UBIQUITOUS KNOWLEDGE PROCESSING

TECHNISCHE UNIVERSITÄT DARMSTADT

SQUARE: TOWARDS MULTI-DOMAIN AND FEW-SHOT COLLABORATING QUESTION ANSWERING AGENTS

Prof. Iryna Gurevych

Haritz Puerto

AGENDA

1 New Possibilities

2 Meta-QA

3 Relevance Feedback In Neural Re-Ranking

4 UKP-SQuARE

5 Future Work

NEW POSSIBILITIES

TECHNISCHE UNIVERSITÄT DARMSTADT

EXPLOSION OF QA DATASETS AND MODELS

How to leverage all this collective effort?

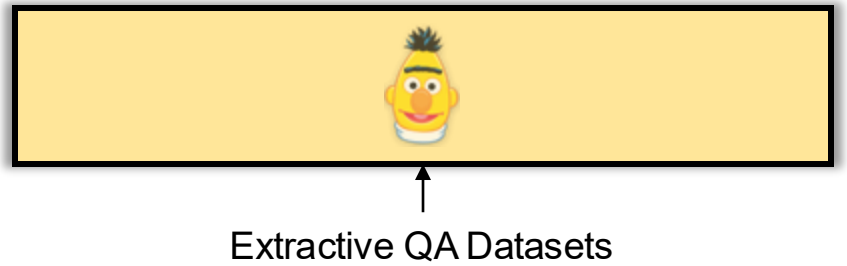
How to study all these models and datasets?

TECHNISCHE UNIVERSITÄT DARMSTADT

MULTI-SKILL QA

MULTI-DATASET MODELS

Train a model on many datasets

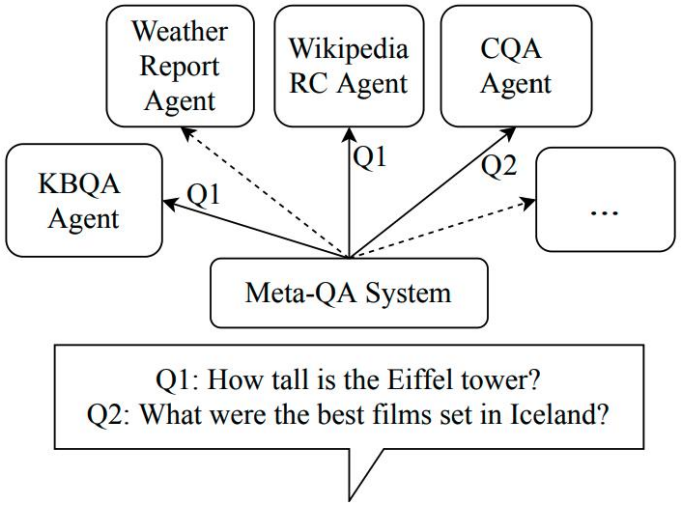


MultiQA (Talmor & Berant, ACL 2019)

UNIFIEDQA (Khashabi et al., Findings 2020)

MULTI-AGENT MODELS

Combine many models

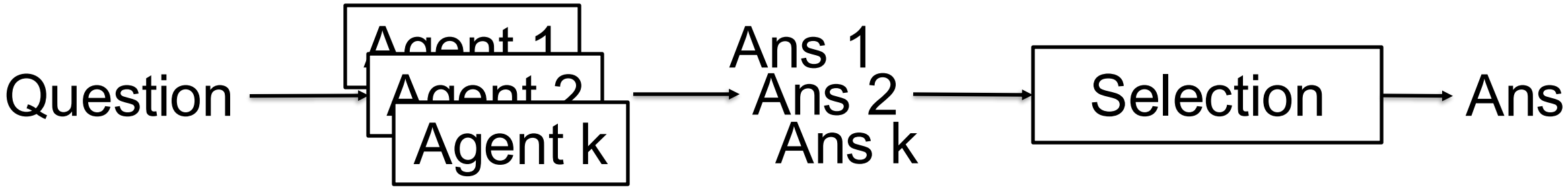


TWEAC: Transformer with Extendable QA Agent Classifiers (Geigle et al., Arxiv 2021)

METAQA

TECHNISCHE UNIVERSITÄT DARMSTADT

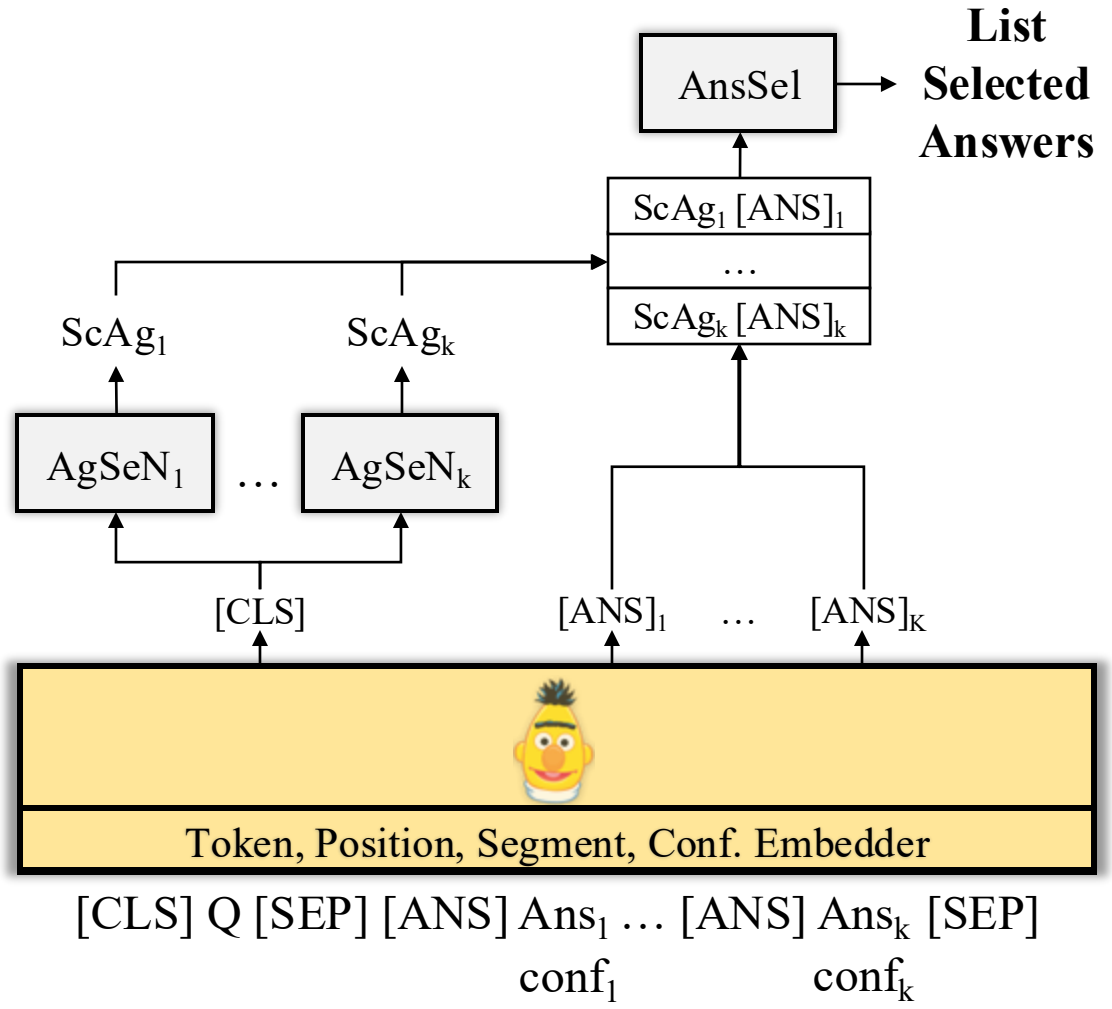
OVERVIEW



TECHNISCHE UNIVERSITÄT DARMSTADT

MODEL

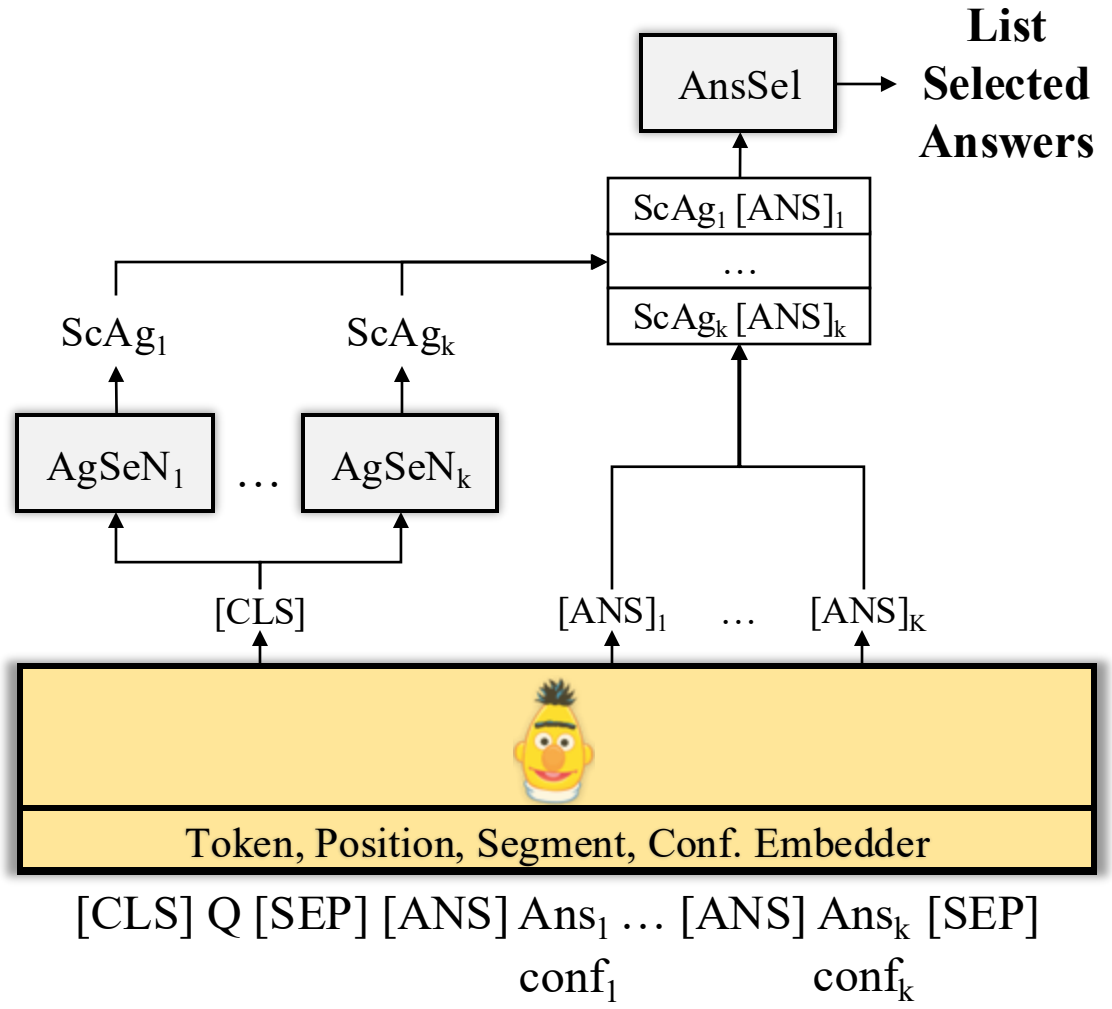
- **Input**: Question, Answers, Confidence Scores
- **Output**: List of selected answers
- **Task**: identify the best answer



TECHNISCHE UNIVERSITÄT DARMSTADT

MULTI-TASK

- **Agent** Selection ≈ Identify the in-domain agent
- **Answer** Selection: selects the best answer based on:
  - Domain of the question
  - Answer semantics



TECHNISCHE UNIVERSITÄT DARMSTADT

DATASETS

Dataset
SQuAD (Rajpurkar et al., 2016)
NewsQA (Trischler et al., 2017)
HotpotQA (Yang et al., 2018)
SearchQA (Dunn et al., 2017)
NQ (Kwiatkowski et al., 2019)
TriviaQA-web (Joshi et al., 2017)
QAMR (Michael et al., 2018)
DuoRC (Saha et al., 2018)

Extractive QA

Dataset
RACE (Lai et al., 2017)
CSQA (Talmor et al., 2019)
BoolQ (Clark et al., 2019)
HellaSWAG (Zellers et al., 2019)
SIQA (Sap et al., 2019)

Multiple-Choice QA

Dataset
DROP (Dua et al., 2019)
NarrativeQA (Kočiský et al., 2018)
HybridQA (Chen et al., 2020)

Abstractive & MultiModal QA

1 agent for each dataset

RESULTS

RESULTS

Datasets	MetaQA	UnifiedQA	Δ
Average (all – 16)	**76.39**	65.9	10.49
Extractive (8)	**77.72**	72.65	5.07
Multiple Choice (5)	**76.77**	61.41	15.36

LIMITATIONS OF MULTI-DATASET MODELS

- Cannot model specific tasks (eg: numerical reasoning)
- Weak in minority domains (eg: Trivia and movie questions)

Dataset	MetaQA	UnifiedQA	Δ
TriviaQA-web	**80.63**	72.34	8.29
DuoRC	**51.24**	34.65	16.59
CSQA	**78.66**	58.43	20.23
HellaSWAG	**73.19**	36.01	37.18
DROP	**73.04**	42.45	30.59

OOD RESULTS

Model	TriviaQA	DuoRC	CSQA	HellaSWAG	SIQA	Δ
UnifiedQA	72.34	34.65	58.43	36.01	**61.62**	-
OOD MetaQA	**75.02**	**50.51**	**58.59**	**52.13**	59.28	-
OOD UnifiedQA	69.33	32.84	50.57	29.35	44.93	-8.12

Best results in bold.
OOD MetaQA outperform in-domain UnifiedQA underlined

COLLABORATION BETWEEN AGENTS

Dataset	In-Domain Agent	OOD Agent
DuoRC	48%	NewsQA (18%)
TriviaQA	80%	DuoRC (3%)
SearchQA	86%	TriviaQA (5%)

Dataset	Question	In-domain Agent	OOD Agent
DuoRC	Who does Rocky Balboa work for as an enforcer?	Adrian	Tony Gazzo (NewsQA Agent)
TriviaQA-web	Who played the character Mr Chips in the 2002 TV adaptation of Goodbye Mr Chips?	Timothy Carroll	MartinClunes (DuoRC Agent)
SearchQA	This short story, written around 1820, contains the line "If I can but reach that bridge... I am safe"	Legend	Legend of Sleepy Hollow (TriviaQA Agent)

EFFICIENCY

Training

- MetaQA is very training efficient
- MetaQA uses 16% of the data used by UnifiedQA

MetaQA learns to match Questions with Answers instead of end-to-end QA

Inference

Model	Time (s) per question
UnifiedQA	2.15 ± 0.02
MetaQA	7.08 ± 0.16

On SQuARE Platform using CPU

https://square.ukp-lab.de

CONCLUSIONS

- We proposed MetaQA, a system to combine multiple expert agents
- Large performance gains wrt multi-dataset models (in-domain and OOD scenarios)
- Agent Collaboration → key to performance
- MetaQA is highly efficient to train
- Inference time remains reasonable

TECHNISCHE UNIVERSITÄT DARMSTADT

MetaQA: Combining Expert Agents for Multi-Skill Question Answering

Haritz Puerto[1], **Gözde Gül Şahin**[2], **Iryna Gurevych**[1]
[1]Ubiquitous Knowledge Processing Lab (UKP Lab)
Department of Computer Science and Hessian Center for AI (hessian.AI)
Technical University of Darmstadt, Germany
[2] Department of Computer Science, Koç University, KUIS AI Lab
Istanbul, Türkiye

EACL 2023

GitHub

TECHNISCHE UNIVERSITÄT DARMSTADT

SQUARE: TOWARDS MULTI-DOMAIN AND FEW-SHOT COLLABORATING QA AGENTS

RELEVANCE FEEDBACK IN NEURAL RE-RANKING

QUERY TYPES [1]

	Navigational / Transactional	Information-Seeking
Example query	“arxiv sentence bert” “acl paper submission”	“origin coronavirus” “neural sentence representation”
# relevant documents	1-few	Many
Scenarios	Navigation Known-Item Retrieval	Scientific Literature Review News Background Case Law Factoid QA

MOTIVATION

Challenges

- Unknown Search Domain
- Difficult Query Formulation

Observation

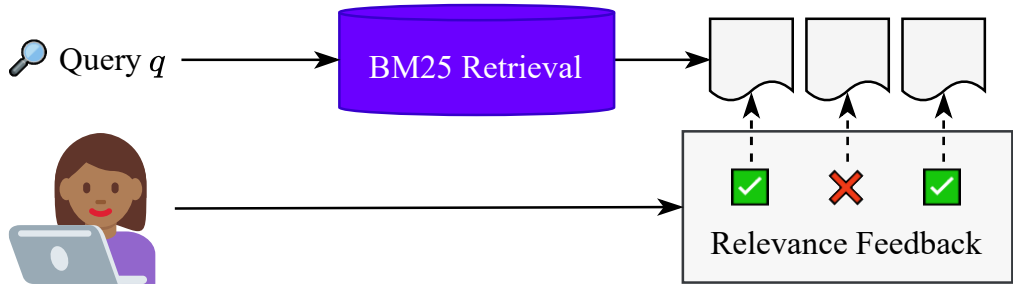
- Judging a document is easier than formulating good queries
- Some relevant documents might be already known to the user

→ **Use relevance feedback to improve search**

BACKGROUND

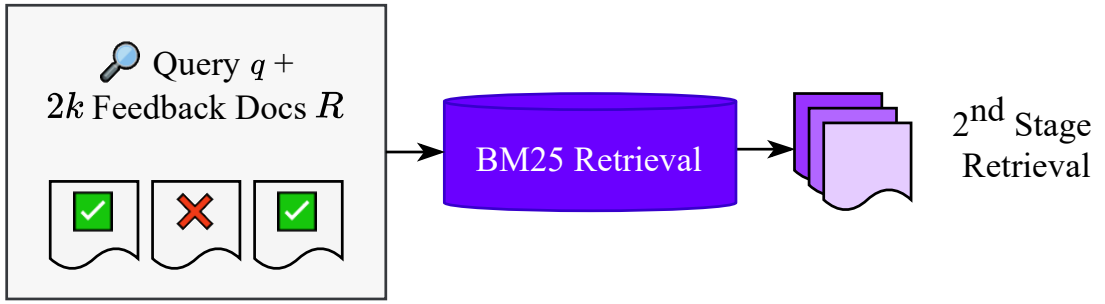
1) Retrieve Documents using the query

2) Obtain feedback on retrieved documents



3) Extract ”expansion terms” from documents

4) Retrieve with query + expansion terms



How to integrate relevance feedback into neural retrieval?

TECHNISCHE UNIVERSITÄT DARMSTADT

TASK SETUP

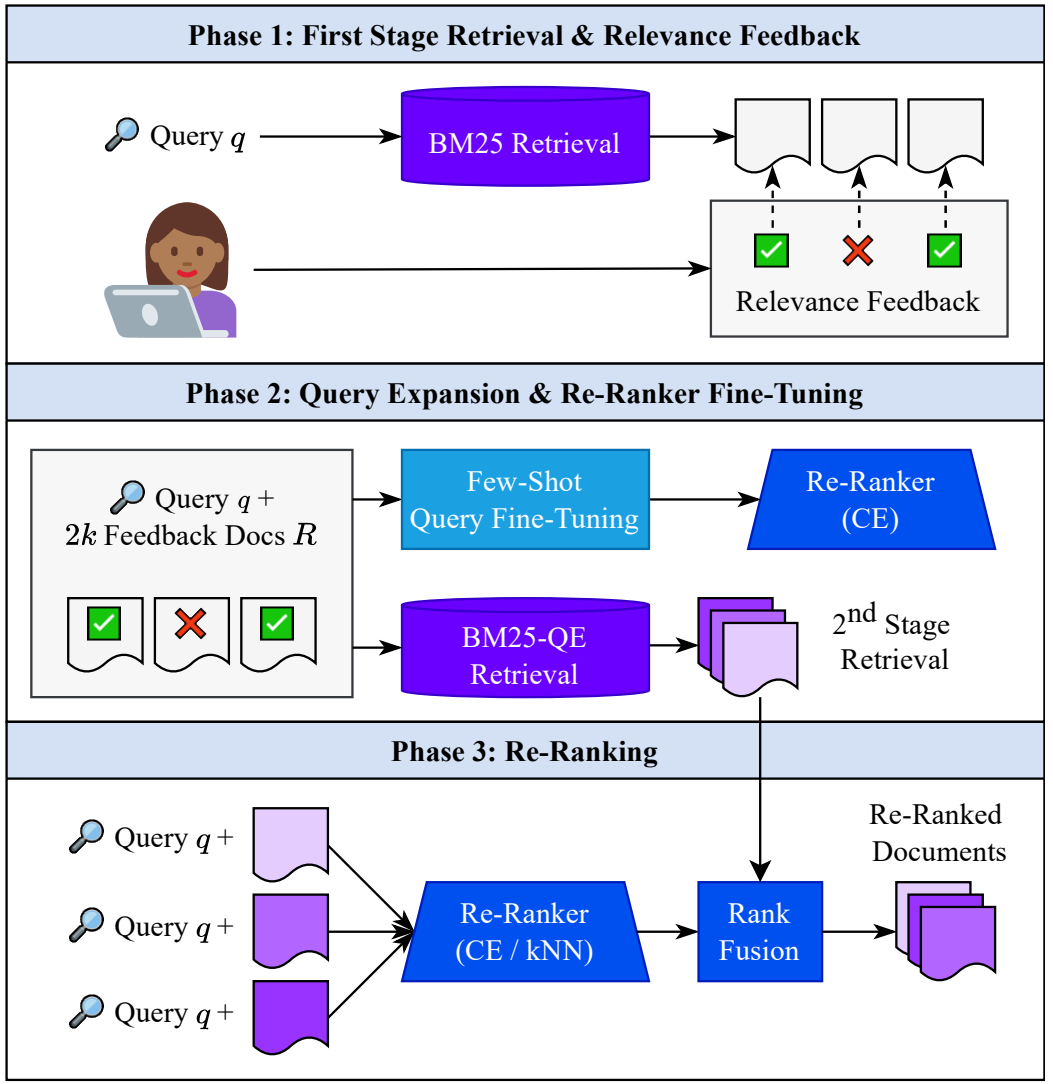
Goal: Re-rank documents from 2. Stage Retrieval with neural re-rankers incorporating relevance feedback

Input

- Query q
- k relevant and non-relevant feedback documents, where $k \in \{2, 4, 8\}$
- 1000 documents from 2. stage retrieval

Evaluation

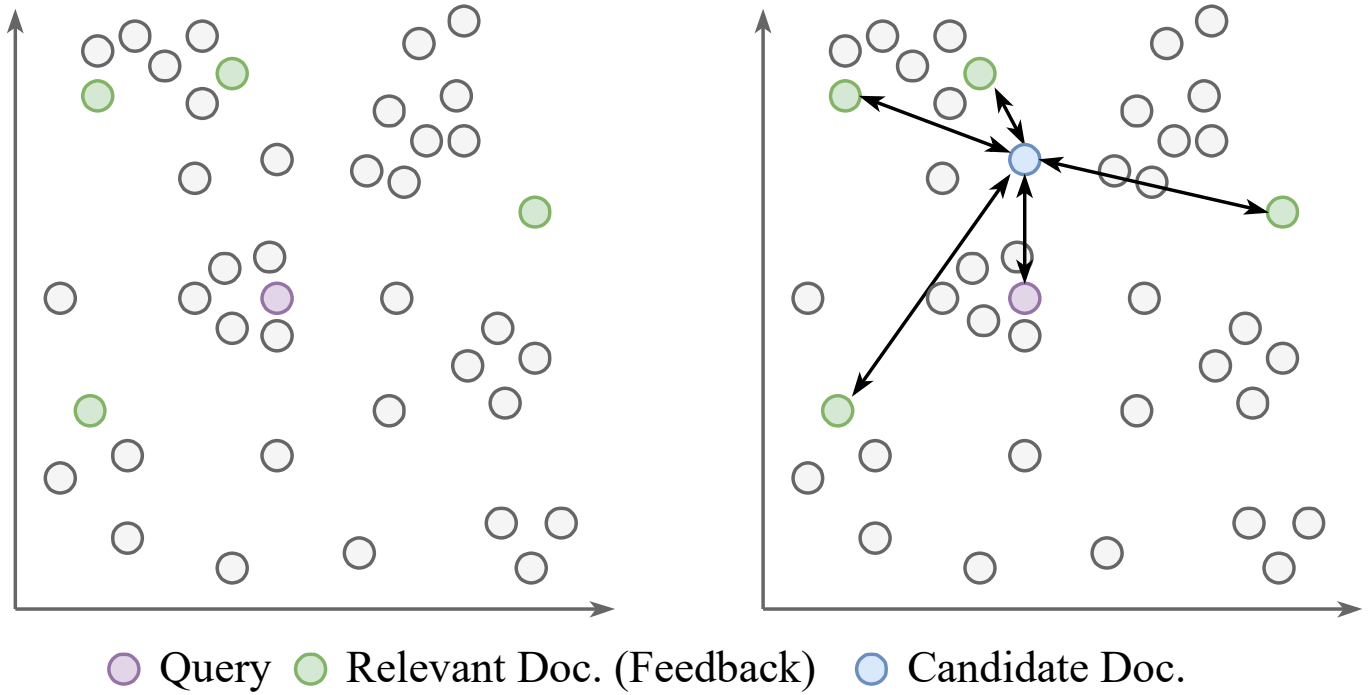
- Re-ranking [nDCG@20]



METHOD: KNN

- Compute document representations $d_i \in D$
- Score documents by summing the similarities between the candidate document d_i, query q and relevant feedback documents $d_j \in R^+$

$$s_i = f(d_i, q) + \sum_{d_j \in R^+} f(d_i, d_j)$$



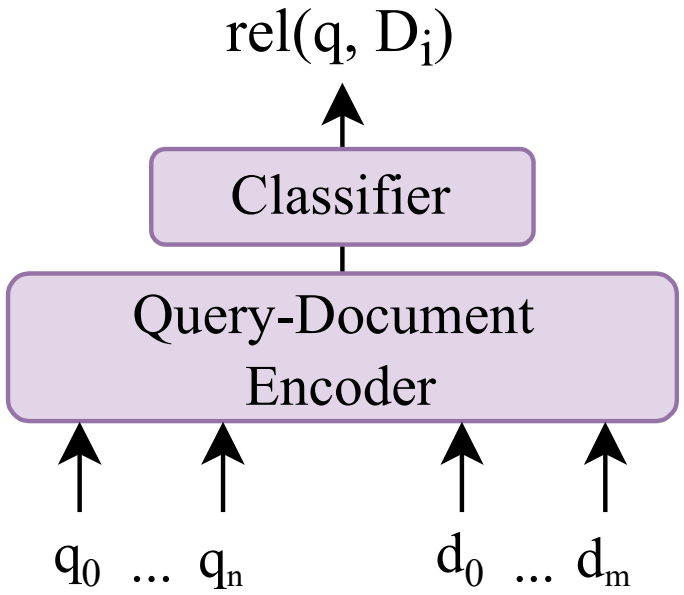
METHOD: CROSS-ENCODER [2]

CE Query Fine-Tuning

Fine-Tune bias layers per query on $2k$ Relevance Feedback Documents

CE MAML + Query Fine-Tuning

1. Fine-Tune bias layers on in-domain annotations with Meta-Learning to obtain ”fast parameters”
2. Fine-Tune per query on on $2k$ Feedback Documents



METHOD: RANK-FUSION [3]

Idea: Merge rankings of different ranking functions $h \in H$

Problem: Different methods produce different scores, simple adding the scores is biased

=> Use ranks instead of raw scores

$$s_i = \sum_{h \in H} \frac{1}{c + h(d_i)}$$

$h(d_i)$ returns the integer rank that method h assigns document d_i

c is constant smoothing the impact of top ranked documents

DATASETS

Dataset	Domain	Docs.	Queries	Judgments
Robust04	News	528k	148	1287 (±501)
TREC-Covid	Biomedical	191k	50	1370 (±323)
TREC-News	News	595k	34	259 (±82)
Touché-2020	Debates	383k	49	50 (±7)

Includes only queries with at least 32 relevant documents.

RESULTS

Method	Robust	Covid	News	Touché	Avg
BM25-QE (2. Stage Retrieval)	**0.496**	0.610	**0.392**	**0.271**	**0.442**
kNN	0.443	0.686	0.365	0.174	0.417
CE Zero-Shot	0.415	**0.702**	0.314	0.176	0.402

=> BM25-QE is hard to beat!

RESULTS

Method	Robust	Covid	News	Touché	Avg
BM25-QE (2. Stage Retrieval)	0.496	0.610	**0.392**	**0.271**	**0.442**
CE Query Fine-Tune	0.484	0.723	0.335	0.198	0.435
CE MAML + Query FT	**0.506**	**0.735**	0.314	0.223	**0.445**

=> Fine-tuning per query helps
=> CE MAML + Query FT is on par with BM25-QE

RESULTS

Method	Robust	Covid	News	Touché	Avg
BM25-QE (2. Stage Retrieval)	0.496	0.610	0.392	**0.271**	0.442
kNN	0.443	0.686	0.365	0.174	0.417
CE Zero-Shot	0.415	0.702	0.314	0.176	0.402
CE Query Fine-Tune	0.484	0.723	0.335	0.198	0.435
CE MAML + Query FT	0.506	0.735	0.314	0.223	0.445
BM25-QE ∩ kNN	0.507	0.707	**0.412**	0.248	0.468
BM25-QE ∩ CE MAML + Query FT	**0.570**	**0.740**	0.405	**0.272**	**0.497**

=> Fusing BM25-QE and neural model is highly effective!

TECHNISCHE UNIVERSITÄT DARMSTADT

Incorporating Relevance Feedback for Information-Seeking Retrieval using Few-Shot Document Re-Ranking

Tim Baumgärtner,[1] Leonardo F. R. Ribeiro,[1*] Nils Reimers,[2] Iryna Gurevych[1]
[1]Ubiquitous Knowledge Processing Lab (UKP Lab),
Department of Computer Science and Hessian Center for AI (hessian.AI),
Technical University of Darmstadt
[2]cohere.ai
www.ukp.tu-darmstadt.de

EMNLP 2022

GitHub

TECHNISCHE UNIVERSITÄT DARMSTADT

SQUARE: TOWARDS MULTI-DOMAIN AND FEW-SHOT COLLABORATING QA AGENTS

UKP-SQUARE

AN ONLINE PLATFORM FOR QA RESEARCH

TECHNISCHE UNIVERSITÄT DARMSTADT

AN ECOSYSTEM FOR QA RESEARCH

- Common **interface**
- Easy **analysis** of the strengths, weaknesses, and biases
- Common **explainability** methods
- Common **adversarial** attacks
- Graph **visualizations**

ACL 2022 AACL 2022

- Running on the **browser** (no configurations, no installations)
- Re-use **data sources**
- Easy to **deploy** new **models**
- Dozens of **models available**

Data icons created by Freepik - Flaticon

ACL 2022

AACL 2022

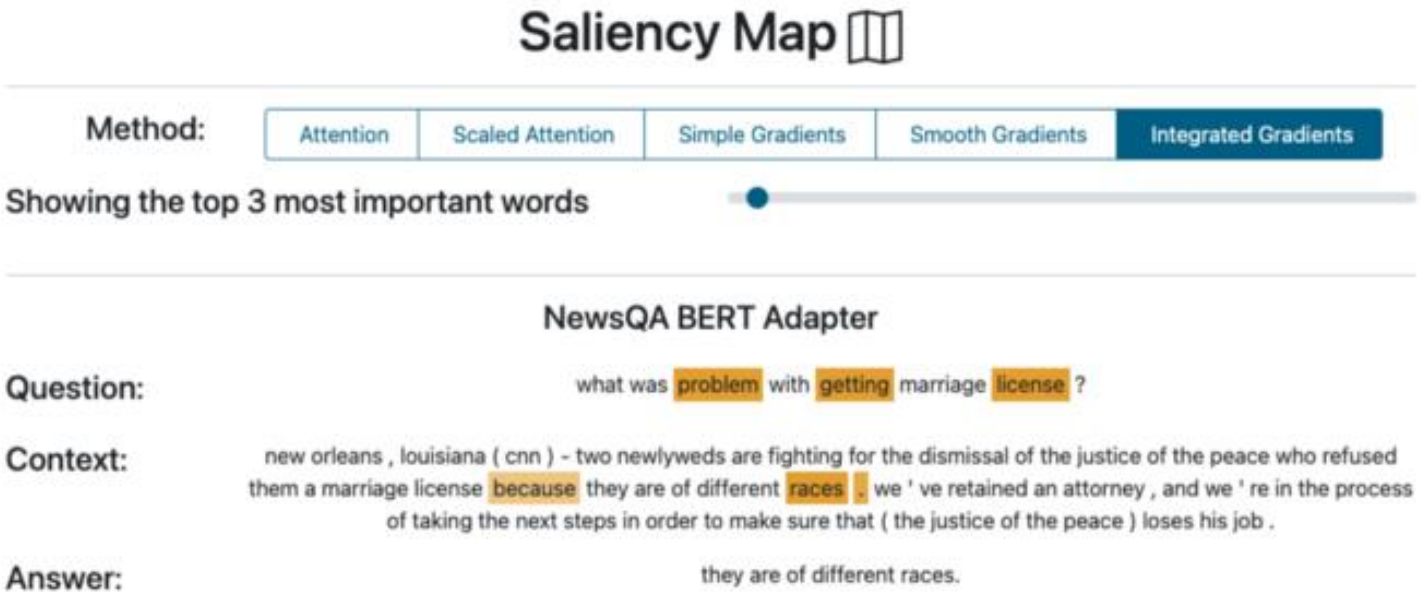
arXiv 2023

TECHNISCHE UNIVERSITÄT DARMSTADT

EXPLAINABILITY

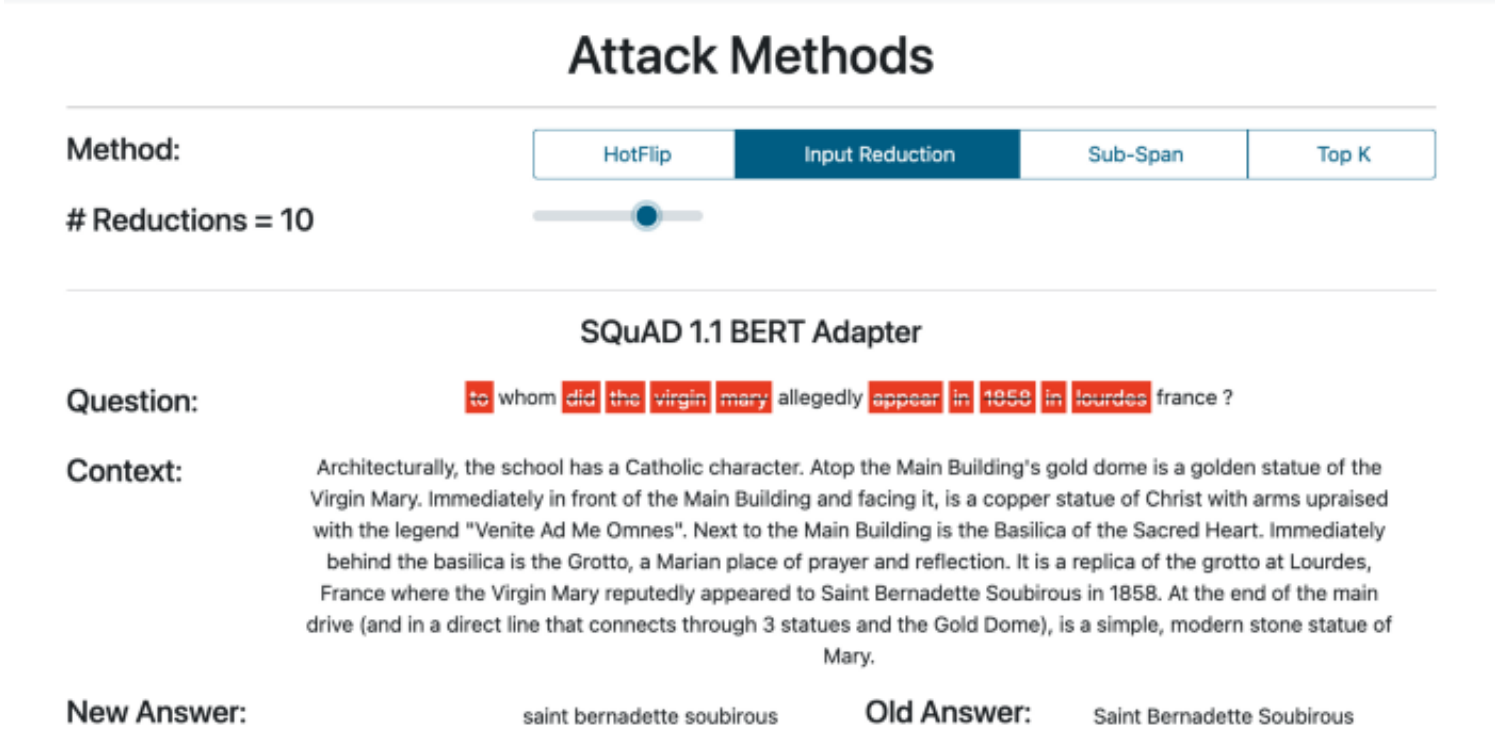
Saliency Maps

- Attention (Jain et al., ACL 2020)
- Scaled Attention (Serrano & Smith, ACL 2019)
- Simple Gradient (Simonyan et al., arXiv 2013)
- Smooth Gradients (Smilkov et al., arXiv 2017)
- Integrated Gradients (Sundararajan et al., PMLR 2017)



ADVERSARIAL ATTACKS

- HotFlip (Ebrahimi et al., ACL 2018)
- Input Reduction (Feng et al., EMNLP 2018)
- Sub-Span
- Top-K

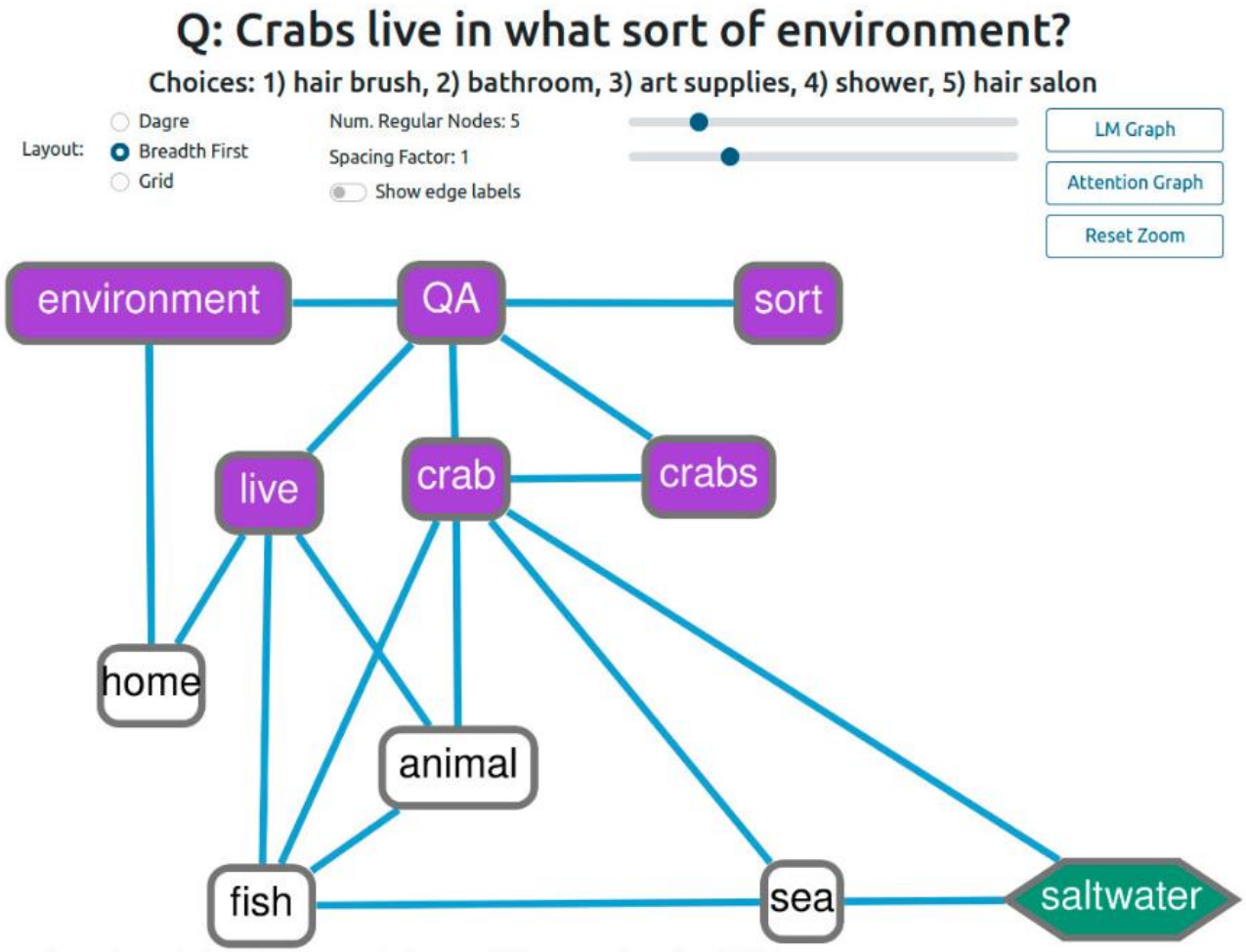


TECHNISCHE UNIVERSITÄT DARMSTADT

GRAPH-BASED MODELS

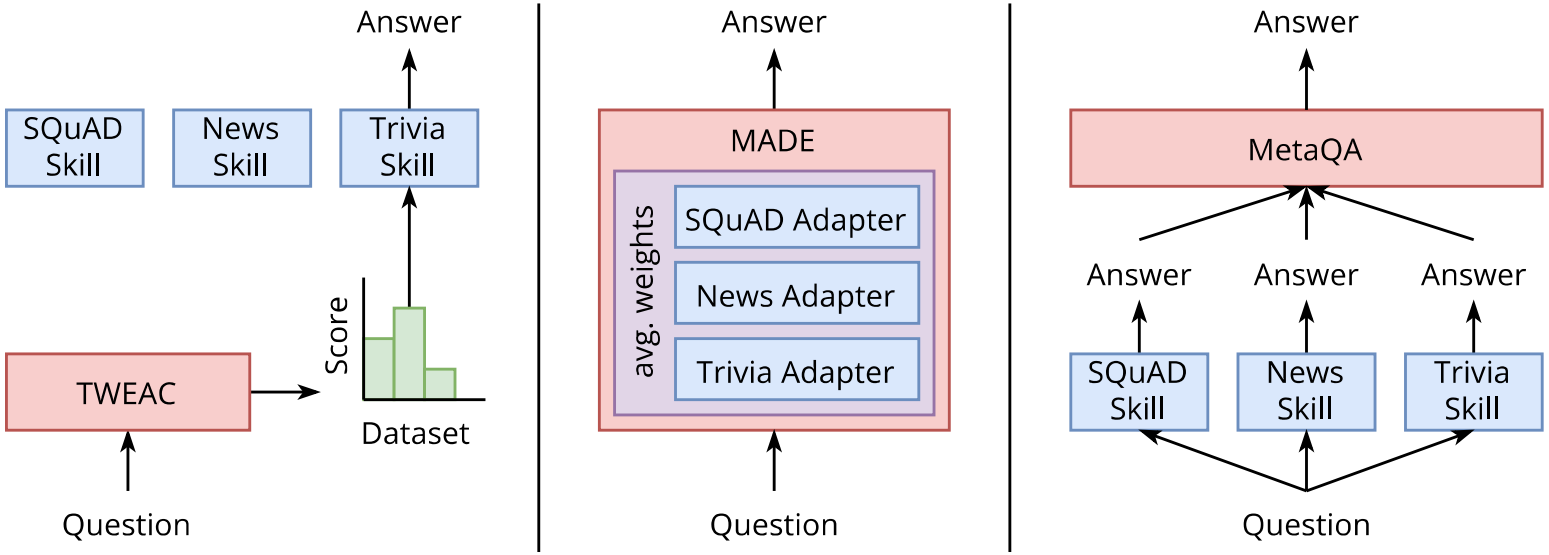
QA-GNN

- LM + KB for common-sense QA
- KB: ConceptNet
- Graph Viz can help us understand the behavior of the model



QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering (Yasunaga et al., NAACL 2021)

MULTI-AGENT SYSTEMS



- Qualitative analysis of multi-agent systems
- Qualitative comparison against multi-dataset systems

DEMO

HTTPS://SQUARE.UKP-LAB.DE

FUTURE WORK

PROMPTING LANGUAGE MODELS

- Comparing multiple LM in parallel
- Sharing prompts easily with the community
- Augmenting LM
  - Documents
  - External tools (eg: calculators)
  - Knowledge Graphs
  - Other (smaller) models

TECHNISCHE UNIVERSITÄT DARMSTADT

THANK YOU!

gurevych@tu-darmstadt.de

https://square.ukp-lab.de

https://twitter.com/UKP_SQuARE

https://github.com/UKP-SQuARE