPR)

- Dense retrieval model
  score $= \texttt{proj}(\mathbf{q}) \cdot \texttt{proj}(\mathbf{d})$

Analyze with explainability methods to derive "WHY"

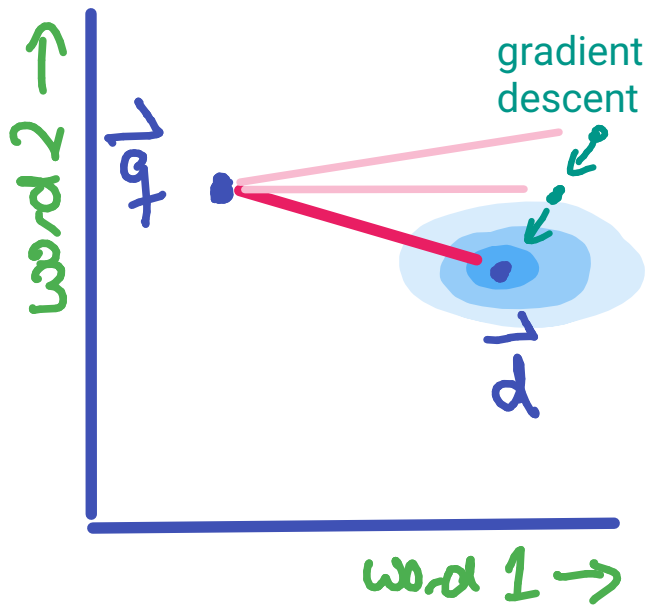
- will identify important terms
- use gradients of the DR model

Exploit important terms within Relevance Feedback framework

- query expansion to expand retrieved set
- use to find relevant symbols
  use symbols for better results

Explainability for Neural Networks

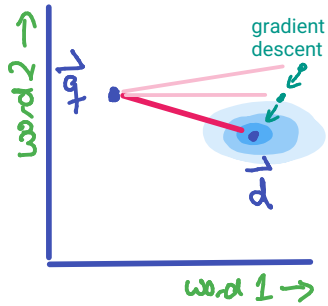
- Model: Dense retrieval $\quad\quad$ score $= \texttt{proj}(\mathbf{q}) \cdot \texttt{proj}(\mathbf{d})$
- Features are embedded words in retrieved documents
- Prediction is the relevance score



Take gradient ∇ of predicted output, w.r.t the input words
High magnitude of gradient indicates the
words need to change the least to affect the prediction the most?

Different Explainability Methods

Gradients $\nabla_{input}(output)$
of the model's output w.r.t. the input features



- **Saliency Maps**
  - absolute value of the gradients $\nabla_{input}(output)$
- **Gradient * Input**
  - Consider Signed gradients and magnitude of input features
  - Input features $\odot$ $\nabla_{input}(output)$
- **Integrated Gradients**
  - Computes the integral of gradients $\nabla_{input}(output)$ along a path from a "bad" baseline input to the actual input
- Attention mechanisms in NLP models (e.g., BiDAF)
- Layer-wise relevance propagation (LRP) for text-based models

Simonyan 2013 Saliency Maps - Shrikumar 2017 Gradient * Input - Sundararajan 2017 Integ

Note: No Ground Truth Required

- Saliency methods can be applied without knowledge of the ground truth
- Asks: Why did the model find this passage/symbol relevant?
- Methods: Trade-off accuracy and computational cost

We should do more research to explore the combination of Explainability, DPR, and PRF

Outline

1. Guiding Entities
2. Coordinated Benchmarks
3. Exploiting Explainability in PRF
4. Retrieval-augmented Generation
5. Conclusion

RAG - With Grounded Hypotheses

How to combine retrieved documents to generate a single answer?

Concatenate Issue: Input too large to process
also: needs to refresh when retrieved docs change

FiD (Izacard & Grave 2020): Generate latent representation of each document, fuse, then generate

RetGen/MoE (Zhang 2022,Cho 2020): From each document, generate a hypothesis (at token level) then integrate with Max Mutual Information

Generated hypotheses are "grounded" in words.

Question: Why not grounded in entities, facts, or other symbols?

Zhang 2022 "Joint Retrieval and Generation Training for Grounded Text Generation.

Neuro-Symbolic Q/A: Facts-as-Experts

- Transformer-based Masked Language Model
- Mask out entities, predict correct one
- To predict maked entities
  1. use context (the usual transformer)
  2. make use of external fact representations (memory)
- Only consider facts that are mentioned in the remaining text.

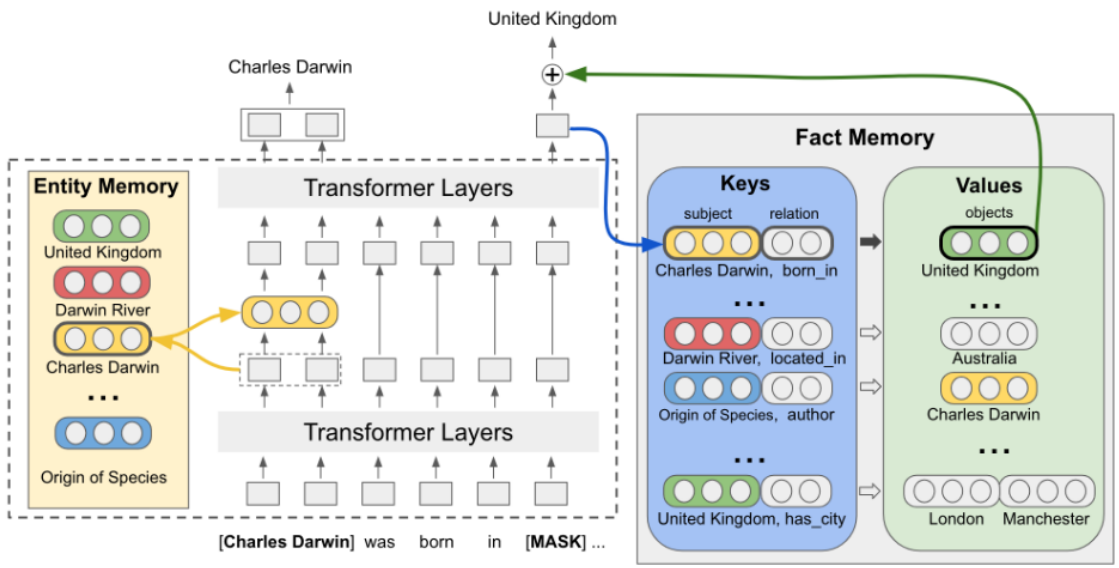


Figure 1: **Fact Injected Language Model architecture.** The model takes a piece of text (a question during fine-tuning or arbitrary text during pre-training) and first contextually encodes it with an entity enriched transformer.

Verga 2021. "Adaptable and interpretable neural memoryover symbolic knowledge."

Neuro-Symbolic Q/A: Facts-as-Experts

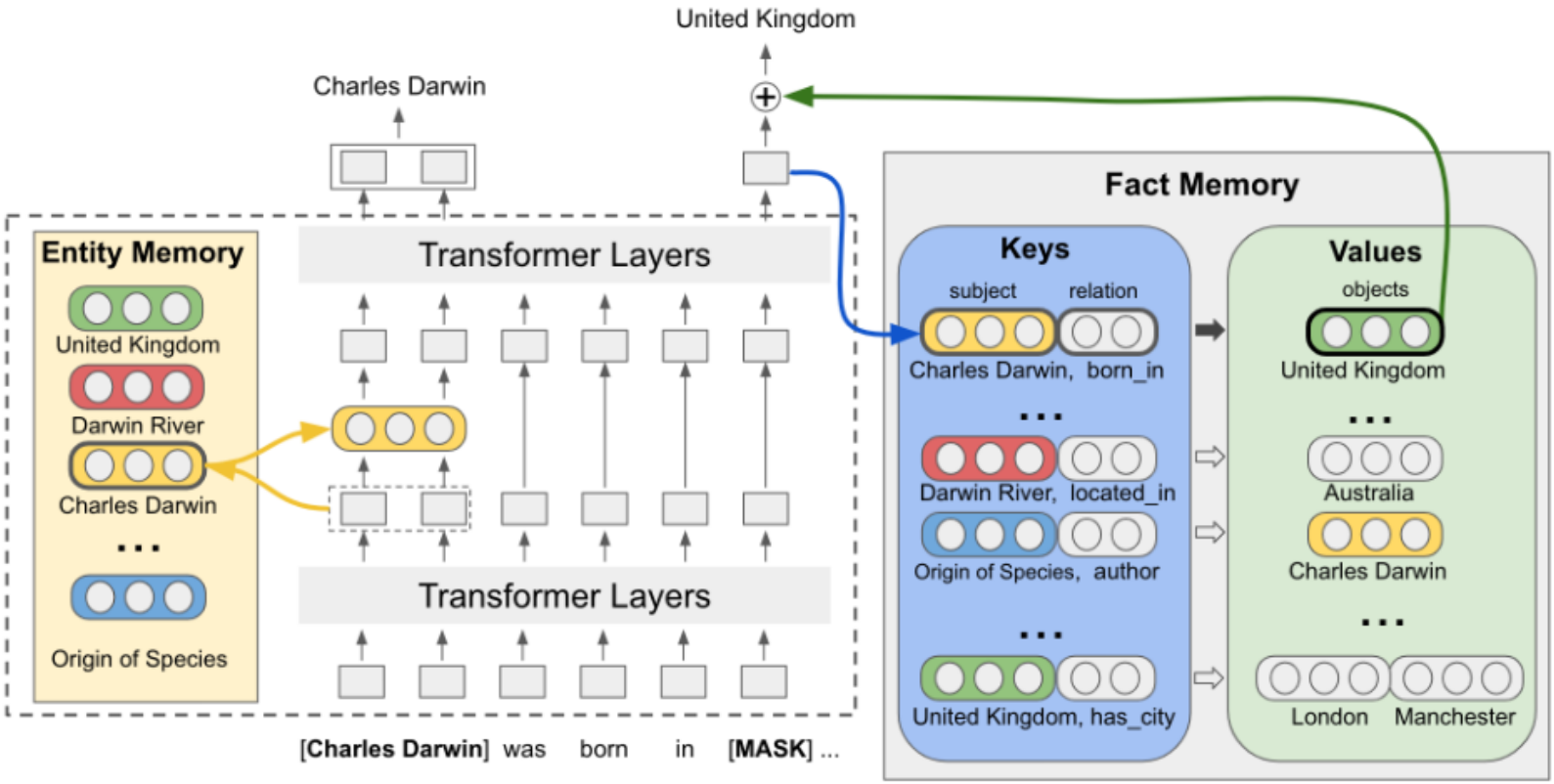


Figure 1: **Fact Injected Language Model architecture.** The model takes a piece of text (a question during fine-tuning or arbitrary text during pre-training) and first contextually encodes it with an entity enriched transformer.

Verga 2021. "Adaptable and interpretable neural memoryover symbolic knowledge."

Results on Open Domain Q/A

	Model	WebQuestionsSP				TriviaQA			
		Full Dataset		Wikidata Answer		Full Dataset		Wikidata Answer	
		Total	No Overlap	Total	No Overlap	Total	No Overlap	Total	No Overlap
Closed-book	**FILM**	**54.7**	**36.4**	**78.1**	**72.2**	**29.1**	**15.6**	**37.3**	**28.4**
	EaE	47.4	25.1	62.4	42.9	19.0	9.1	24.4	17.1
	T5-11B	49.7	31.8	61.0	48.5	–	–	–	–
	BART-Large	30.4	5.6	36.7	8.3	26.7	0.8	30.6	1.0
Open-Book	RAG	**50.1**	**30.7**	**62.5**	**45.1**	56.8	29.2	64.9	45.2
	DPR	48.6	34.1	56.9	45.1	57.9	31.6	66.3	48.8
	FID	–	–	–	–	**67.6**	**42.8**	**76.5**	**64.5**
	EmQL†	75.5	-	74.6	-	-	-	-	-

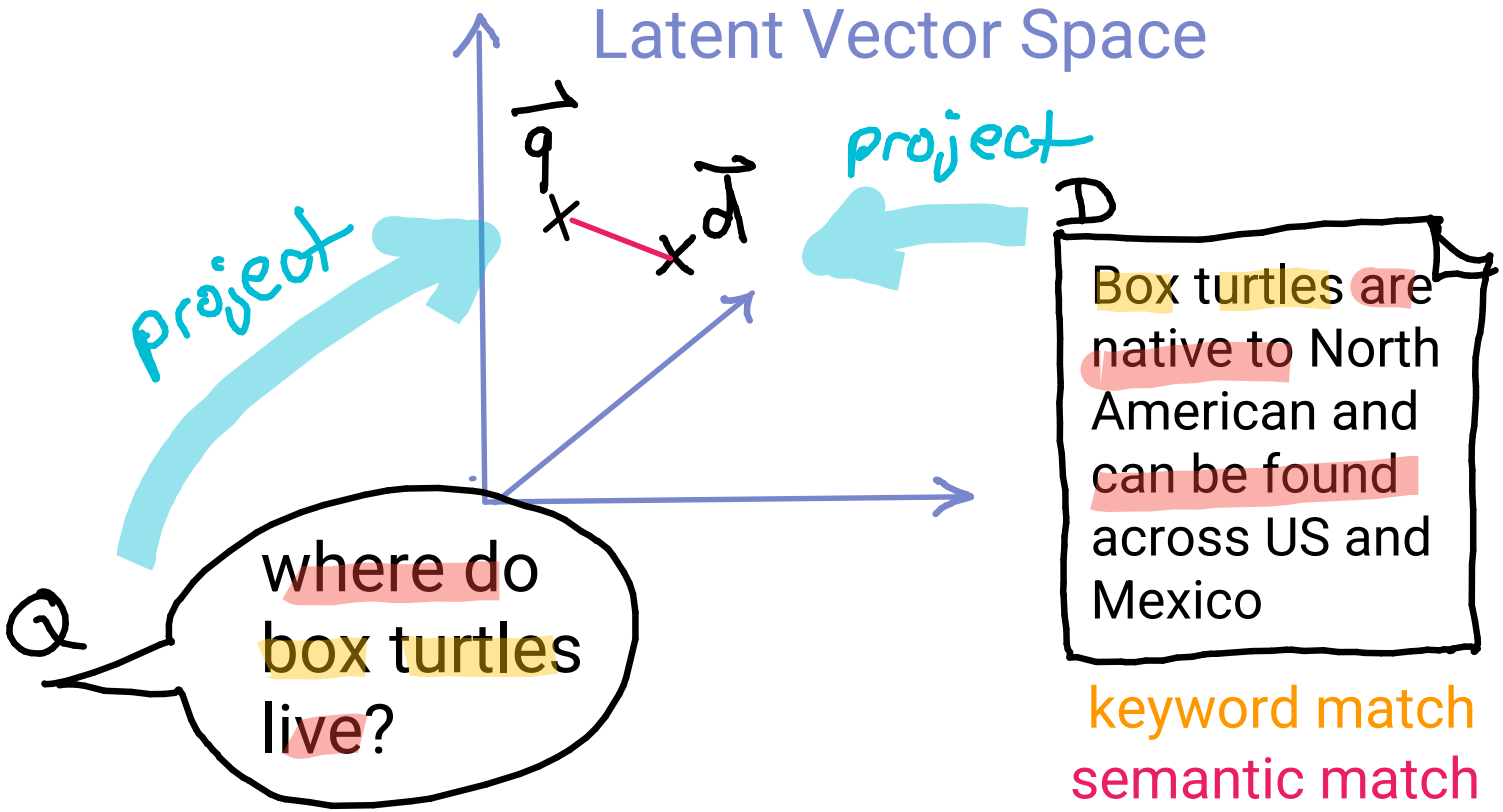
Verga 2021. "Adaptable and interpretable neural memoryover symbolic knowledge."

Outline

1. Guiding Entities
2. Coordinated Benchmarks
3. Exploiting Explainability in PRF
4. Retrieval-augmented Generation
5. Conclusion

SOTA: ad hoc Text Ranking

Dense Retrievel Approach:



Latent space and projections are trained, so that $\mathbf{q}$ and $\mathbf{d}$ are close whenever documents are relevant for the query.

Opportunities of Neuro-Symbolic Approaches

- Explainability and interpretability:
  - promising direction to exploit "model knowledge"
  - symbolic logic providing insights into the reasoning process
  - make it easier for users to trust the generated results
- Robustness to noise and ambiguity:
  - Use symbolic reasoning to when neural components struggle
  - more reliable for relevance and topics
  - dealing with incomplete or noisy data
- Transfer learning and generalization:
  - Leverage compositional nature of symbolic representation
  - Generalize to unseen queries or domains
  - Exploit known knowledge — rather than learning first principles

Future Research Opportunities for IR

Query Processing Improve the interpretation of user queries

- Expand the query with semantically related terms/symbols
- Decide what is not relevant
- Guess where to find non-obvious relevant info

Matching Improve the relevance of retrieved documents

- Identify similarity of semantically related information
- Reason about which connections are meaningful

Ranking Relative ordering of retrieved documents

- Capture the importance and relevance of entities to the query and the document
- Provide explanations and relevant background
- Fill knowledge graps in retrieved results