Tencent AI Lab

Towards Better Retrievers for Knowledge-intensive Tasks

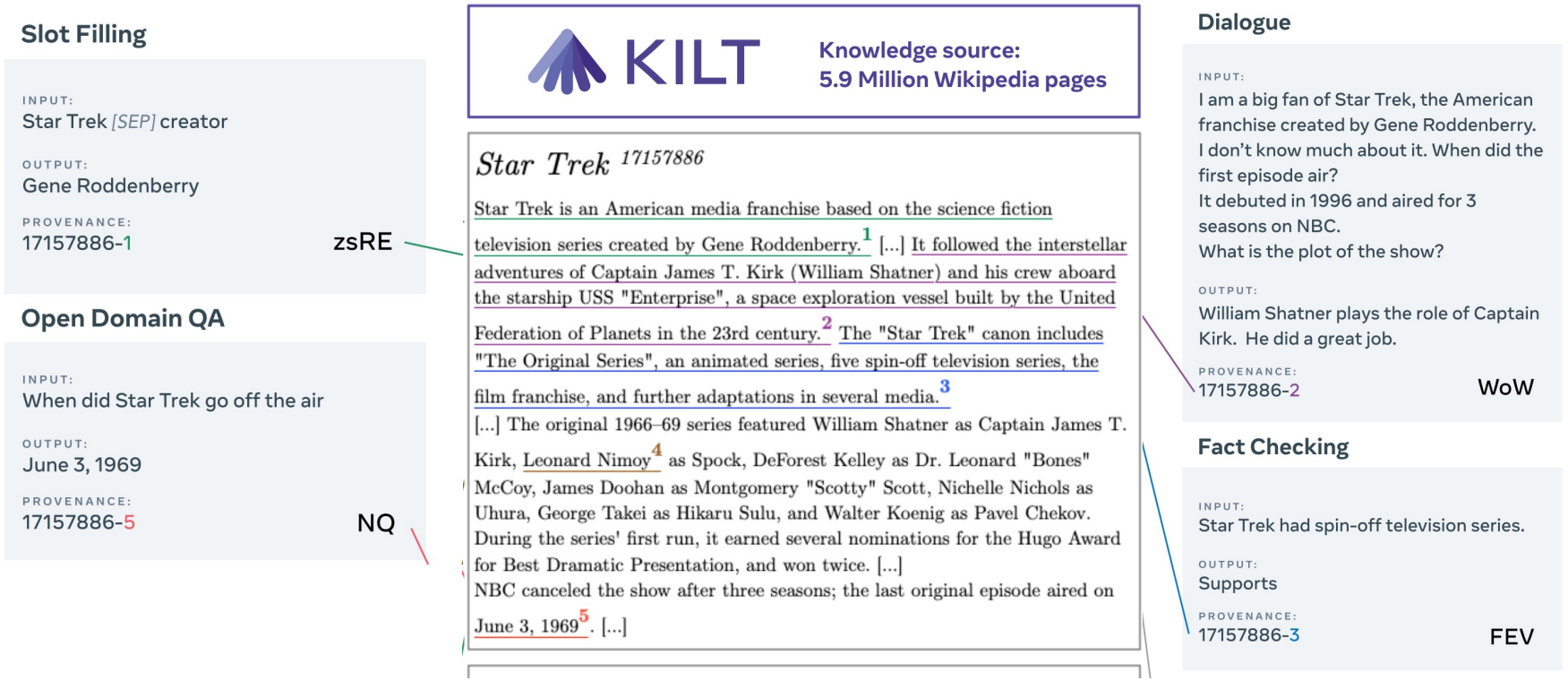
Xinting Huang

2022.03.16

Contents

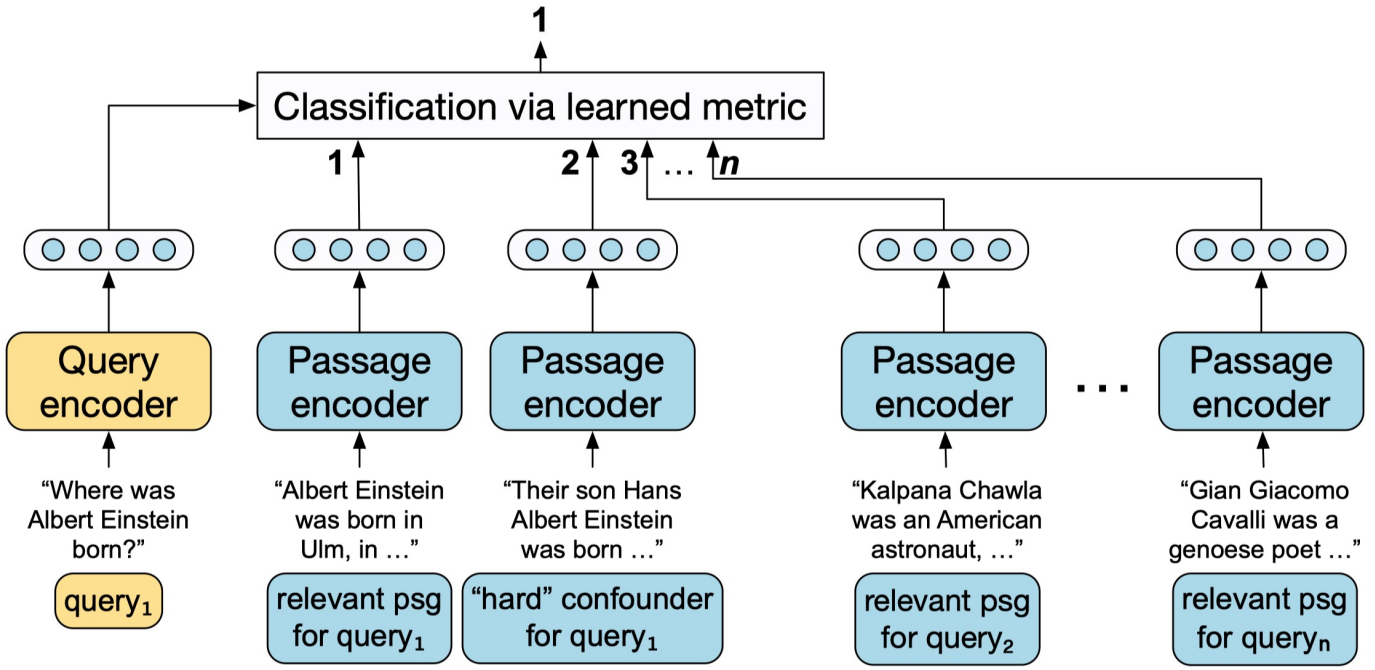
Background: Knowledge-Intensive Tasks

- Knowledge-Intensive Tasks
  - Tasks that require access to large, external knowledge sources
  - The knowledge sources can take various forms: plain text (e.g., Wikipedia passages), structured knowledge (e.g., tables)
  - Solving knowledge-intensive tasks commonly follow a retrieve-and-read/generate paradigm.



Background: Retriever Learning

- Dense Retriever for Knowledge Selection
  - Query and each passage are encoded, relevancy is determined by the distance of their vectors
  - Training is based on a contrastive loss, using triplets of query, a relevant passage, and a set of irrelevant passages
  - Training involves both self-supervised and supervised learning



Challenges: Test-Train Overlapping

Tencent AI Lab

- Investigating Test-Train Overlapping (21EACL, Facebook)
  - This work explores the test sets of three open-domain QA datasets
  - Two types of test-train overlap are identified: question and answer
    - Question: a random subset annotated by humans;
    - Answer: lexical similarities;
  - Experiments are conducted on NaturalQuestions, TriviaQA, and WebQuestions; The knowledge source is Wikipedia passages
  - All models perform significantly higher on memorizable questions;
  - The drop is severe for almost all models, with an average absolute drop of 25% with respect to total performance.

Dataset	% Answer overlap	% Question overlap
NaturalQuestions	63.6	32.5
TriviaQA	71.7	33.6
WebQuestions	57.9	27.5

Table 1: Test-train set overlap for the datasets.

Model		Open NaturalQuestions			
		Total	Question Overlap	Answer Overlap Only	No Overlap
Open book	RAG	44.5	70.7	34.9	24.8
	DPR	41.3	69.4	34.6	19.3
	FID	51.4	71.3	48.3	34.5
Closed book	T5-11B+SSM	35.2	76.5	21.0	8.4
	BART	26.5	67.6	10.2	0.8
Nearest Neighbor	Dense	26.7	69.4	7.0	0.0
	TF-IDF	22.2	56.8	4.1	0.0

Lewis et al. (2021). Question and Answer Test-Train Overlap in Open-Domain Question Answering Datasets. *EACL*.

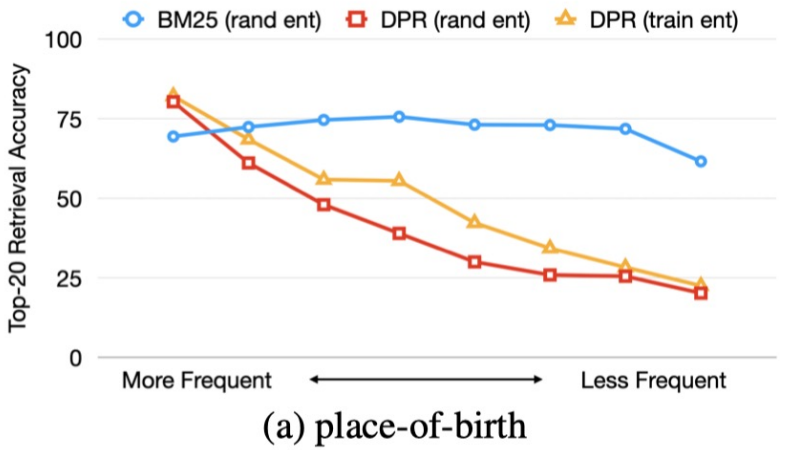
Challenges: Simple Query Rewriting

- Simple Query Rewriting (21EMNLP, Princeton)
  - Exploring the retriever abilities to handle simple, entity-centric questions
  - Collecting an evaluation benchmark using manually defined templates to convert Wikipedia facts, i.e., <subject, relation, object>
  - It still clearly DPR is outperformed by BM25. The gaps are especially large on questions about person entities.

  - The effects of entity frequency is further explored.
  - DPR representations are much better at representing the most common entities as well as entities observed during training

  - What if the challenging queries can be seen during training?

	DPR (NQ)	DPR (multi)	BM25 -
Natural Questions	**80.1**	79.4	64.5
EntityQuestions (this work)	49.7	56.7	**71.2**
What is the capital of [E]?	77.3	78.9	**90.0**
Who is [E] married to?	35.6	48.1	**85.9**
Where is the headquarter of [E]?	70.0	72.0	**85.0**
Where was [E] born?	25.4	41.8	**75.2**
Where was [E] educated?	26.4	41.8	**73.0**
Who was [E] created by?	54.1	57.7	**71.7**
Who is [E]'s child?	19.2	33.8	**82.9**
(17 more types of questions)	…	…	…

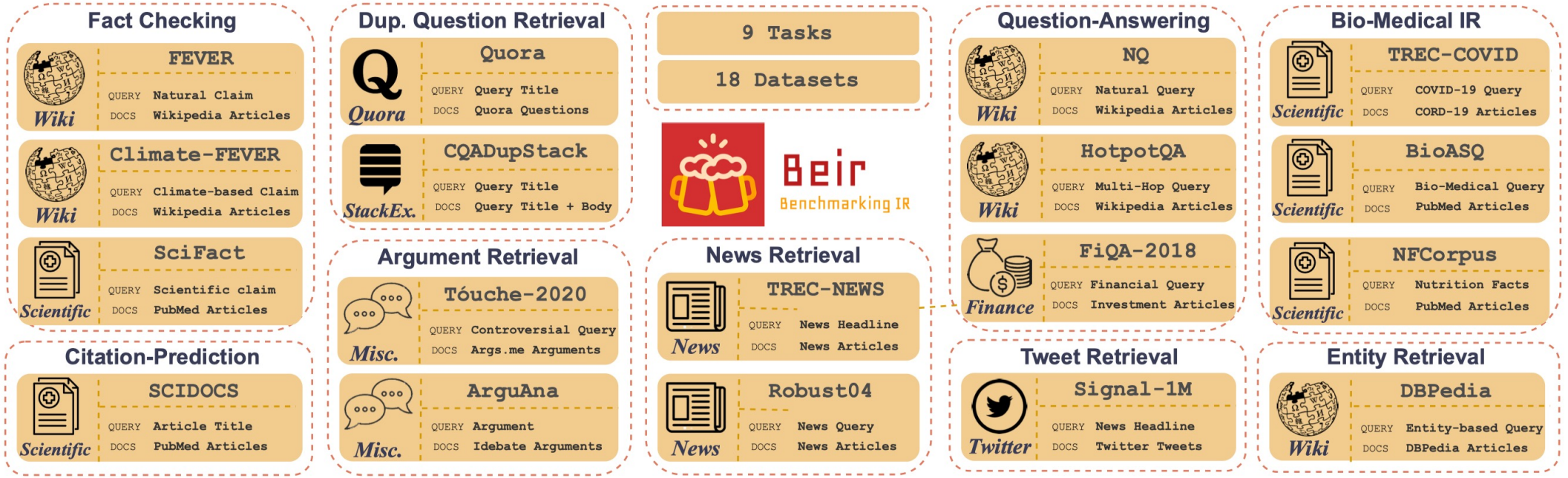


(a) place-of-birth

Sciavolino et al. (2021). Simple Entity-Centric Questions Challenge Dense Retrievers. *EMNLP*.

Challenges: Zero-shot Retrieval

- BEIR: Zero-shot Retrieval Benchmark (21NeuIPS, UKP Lab)
  - Organized a robust and heterogeneous benchmark comprising of 18 retrieval datasets for comparison and evaluation
  - diverse text domains, broad topics, different query lengths (between 3 and 192 words) and document lengths (between 11 and 635 words)



Thakur et al. (2021). BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models. NeuIPS.

Challenges: Zero-shot Retrieval

Tencent AI Lab

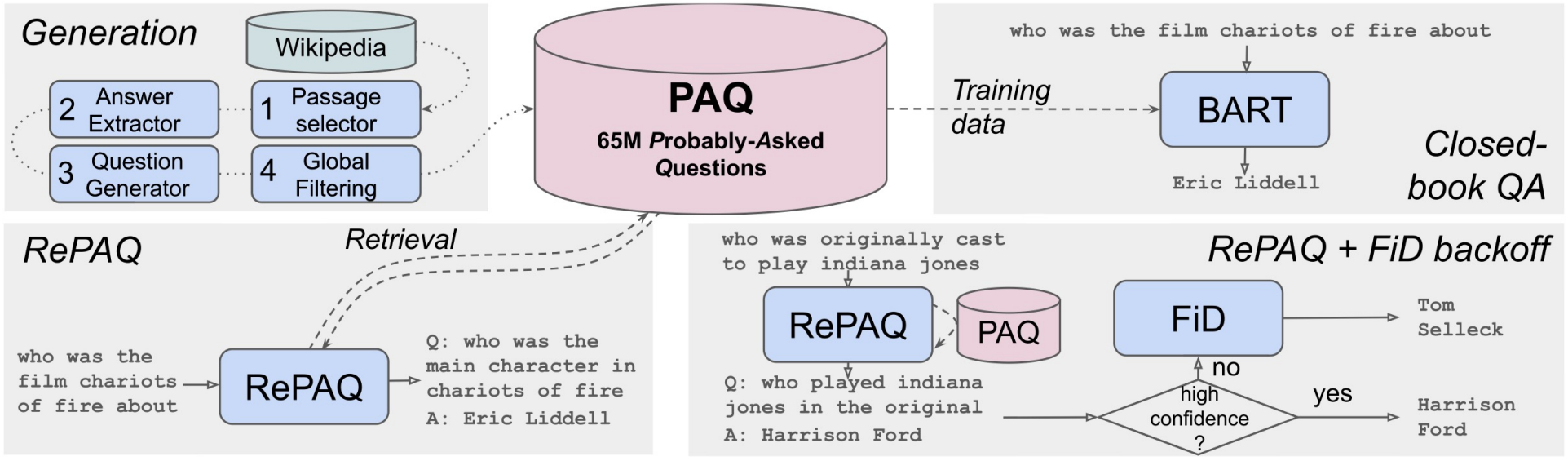
- BEIR: Zero-shot Retrieval Benchmark (21NeuIPS, UKP Lab)
  - In-domain performance is not a good indicator for out-of-domain generalization.
  - Dense retrieval models with issues for out-of-distribution data.
  - Tradeoff between performance and retrieval latency

Model (→)	Lexical	Sparse			Dense				Late-Interaction	Re-ranking
Dataset (↓)	**BM25**	**DeepCT**	**SPARTA**	**docT5query**	**DPR**	**ANCE**	**TAS-B**	**GenQ**	**ColBERT**	**BM25+CE**
MS MARCO	0.228	0.296‡	0.351‡	0.338‡	0.177	0.388‡	0.408‡	0.408‡	**0.425**‡	0.413‡
TREC-COVID	0.656	0.406	0.538	0.713	0.332	0.654	0.481	0.619	0.677	**0.757**
BioASQ	0.465	0.407	0.351	0.431	0.127	0.306	0.383	0.398	0.474	**0.523**
NFCorpus	0.325	0.283	0.301	0.328	0.189	0.237	0.319	0.319	0.305	**0.350**
NQ	0.329	0.188	0.398	0.399	0.474‡	0.446	0.463	0.358	0.524	**0.533**
HotpotQA	0.603	0.503	0.492	0.580	0.391	0.456	0.584	0.534	0.593	**0.707**
FiQA-2018	0.236	0.191	0.198	0.291	0.112	0.295	0.300	0.308	0.317	**0.347**
Signal-1M (RT)	0.330	0.269	0.252	0.307	0.155	0.249	0.289	0.281	0.274	**0.338**
TREC-NEWS	0.398	0.220	0.258	0.420	0.161	0.382	0.377	0.396	0.393	**0.431**
Robust04	0.408	0.287	0.276	0.437	0.252	0.392	0.427	0.362	0.391	**0.475**

Thakur et al. (2021). BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models. NeuIPS.

Attempts: Synthetic data generation

- Large-scale Query Generation (21TACL, Facebook)
  - PAQ (Probably-Asked Questions), automatically constructed using a question generation model and Wikipedia.
  - The synthetic data can be used for training both open-book and closed-book QA
  - The query generation pipeline includes: passage selection model; answer extraction; question generator, and filtering



Lewis et al. (2021). PAQ: 65 Million Probably-Asked Questions and What You Can Do With Them. TACL.

Attempts: Synthetic data generation

- Large-scale Query Generation (21TACL, Facebook)
  - PAQ (Probably-Asked Questions), automatically constructed using a question generation model and Wikipedia.
  - The synthetic data can be used for training both open-book and closed-book QA
  - Further address open-domain QA via question matching: directly using QA pair seen during training

Input: who was the film chariots of fire about	**A**: Eric Liddell	
who was the main character in chariots of fire	**A**: *Eric Liddell*	✓
who starred in the movie chariots of fire	**A**: Ian Charleson	✗
which part did straan rodger play in chariots of fire	**A**: Sandy McGrath	✗
who played harold in the 1981 film chariots of fire	**A**: Ben Cross	✗
who is the main character in chariots of fire	**A**: Eric Liddell	✓
Input: what is the meaning of the name didymus	**A**: twin	
what language does the name didymus come from	**A**: Greek	✗
where does the name didymus come from in english	**A**: Greek	✗
what does the word domus mean in english	**A**: home	✗
how long has the term domus been used	**A**: 1000s of years	✗
what does the greek word didyma mean	**A**: *twin*	✓
Input: what is the name of a group of llamas	**A**: herd	
what are llamas and alpacas considered to be	**A**: domesticated	✗
what are the figures of llamas in azapa valley	**A**: Atoca	✗
what are the names of the llamas in azapa valley	**A**: Atoca	✗
what is the scientific name for camels and llamas	**A**: *Camelidae*	✗
are llamas bigger or smaller than current forms	**A**: larger	✗

Table 6: Examples of top 5 retrieved QA-pairs for NQ. Italics indicate QA-pairs chosen by reranker.

Lewis et al. (2021). PAQ: 65 Million Probably-Asked Questions and What You Can Do With Them. TACL.

Attempts: Synthetic data generation

Tencent AI Lab

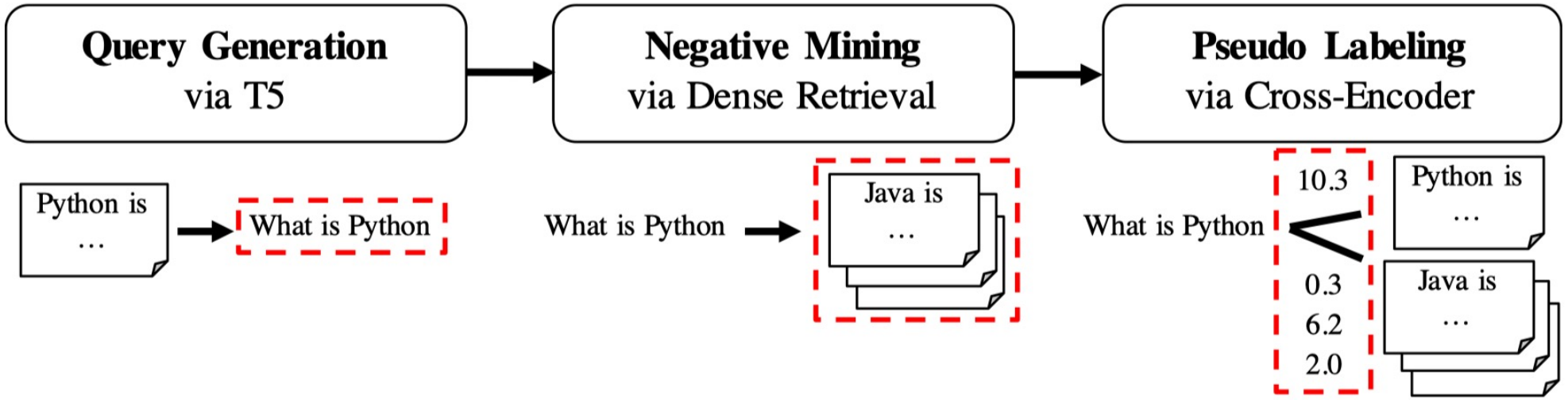
- Large-scale Query Generation (21TACL, Facebook)
  - PAQ (Probably-Asked Questions), automatically constructed using a question generation model and Wikipedia.
  - Experiments

#	Model Type	Model	NaturalQuestions
1	Closed-book	T5-11B-SSM (Roberts et al., 2020)	35.2
2	Closed-book	BART-large (Lewis et al., 2021)	26.5
3	QA-pair retriever	Dense retriever (Lewis et al., 2021)	26.7
4	Open-book, retrieve-and-read	RAG-Sequence (Lewis et al., 2020b)	44.5
5	Open-book, retrieve-and-read	FiD-large, 100 docs (Izacard and Grave, 2021)	51.4
6	Open-book, phrase index	DensePhrases (Lee et al., 2021)	40.9
7	Closed-book	BART-large, pre-finetuned on PAQ	32.7
8	QA-pair retriever	RePAQ (retriever only)	41.2
9	QA-pair retriever	RePAQ (with reranker)	47.7
10	QA-pair retriever	RePAQ-multitask (retriever only)	41.7
11	QA-pair retriever	RePAQ-multitask (with reranker)	47.6
12	QA-pair retriever	RePAQ-multitask w/ FiD-Large Backoff	**52.3**

Lewis et al. (2021). PAQ: 65 Million Probably-Asked Questions and What You Can Do With Them. TACL.

Attempts: Synthetic data generation

- Fined-Grained Pseudo Labeling (21Arxiv, UKP Lab)
  - Query Generator used T5 model trained on MS-MARCO
  - The generated synthetic data are used to provide pairwise supervision, instead of directly used for retriever learning
  - This technique is claimed to alleviate false positive issues.



Wang et al. (2021). GPL: Generative Pseudo Labeling for Unsupervised Domain Adaptation of Dense Retrieval. ArXiv.

Attempts: Synthetic data generation

Tencent AI Lab

- Fined-Grained Pseudo Labeling (21Arxiv, UKP Lab)
  - Query Generator used T5 model trained on MS-MARCO
  - The proposed GPL outperforms Qgen (direct data augmentation), and can be combined with other target-domain approaches (e.g., Inverse Cloze Task, denoising auto-encoding)

Method \ Dataset	FiQA	SciFact	BioASQ	TRECC.	CQADup.	Robust04	Avg.
Zero-Shot Models							
MS MARCO	26.7	57.1	52.9	66.1	29.6	39.0	45.2
PAQ	15.2	53.3	44.0	23.8	24.5	31.9	32.1
PAQ + MS MARCO	26.7	57.6	53.8	63.4	30.6	37.2	44.9
$\text{TSDAE}_{\text{MS MARCO}}$	26.7	55.5	51.4	65.6	30.5	36.6	44.4
BM25	23.9	66.1	70.7	60.1	31.5	38.7	48.5
Generation-based Domain Adaptation (Previous State-of-the-Art)							
QGen	28.2	61.7	60.0	72.8	33.6	38.5	49.1
QGen (w/ Hard Negatives)	26.0	59.6	57.7	65.0	33.2	36.5	46.3
TSDAE + QGen (Ours)	30.3	64.7	60.5	73.8	**35.1**	38.4	50.5
Proposed Method: Generative Pseudo Labeling							
GPL	33.1	65.2	61.6	71.7	34.4	42.1	51.4
TSDAE + GPL	**33.3**	**67.3**	**62.8**	**74.0**	**35.1**	**42.1**	**52.4**

Wang et al. (2021). GPL: Generative Pseudo Labeling for Unsupervised Domain Adaptation of Dense Retrieval. ArXiv.

Potentials: Diversified Query Generation

Tencent AI Lab

- Diversification (20ACL, IBM)
  - The knowledge and query usually have many-to-many characteristics. This work focus on data augmentation for QA, instead of retrieval;
  - Diverse yet accurate samples might yield better QA results than the current approach of optimizing for the "most likely" question;
  - The diversified examples are generated using nucleus sampling

On Tesla's 75th birthday in 1931, Time magazine put him on its cover. The cover caption "All the world's his power house" noted his contribution to electrical power generation. He received congratulatory letters from more than 70 pioneers in science and engineering, including Albert Einstein.

- *Who appeared on Time magazine's cover on his 75th birthday?*
- *Which famous scientist was in the cover of Time Magazine in 1931?*
- *Which mad scientist received more than a 70 people congratulating him on his birthday?*
- *What famous scientist was also 75?*

Figure 1: A passage with an underlined answer span ("Tesla"), and corresponding questions generated by our model. The generated questions exhibit both lexical and factual diversity.

Sultan et al. (2020). On the Importance of Diversity in Question Generation for QA. *ACL*.

Potentials: Diversified Query Generation

Tencent AI Lab

- Diversification
  - The diversified examples are generated using nucleus sampling
  - Diversity-promoting question generation increase downstream QA performances
  - QG metrics that only reward similarity with GT are negatively correlated with diversity

%train	generator	B_1	R_4	MT	QA F_1	B_1	R_4	MT	QA F_1
		base model				*large* model			
50	$b = 5$	**39.1**	**11.9**	**44.4**	82.8	**40.6**	**12.6**	**45.4**	84.3
	$p = .1$	37.8	10.3	43.4	83.6	39.6	11.2	44.7	84.8
	$p = .5$	37.4	10.0	42.9	83.8	39.4	11.1	44.4	84.9
	$p = .75$	35.4	8.8	40.2	84.3	38.2	10.3	42.8	85.3
	$p = .95$	31.4	6.3	35.2	**84.8**	33.6	7.5	37.2	**85.7**
100	$b = 5$	**40.3**	**12.6**	**45.8**	83.6	**41.6**	**13.4**	**46.7**	84.5
	$p = .1$	38.9	11.0	44.6	83.9	40.6	12.1	46.1	84.9
	$p = .5$	38.5	10.7	44.1	84.3	40.3	11.9	45.7	85.0
	$p = .75$	36.7	9.6	41.7	84.8	38.8	10.8	43.7	85.5
	$p = .95$	32.5	6.9	36.4	**85.3**	34.4	7.6	38.3	**86.1**

Sultan et al. (2020). On the Importance of Diversity in Question Generation for QA. *ACL*.

Potentials: Compositional Generalization

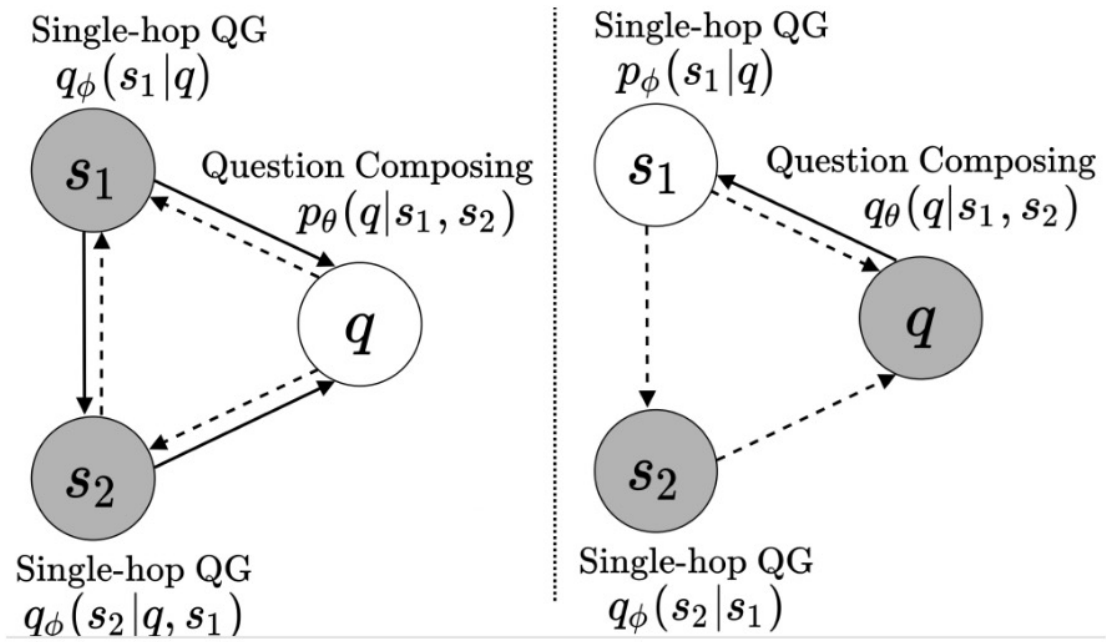
- Style transfer for compositional generalization (21ACL Finding)
  - Multi-hop questions requires identifying and aggregating information from multiple documents to derive the question
  - Modeling the reasoning process by decomposing a multi-hop question into several sub-questions

Supportive Evidence	**Paragraph A. Dario Franchitti**
	After Franchitti was contracted by the AMG team to compete in touring cars in the DTM and its successor — International Touring Car Championship.
	Paragraph B. Mercedes-AMG
	Mercedes-AMG is headquartered in Affalterbach, Germany.
Groundtruth QA Pair	After he was contracted by the team that is headquartered in Affalterbach, Germany, Dario Franchitti competed in what series? (Answer: DTM)

Huang et al. (2021). Latent Reasoning for Low-Resource Question Generation. *ACL Findings*.

Potentials: Compositional Generalization

- Style transfer for compositional generalization (21ACL Finding)
  - Introduced non-parallel single-hop corpuses for joint optimization.
  - Single-hop questions regularize the behavior of sub-question generation;
  - We model single- and multi-hop QG in a unified way via backtranslation.



Huang et al. (2021). Latent Reasoning for Low-Resource Question Generation. *ACL Findings*.

Take away

Tencent AI Lab

1. The retriever is essential for knowledge-intensive tasks, while the gap between "query" and "documents" is not well addressed’;

2. Query generation is a promising direction to enhance retriever generalization abilities, while existing attempts mainly focus on question-type generation;

3. The task formulation and learning objectives of query generation model is under-explored, where diversification and compositional generalization could have potentials;

4. The ultimate goal is to utilize "universal knowledge" for the tasks of interest.