CS11-737 Multilingual NLP

Multilingual Question Answering

Vijay Viswanathan and Graham Neubig

http://phontron.com/class/multiling2022/

Carnegie Mellon University
Language Technologies Institute

Question Answering

➔ Humans have dreamed of asking questions to computers

Image Source: The Hitchhiker's Guide to the Galaxy (2005)

Question Answering

➔ Real examples of question answering systems and tasks you may know

Question Answering

- Real examples of question answering systems and tasks you may know
  - LUNAR

Question Answering

- ➔ Real examples of question answering systems and tasks you may know
  - ◆ LUNAR
  - ◆ IBM Watson

Question Answering

- Real examples of question answering systems and tasks you may know
  - LUNAR
  - IBM Watson
  - SQuAD

Question Answering

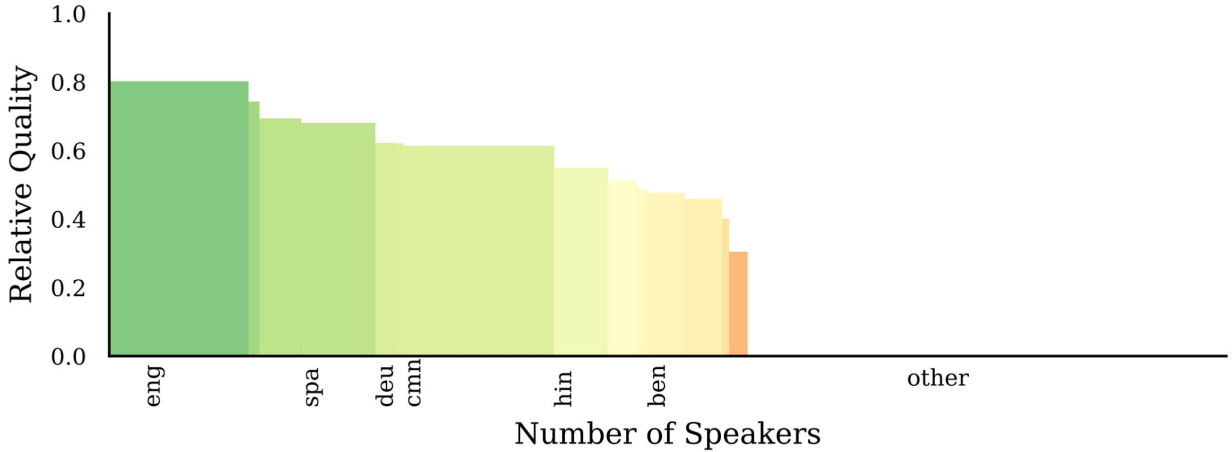
- ➔ Real examples of question answering systems and tasks you may know
  - LUNAR
  - IBM Watson
  - SQuAD
- ➔ **All English-language systems and tasks**

Question Answering

- ➔ Real examples of question answering systems and tasks you may know
  - ◆ LUNAR
  - ◆ IBM Watson
  - ◆ SQuAD
- ➔ **All English-language systems and tasks**
- ➔ Raise your hand if you have tried using a digital assistant to answer questions in a language other than English

Multilingual Question Answering

➔ How to help *all people* access information easily?



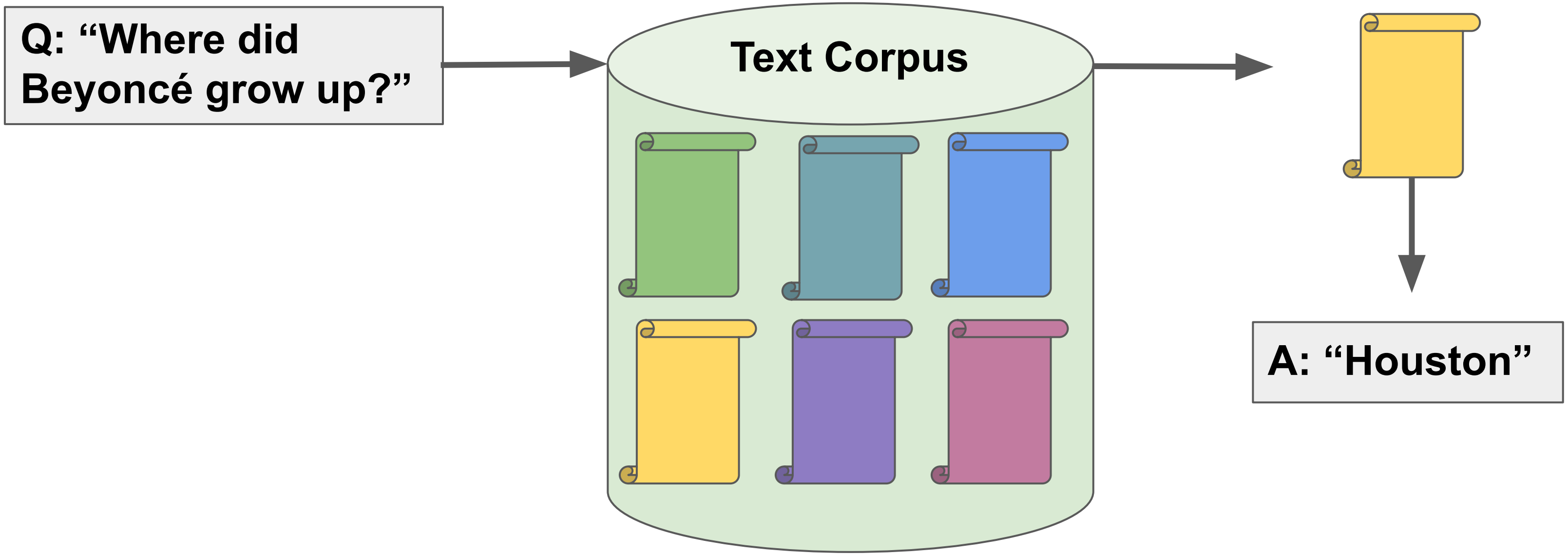
Systematic Inequalities in Language Technology Performance across the World's Languages. Blasi et al. 2021.

Question Answering: Two Kinds

1. Open-Retrieval Question Answering (aka Open Domain Question Answering)

Question Answering: Two Kinds

1. Open-Retrieval Question Answering (aka Open Domain Question Answering)
    - Given a question, find an answer (if one exists) from a text corpus

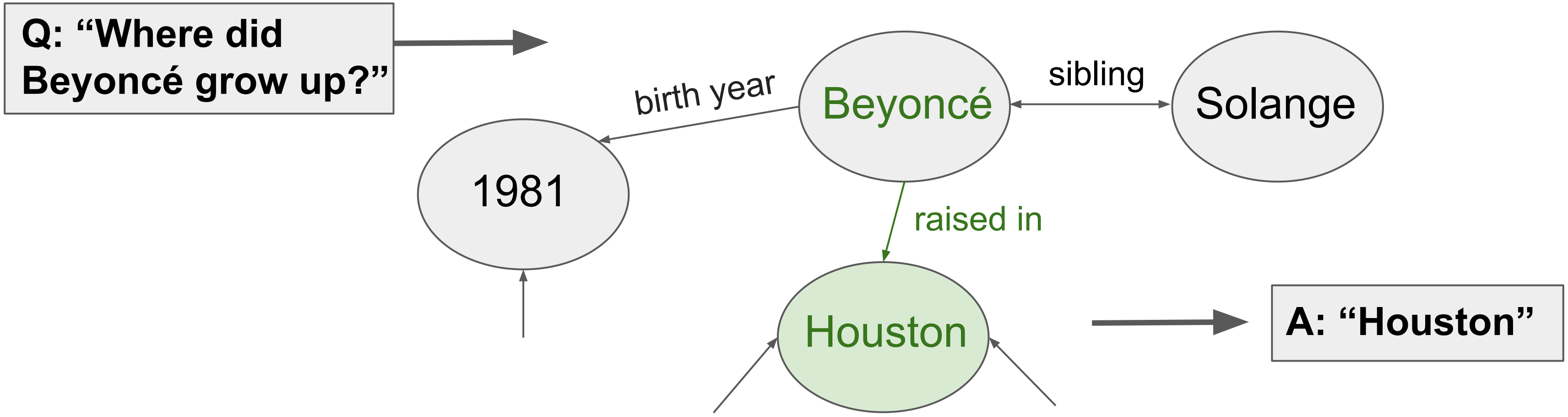


Question Answering: Two Kinds

1. Open-Retrieval Question Answering (aka Open Domain Question Answering)
   - Given a question, find an answer (if one exists) from a text corpus
2. Knowledge Graph Question Answering

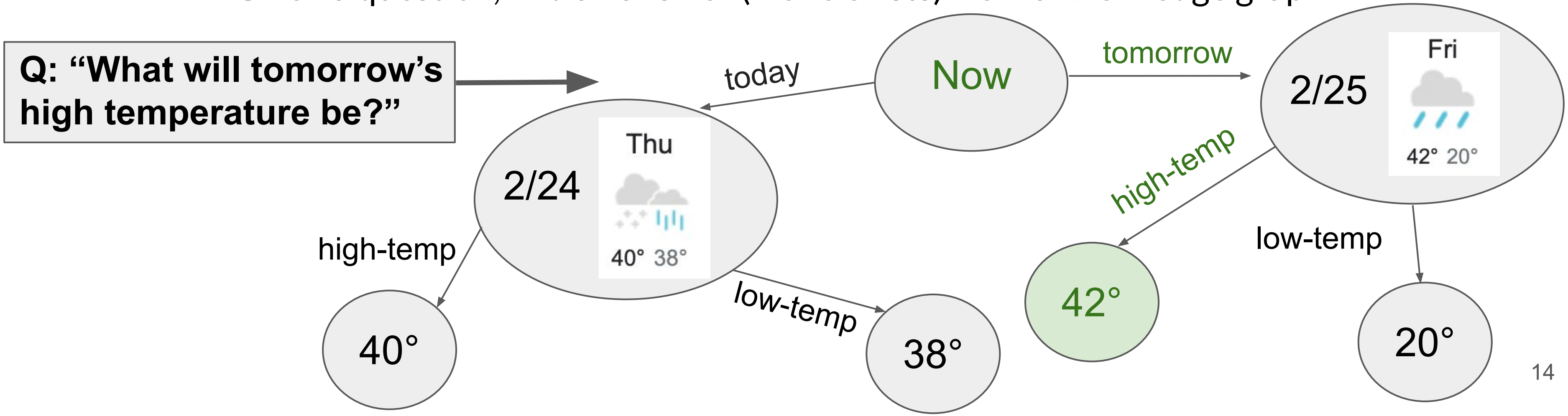
Question Answering: Two Kinds

1. Open-Retrieval Question Answering (aka Open Domain Question Answering)
   - Given a question, find an answer (if one exists) from a text corpus
2. Knowledge Graph Question Answering
   - Given a question, find an answer (if one exists) from a knowledge graph



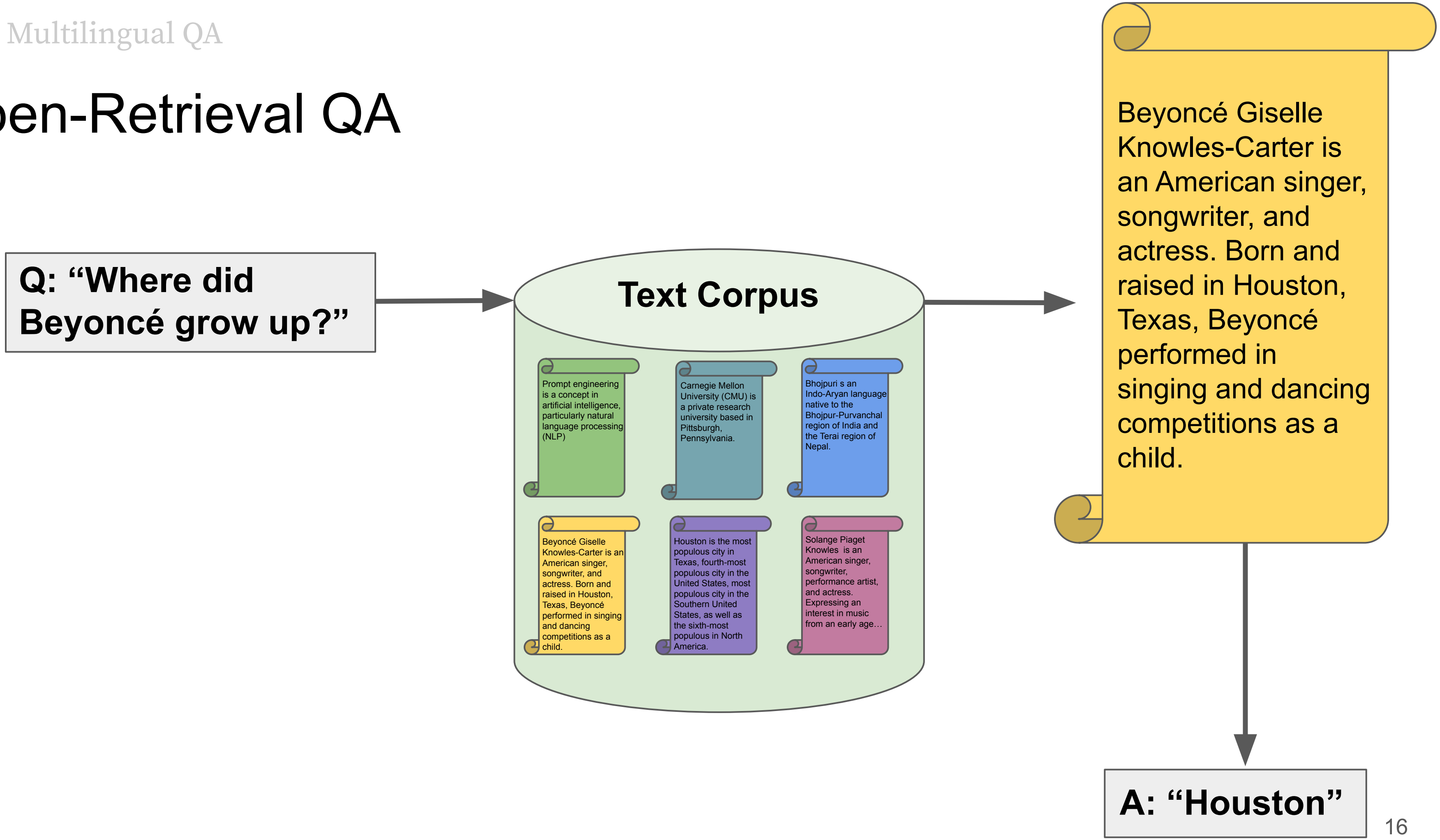
Question Answering: Two Kinds

1. Open-Retrieval Question Answering (aka Open Domain Question Answering)
   - Given a question, find an answer (if one exists) from a text corpus
2. Knowledge Graph Question Answering
   - Given a question, find an answer (if one exists) from a knowledge graph

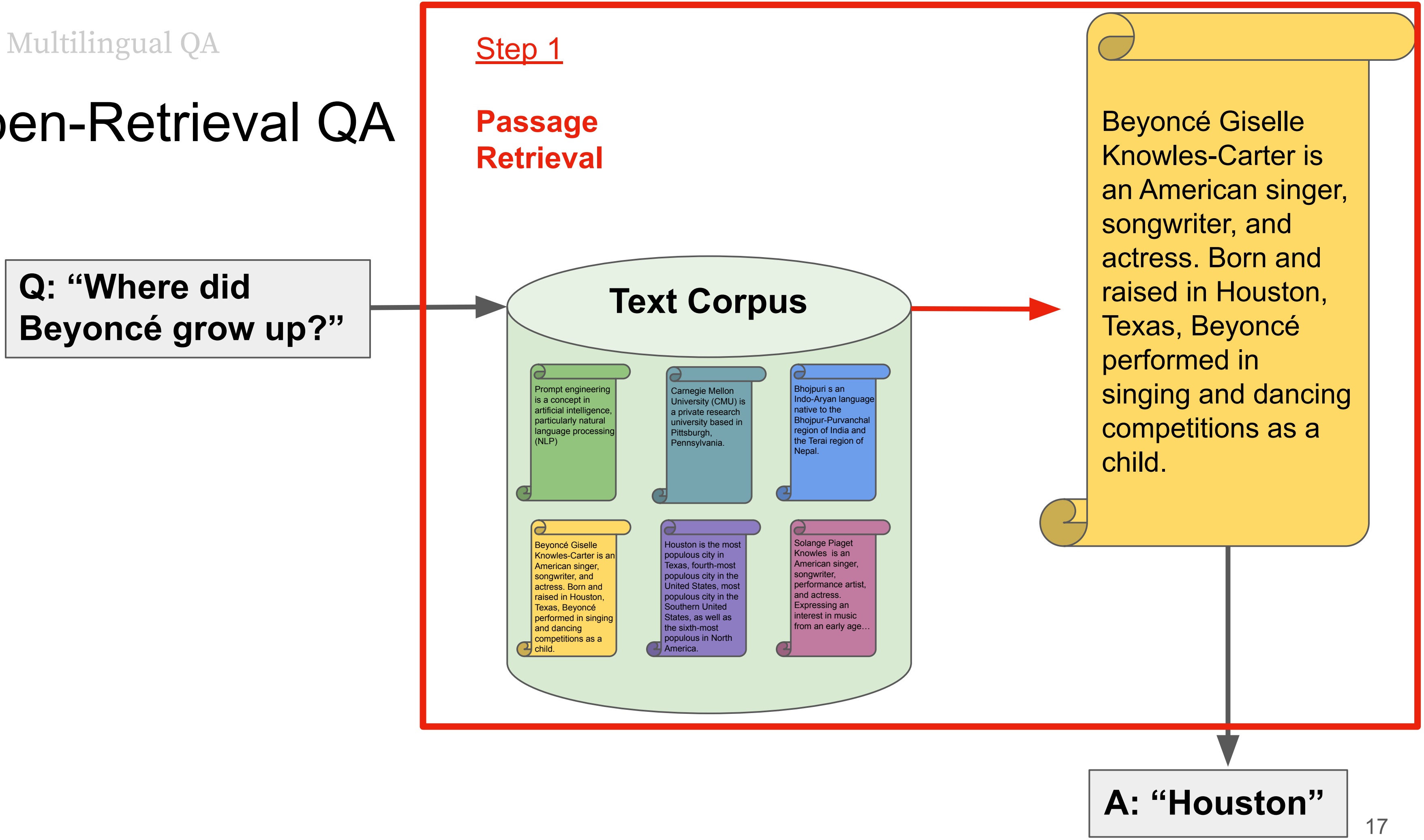


Open-Retrieval QA

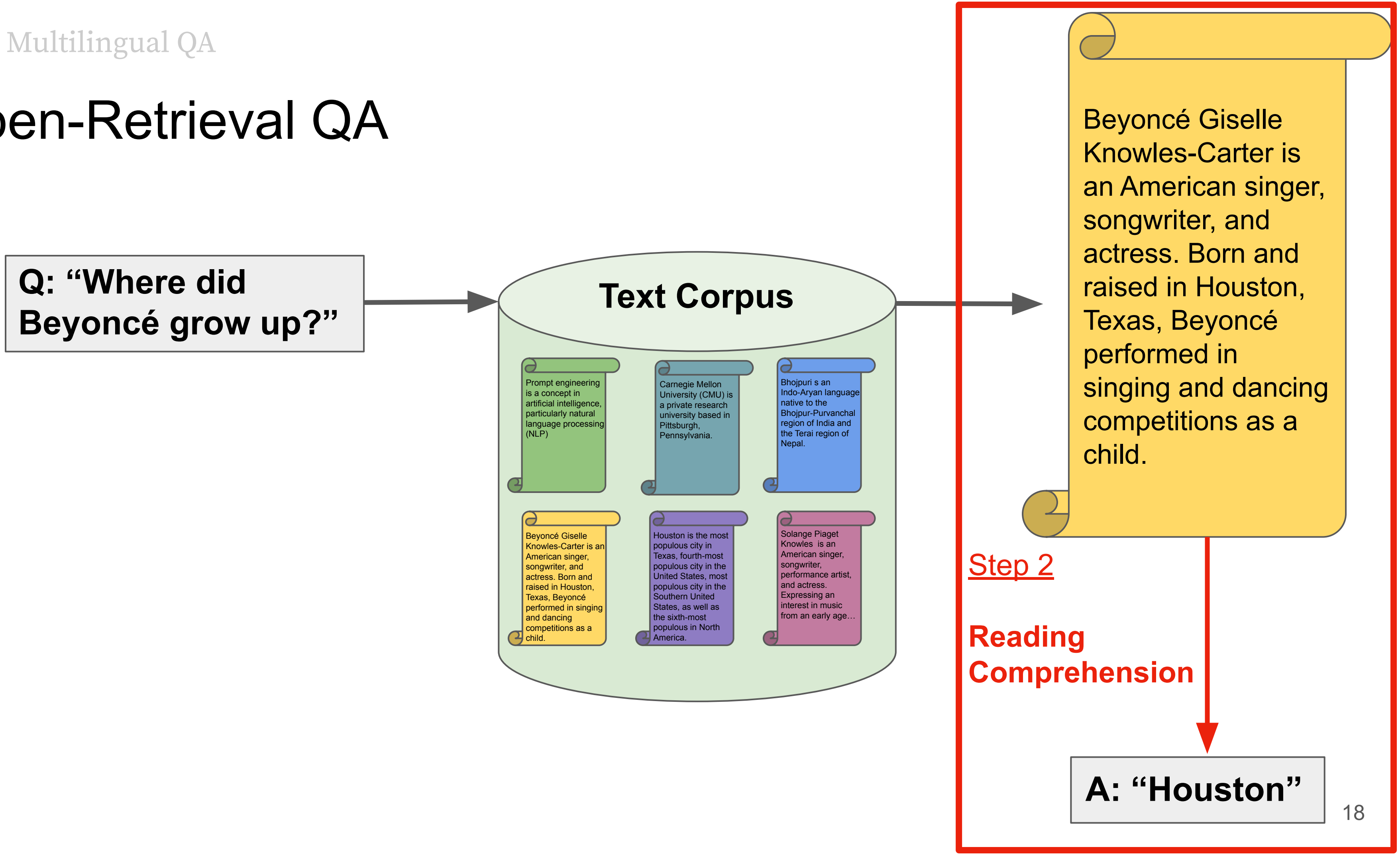
Open-Retrieval QA



Open-Retrieval QA



Open-Retrieval QA



Passage Retrieval

- Given a query q and a collection of documents $D = \{d_1, \ldots, d_N\}$, return the most relevant document $d \in D$ for q
  - *Challenge*: How to define *score(q, d)*

Passage Retrieval: tf-idf

- Classic approach (Spärck Jones 1972)
- Two ingredients:
  - Term frequency (TF)
    - how often term appears in a document
  - Inverse document frequency (IDF)
    - how many total documents contain a term
- $$\text{score}(q, d) = \sum_{t \in q} \frac{tf_{t,d}}{df_t}$$

Passage Retrieval: BM25

- ➔ More complex variant of tf-idf (Robertson and Walker 1994)

- ➔ $$\text{score}(q,d) = \sum_{t \in q} \log\left(\frac{N}{df_t}\right) \frac{tf_{t,d}}{k\left(1 - b + b\left(\frac{|d|}{|d_{\text{avg}}|}\right)\right) + tf_{t,d}}$$

Speech and Language Processing, 3rd edition (draft). Jurafsky and Martin 2021.

Passage Retrieval: BM25

- ➔ More complex variant of tf-idf (Robertson and Walker 1994)

IDF Term Weight

- ➔ $$\text{score}(q,d) = \sum_{t \in q} \log\left(\frac{N}{df_t}\right) \frac{tf_{t,d}}{k\left(1 - b + b\left(\frac{|d|}{|d_{\text{avg}}|}\right)\right) + tf_{t,d}}$$

Speech and Language Processing, 3rd edition (draft). Jurafsky and Martin 2021.

Passage Retrieval: BM25

➔ More complex variant of tf-idf (Robertson and Walker 1994)

IDF Term Weight

➔ $$\text{score}(q,d) = \sum_{t \in q} \log\left(\frac{N}{df_t}\right) \frac{tf_{t,d}}{k\left(1 - b + b\boxed{\left(\frac{|d|}{|d_{\text{avg}}|}\right)}\right) + tf_{t,d}}$$

Document Length Normalization

Speech and Language Processing, 3rd edition (draft). Jurafsky and Martin 2021.

Passage Retrieval: BM25

➔ More complex variant of tf-idf (Robertson and Walker 1994)

➔ $$\text{score}(q,d) = \sum_{t \in q} \log\left(\frac{N}{df_t}\right) \frac{tf_{t,d}}{k\left(1 - b + b\left(\frac{|d|}{|d_{\text{avg}}|}\right)\right) + tf_{t,d}}$$

Speech and Language Processing, 3rd edition (draft). Jurafsky and Martin 2021.

Passage Retrieval: BM25

➔ More complex variant of tf-idf (Robertson and Walker 1994)

➔ $$\text{score}(q,d) = \sum_{t \in q} \log\left(\frac{N}{df_t}\right) \frac{tf_{t,d}}{k\left(1 - b + b\left(\frac{|d|}{|d_{\text{avg}}|}\right)\right) + tf_{t,d}}$$

➔ Like tf-idf, requires no supervision

Speech and Language Processing, 3rd edition (draft). Jurafsky and Martin 2021.

Passage Retrieval: BM25

	Unsupervised		
BM25	18.4	62.9, 78.3	76.4, 83.2
ICT	-	50.9, 66.8	57.5, 73.6
MSS	-	59.8, 74.9	68.2, 79.4
Contriever	-	67.2, 81.3	74.2, 83.2
cpt-text S	19.9	65.5, 77.2	75.1, 81.7
cpt-text M	20.6	68.7, 79.6	78.0, 83.8
cpt-text L	21.5	73.0, 83.4	80.0, 86.8
cpt-text XL	**22.7**	**78.8, 86.8**	**82.1, 86.9**

Text and Code Embeddings by Contrastive Pre-Training. Neelakantan et. al. 2021.

Passage Retrieval: BM25

Cost to encode 6M articles:
$17,000
$25,000
$126,000
$1,260,000

	Unsupervised		
BM25	18.4	62.9, 78.3	76.4, 83.2
ICT	-	50.9, 66.8	57.5, 73.6
MSS	-	59.8, 74.9	68.2, 79.4
Contriever	-	67.2, 81.3	74.2, 83.2
`cpt-text` S	19.9	65.5, 77.2	75.1, 81.7
`cpt-text` M	20.6	68.7, 79.6	78.0, 83.8
`cpt-text` L	21.5	73.0, 83.4	80.0, 86.8
`cpt-text` XL	**22.7**	**78.8, 86.8**	**82.1, 86.9**

Text and Code Embeddings by Contrastive Pre-Training. Neelakantan et. al. 2021.

Dense Passage Retrieval

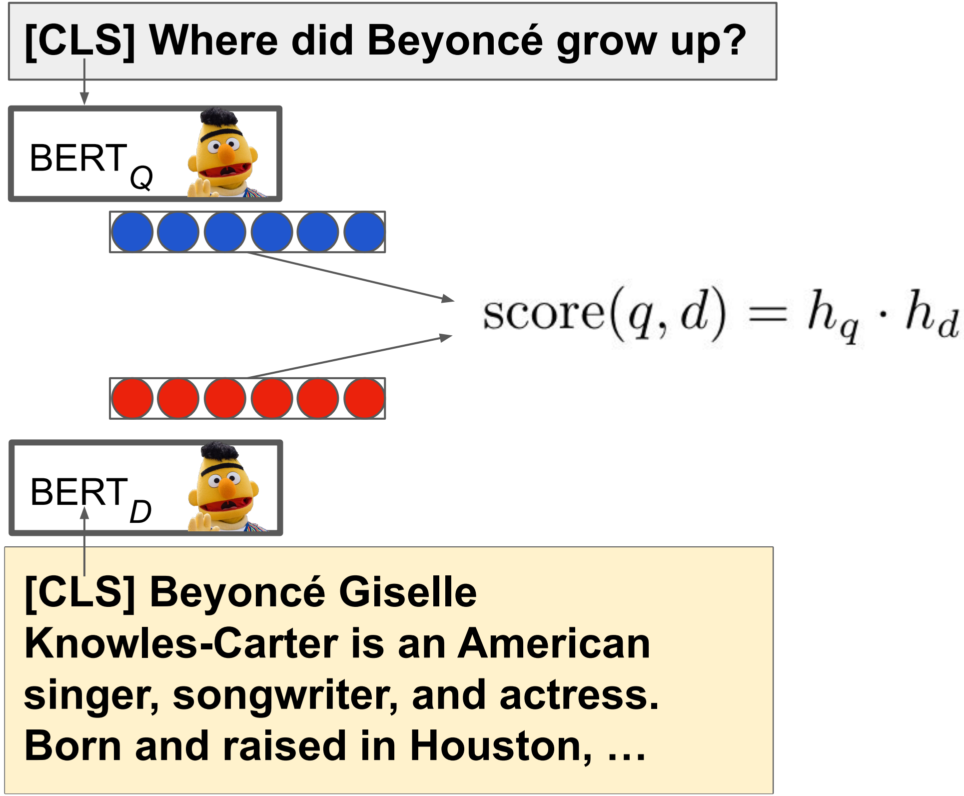
- Use two deep encoders
- $$h_q = \text{cls}(\texttt{BERT}_Q(q))$$
  $$h_d = \text{cls}(\texttt{BERT}_D(d))$$

Dense Passage Retrieval for Open-Domain Question Answering. Karpukhin et al. 2020.
Speech and Language Processing, 3rd edition (draft). Jurafsky and Martin 2021.

Dense Passage Retrieval

- ➔ Use two deep encoders
- ➔ $h_q = \text{cls}(\text{BERT}_Q(q))$

$$h_d = \text{cls}(\text{BERT}_D(d))$$



Dense Passage Retrieval for Open-Domain Question Answering. Karpukhin et al. 2020.
Speech and Language Processing, 3rd edition (draft). Jurafsky and Martin 2021.

Dense Passage Retrieval

- ➔ Use two deep encoders
- ➔ $$h_q = \text{cls}(\text{BERT}_Q(q))$$
  $$h_d = \text{cls}(\text{BERT}_D(d))$$

$$\text{score}(q, d) = h_q \cdot h_d$$

- ➔ Index search corpus offline
- ➔ For a new query:
  - ◆ Encode the query using BERT
  - ◆ Perform “maximum inner product search” against search corpus

Dense Passage Retrieval for Open-Domain Question Answering. Karpukhin et al. 2020.
Speech and Language Processing, 3rd edition (draft). Jurafsky and Martin 2021.

Dense Passage Retrieval

- ➔ Use two deep encoders
- ➔ $$h_q = \mathrm{cls}(\mathrm{BERT}_Q(q))$$
  $$h_d = \mathrm{cls}(\mathrm{BERT}_D(d))$$
  $$\mathrm{score}(q, d) = h_q \cdot h_d$$
- ➔ Trained on a set of questions tagged with known relevant and non-relevant documents
  - Encoders are fine-tuned on this dataset
  - Large dataset is required, creating a challenge for low-resource languages

Dense Passage Retrieval for Open-Domain Question Answering. Karpukhin et al. 2020.
Speech and Language Processing, 3rd edition (draft). Jurafsky and Martin 2021.

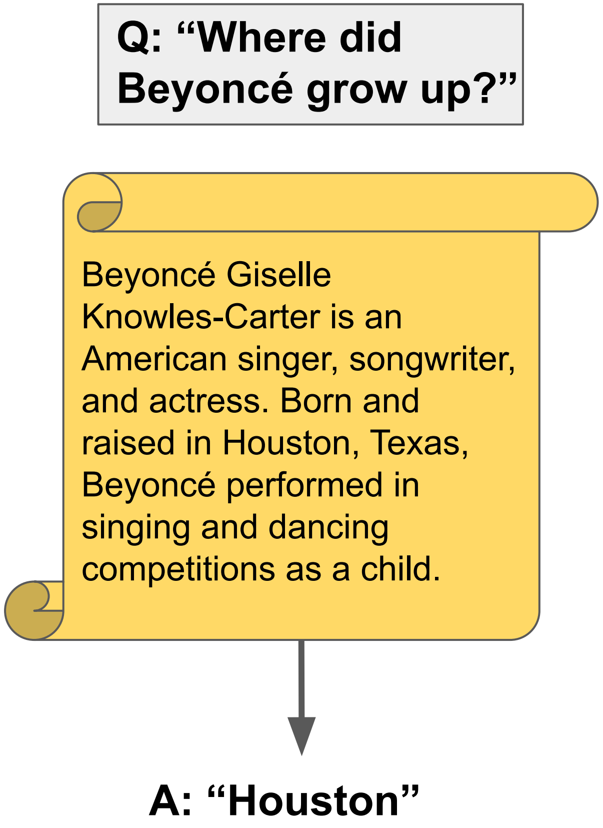
Dense Passage Retrieval

“Natural Questions” Dataset Leaderboard

Rank	Model	Precision@100	Precision@20	Extra Training Data	Paper
1	**DPR-PAQ**	89.22	84.68	✓	Domain-matched Pre-training Tasks for Dense Retrieval
2	**RocketQA**	88.5	82.7	×	RocketQA: An Optimized Training Approach to Dense Passage Retrieval for Open-Domain Question Answering
3	**DPR+ELECTRA-large-extreader-reranker**	88.25	85.26	×	R2-D2: A Modular Baseline for Open-Domain Question Answering
4	**DPR+RoBERTa-base-crossencoder-reranker**	88.03	84.46	×	R2-D2: A Modular Baseline for Open-Domain Question Answering
5	**ANCE**	87.5	81.9	×	Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval
6	**DPR**	86	79.4	×	Dense Passage Retrieval for Open-Domain Question Answering
7	**BM25+RM3**	79.6	64.2	×	Generation-Augmented Retrieval for Open-domain Question Answering

Natural Questions: a Benchmark for Question Answering Research. Kwiatkowski et al. 2019.
Dense Passage Retrieval for Open-Domain Question Answering. Karpukhin et al. 2020.
Speech and Language Processing, 3rd edition (draft). Jurafsky and Martin 2021.

Reading Comprehension

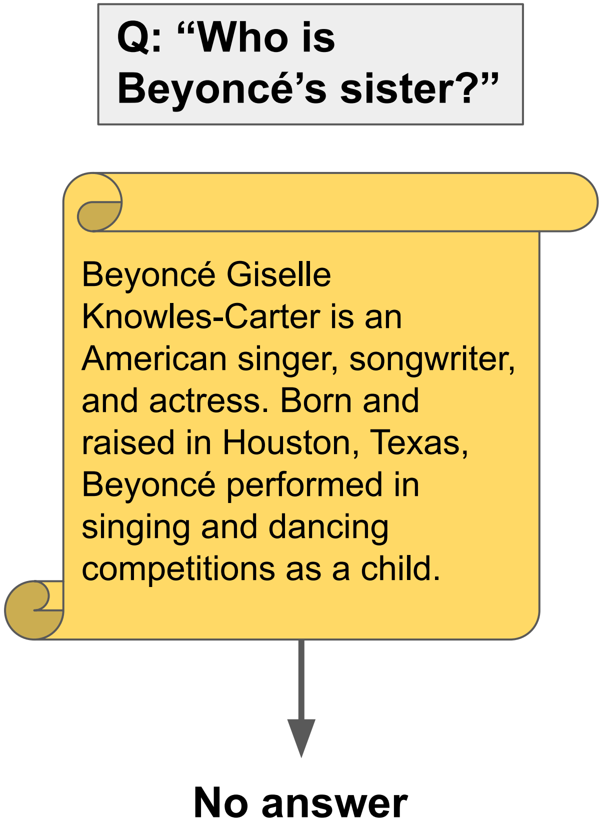


- ➔ Given a question *q* and a passage *p*, return an answer *s*
  - ◆ (or determine that no answer exists in the passage)

Speech and Language Processing, 3rd edition (draft). Jurafsky and Martin 2021.

Reading Comprehension

- ➔ Given a question *q* and a passage *p*, return an answer *s*
  - ◆ (or determine that no answer exists in the passage)

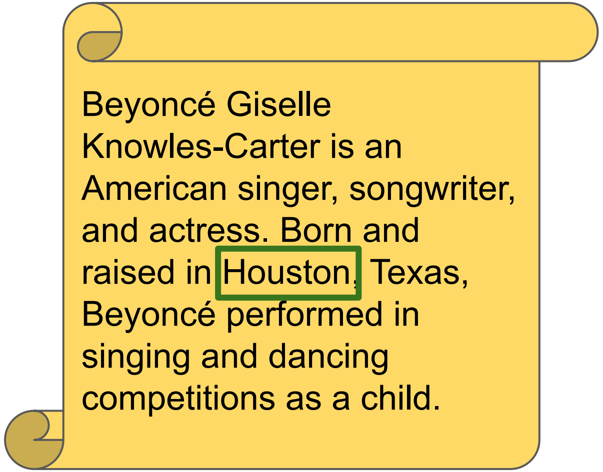


Speech and Language Processing, 3rd edition (draft). Jurafsky and Martin 2021.

Reading Comprehension

Q: “Where did Beyoncé grow up?”

- Two kinds of reading comprehension:
  a. Extractive QA: answer is assumed to be a *span* in the passage *p*

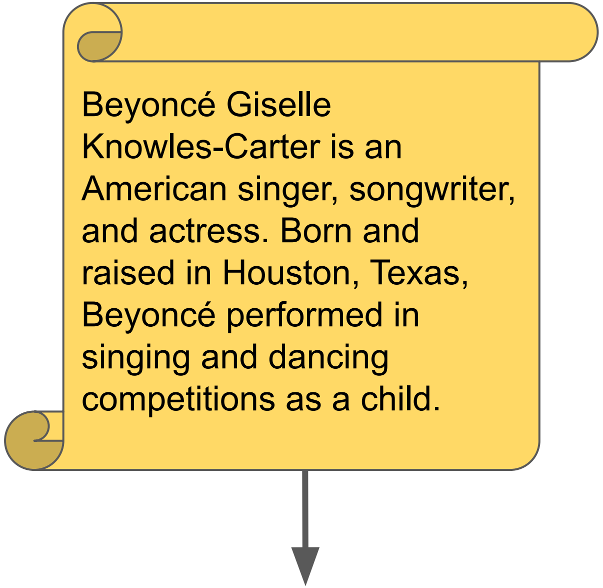


Speech and Language Processing, 3rd edition (draft). Jurafsky and Martin 2021.

Reading Comprehension

- Two kinds of reading comprehension:
    a. Extractive QA: answer is assumed to be a *span* in the passage *p*
    b. Generative QA: answer is freely generated, given the passage *p*

Q: "Is Beyoncé from the Southern United States?"



A: "Yes"

Speech and Language Processing, 3rd edition (draft). Jurafsky and Martin 2021.

Reading Comprehension: Extractive Baseline

- ➔ Concatenate the question and passage
- ➔ Encode each token in the passage
- ➔ Predict P(start) and P(end) at each token (*locally normalized*)

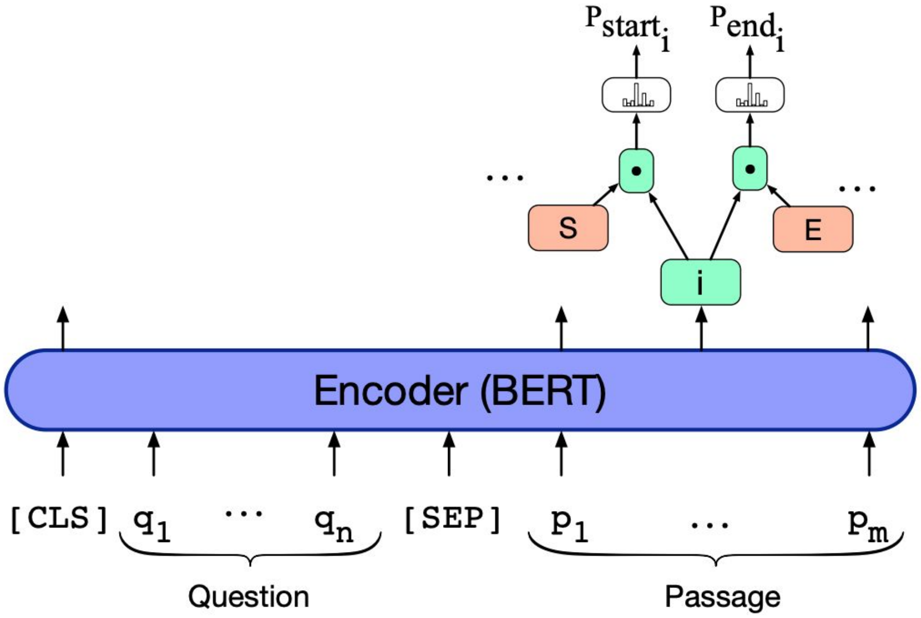


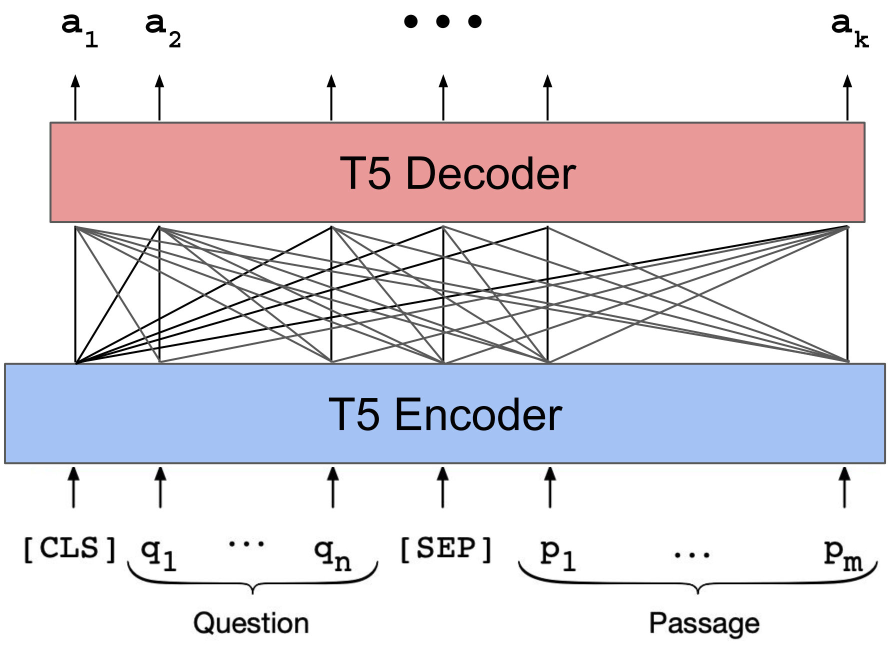
Diagram from Jurafsky and Martin, 3rd ed.

A BERT Baseline for the Natural Questions. Alberti, Lee, and Collins 2019.
Speech and Language Processing, 3rd edition (draft). Jurafsky and Martin 2021.

Reading Comprehension: Generative Baseline

- ➔ Concatenate the question and passage
- ➔ Using a pretrained sequence-to-sequence model, generate an *answer* that maximizes

$$P(\text{answer} \mid [\text{question, passage}])$$



Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Raffel et al. 2020.

Multilinguality

➔ Most Open-Retrieval QA datasets are in English

Data Type	Language
Questions	English
Answers	English
Text Corpus	English

Benchmarks

These leaderboards are used to track progress in Question Answering

Trend	Dataset	Best Model
	SQuAD1.1	🏆 {ANNA} (single model)
	SQuAD1.1 dev	🏆 T5-11B
	SQuAD2.0	🏆 IE-Net (ensemble)
	HotpotQA	🏆 BigBird-etc
	WikiQA	🏆 TANDA-RoBERTa (ASNQ, WikiQA)
	Quora Question Pairs	🏆 XLNet (single model)
	CNN / Daily Mail	🏆 GA+MAGE (32)
	TriviaQA	🏆 SpanBERT
	SQuAD2.0 dev	🏆 XLNet (single model)
	bAbi	🏆 STM
	Natural Questions (short)	🏆 BERTwwm + SQuAD 2

Multilinguality

➔ Most Open-Retrieval QA datasets are in English

Data Type	Language
Questions	English
Answers	English
Text Corpus	English

➔ Can we support *questions* in another language

➔ Can we search against a *corpus* in another language?

Multilinguality

- ➔ Two related problem settings in Open-Retrieval QA:
    1) Multilingual QA
    2) Crosslingual QA

Multilinguality

➔ Two related problem settings:

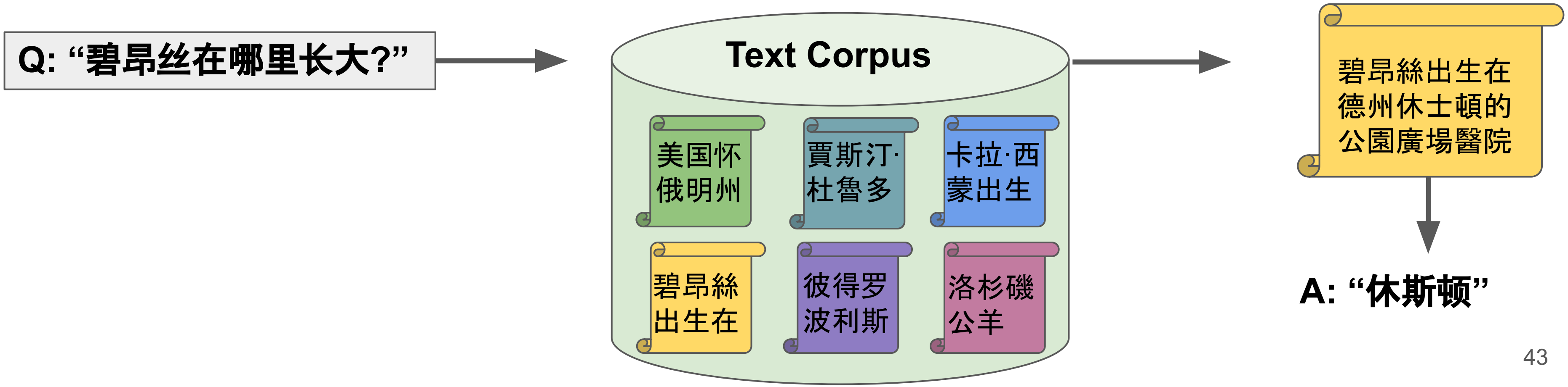
1) Multilingual QA

Data Type	Language
Questions	Language X
Answers	Language X
Text Corpus	Language X

Multilinguality

- ➔ Two related problem settings:
  1) Multilingual QA

Data Type	Language
Questions	Chinese
Answers	Chinese
Text Corpus	Chinese



Multilinguality

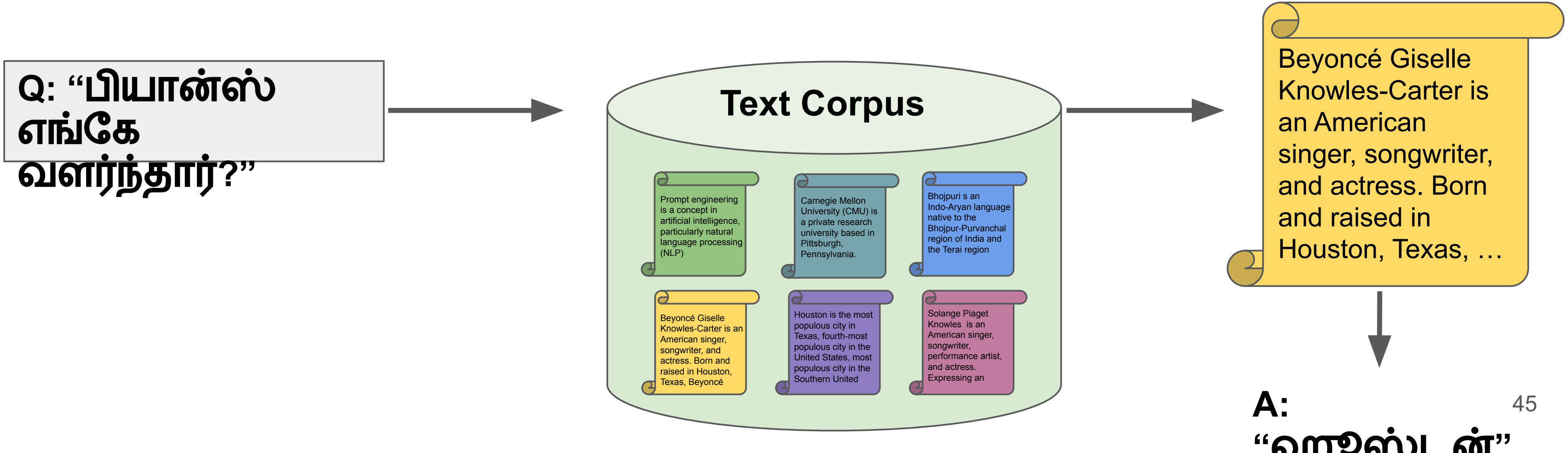
➔ Two related problem settings:
1) Multilingual QA
2) Crosslingual QA

Data Type	Language
Questions	Language X
Answers	Language Y
Text Corpus	Language Z

Multilinguality

- Two related problem settings:
  1) Multilingual QA
  2) Crosslingual QA

Data Type	Language
Questions	Tamil
Answers	Tamil
Text Corpus	English



Multilingual Open-Retrieval QA

- Approach 1: **Zero-Shot Transfer**
  - Choose a multilingual encoder (e.g. XLM-R)
  - Finetune it on an English-language QA dataset (e.g. SQuAD)
  - Transfer the encoder to a new language (e.g. Tamil)

Multilingual Open-Retrieval QA

- ➔ Approach 1: **Zero-Shot Transfer**
  - ◆ Choose a multilingual encoder (e.g. XLM-R)
  - ◆ Finetune it on an English-language QA dataset (e.g. SQuAD)
  - ◆ Transfer the encoder to a new language (e.g. Tamil)
- ➔ **Problem:** zero-shot transfer usually doesn't work great
  - ◆ Just giving a few target-language examples helps a lot

Task	Model	$k=0$	score	Δ	score	Δ	score	Δ	score	Δ	score	Δ
			$k=2$		$k=4$		$k=6$		$k=8$		$k=10$	
XQUAD	mBERT	45.62	48.12	2.50	48.66	3.04	49.34	3.72	49.91	4.29	50.19	4.57
	XLM-R	53.68	53.73	0.05	53.84	0.17	54.76	1.08	55.56	1.88	55.78	2.10

From Zero to Hero: On the Limitations of Zero-Shot Language Transfer with Multilingual Transformers. Lauscher et al. 2021.

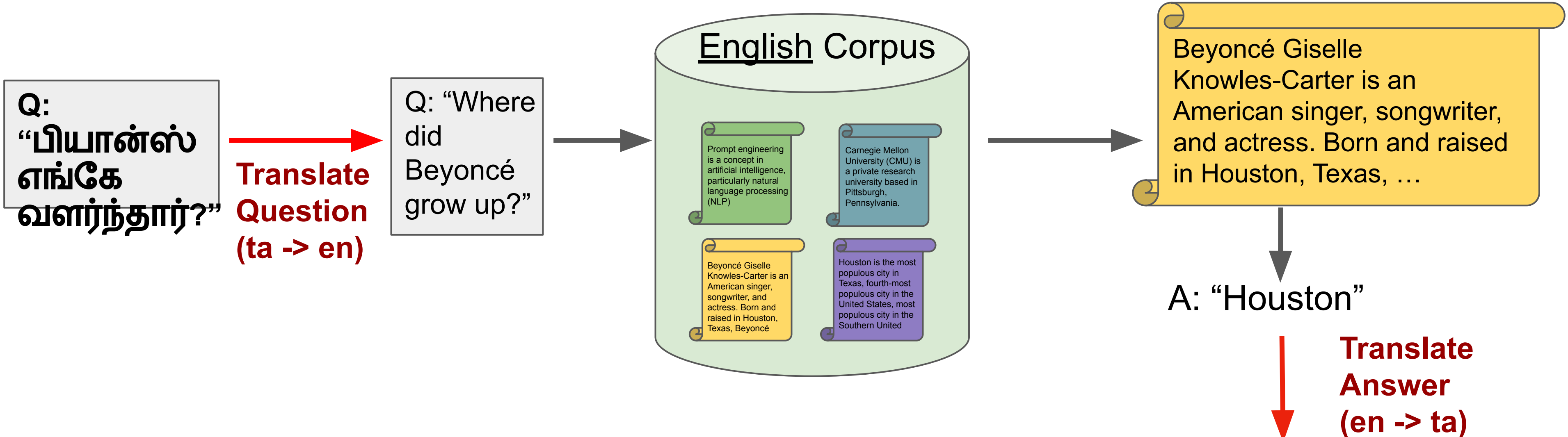
Multilingual Open-Retrieval QA

- Approach 2: **Translation-Based Adaptation**
  - "Translate-Test"
    - Translate question to English
    - Apply an off-the-shelf English QA system against an English text corpus
    - Translate the answer into the language of your choice
  - Discussed in the Ruder reading

Bootstrap Pattern Learning for Open-Domain CLQA. Shima and Mitamura 2010.
Multi-domain Multilingual Question Answering [blog post]. Sebastian Ruder 2021.

Multilingual Open-Retrieval QA

- ➔ Approach 2: **Translation-Based Adaptation**
  - ◆ “Translate-Test”



Bootstrap Pattern Learning for Open-Domain CLQA. Shima and Mitamura 2010.
Multi-domain Multilingual Question Answering [blog post]. Sebastian Ruder 2021.

Multilingual Open-Retrieval QA

Data Type	Language
Questions	Tamil
Answers	Tamil
Text Corpus	English

- Approach 2: **Translation-Based Adaptation**
  - "Translate-Test"
    - Suffers from error propagation from MT systems + QA system
    - Answers must be found in an English corpus
      - Leads to anglocentric QA systems

Bootstrap Pattern Learning for Open-Domain CLQA. Shima and Mitamura 2010.
Multi-domain Multilingual Question Answering [blog post]. Sebastian Ruder 2021.

Multilingual Open-Retrieval QA

Data Type	Language
Questions	Tamil
Answers	Tamil
Text Corpus	English

➔ Approach 2: **Translation-Based Adaptation**
- "Translate-Train"
  - Translate full training data (questions, answers, and text corpus/passages) to target language
  - Train model in target language

Bootstrap Pattern Learning for Open-Domain CLQA. Shima and Mitamura 2010.
Multi-domain Multilingual Question Answering [blog post]. Sebastian Ruder 2021.

Multilingual Open-Retrieval QA

Data Type	Language
Questions	Tamil
Answers	Tamil
Text Corpus	English

- ➔ Approach 2: **Translation-Based Adaptation**
  - ◆ "Translate-Train"
    - Translate full training data (questions, answers, and text corpus/passages) to target language
    - Train model in target language
    - At test time, run open-retrieval QA system on translated text corpus

Bootstrap Pattern Learning for Open-Domain CLQA. Shima and Mitamura 2010.
Multi-domain Multilingual Question Answering [blog post]. Sebastian Ruder 2021.

Multilingual Open-Retrieval QA

Data Type	Language
Questions	Tamil
Answers	Tamil
Text Corpus	English

- ➔ Approach 2: **Translation-Based Adaptation**
  - ◆ "Translate-Train"
    - Requires translating full text corpus (e.g. English Wikipedia)
    - Text corpus (and training data) are noisy due to MT errors

Bootstrap Pattern Learning for Open-Domain CLQA. Shima and Mitamura 2010.
Multi-domain Multilingual Question Answering [blog post]. Sebastian Ruder 2021.

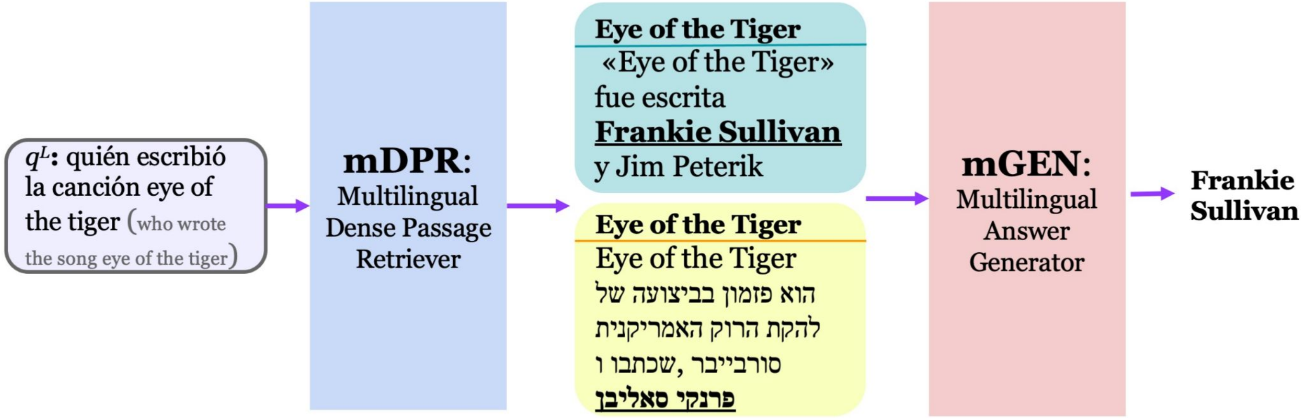
Multilingual Open-Retrieval QA

- Approach 3: **multilingual retriever-generator** (Asai et al. 2021)
  - Method for cross-lingual QA *without doing any translation*
  - Search for answers from corpora from multiple languages

One Question Answering Model for Many Languages with Cross-lingual Dense Passage Retrieval. Asai et al. 2021.

Multilingual Open-Retrieval QA

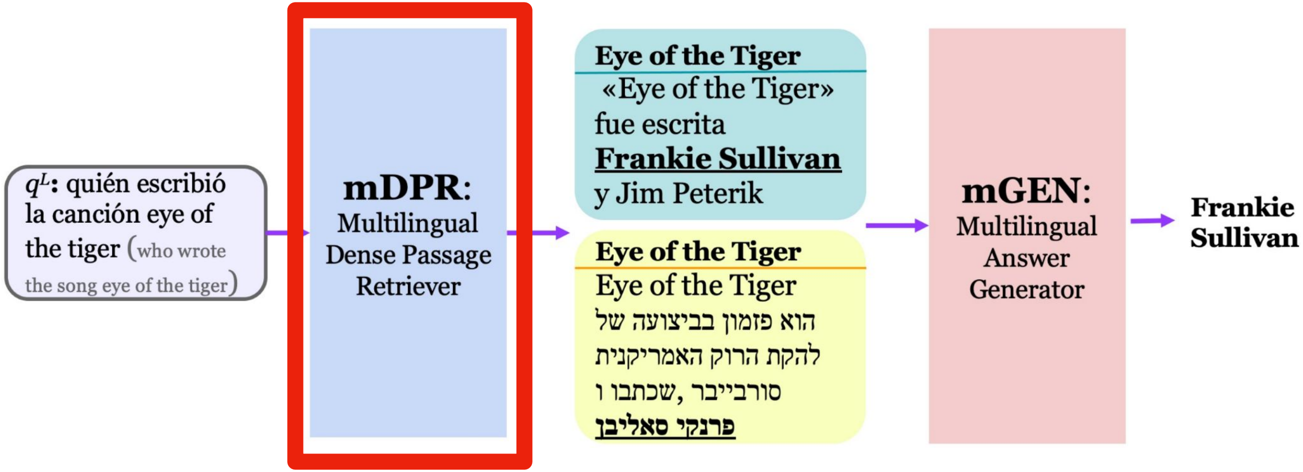
➔ Approach 3: **multilingual retriever-generator** (Asai et al. 2021)



One Question Answering Model for Many Languages with Cross-lingual Dense Passage Retrieval. Asai et al. 2021.

Multilingual Open-Retrieval QA

➔ Approach 3: **multilingual retriever-generator** (Asai et al. 2021)



Leverage “multilingual DPR” to retrieve passages of any language from queries of any language

One Question Answering Model for Many Languages with Cross-lingual Dense Passage Retrieval. Asai et al. 2021.

Multilingual Open-Retrieval QA

➔ Approach 3: **multilingual retriever-generator** (Asai et al. 2021)

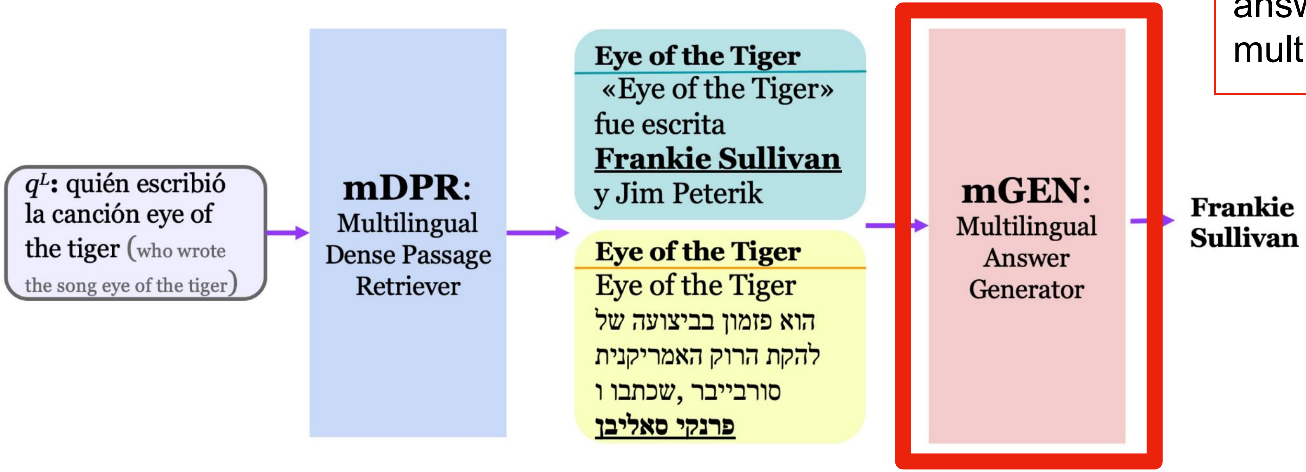
Retrieve relevant passages from multiple languages' text corpora

One Question Answering Model for Many Languages with Cross-lingual Dense Passage Retrieval. Asai et al. 2021.

Multilingual Open-Retrieval QA

Use multilingual answer generator (“mT5”) to generate target-language answer from several multilingual passages.

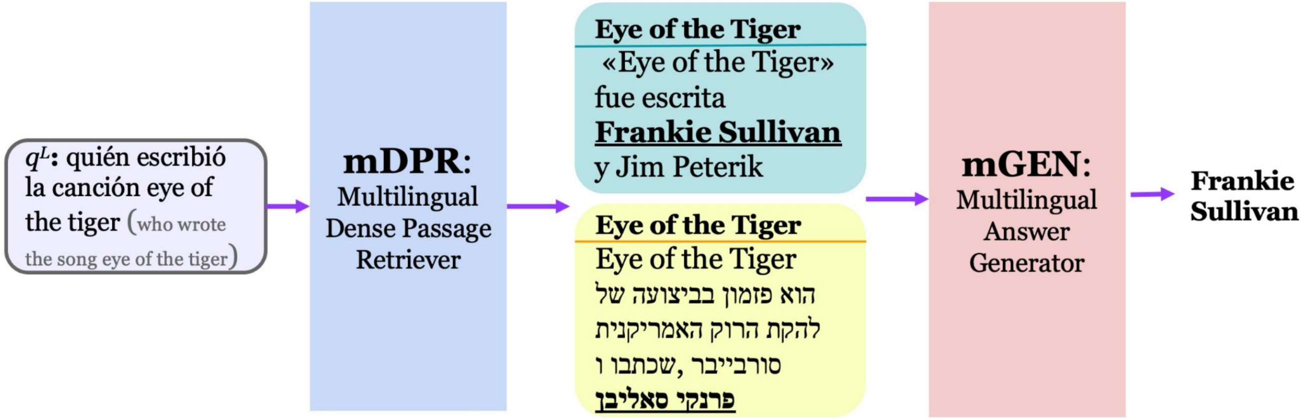
➔ Approach 3: **multilingual retriever-generator** (Asai et al. 2021)



One Question Answering Model for Many Languages with Cross-lingual Dense Passage Retrieval. Asai et al. 2021.
mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. Xue et al. 2021.

Multilingual Open-Retrieval QA

➔ Approach 3: **multilingual retriever-generator** (Asai et al. 2021)

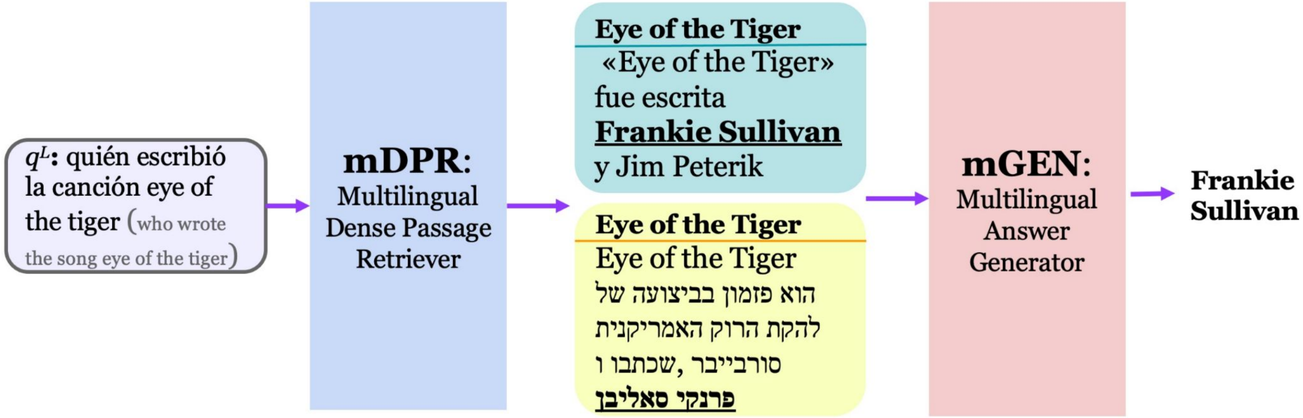


◆ Challenge: lack of training data for mDPR and mGEN modules

One Question Answering Model for Many Languages with Cross-lingual Dense Passage Retrieval. Asai et al. 2021.

Multilingual Open-Retrieval QA

➔ Approach 3: **multilingual retriever-generator** (Asai et al. 2021)



- ◆ Challenge: lack of training data for mDPR and mGEN modules
    - Solution: iterative self-training

One Question Answering Model for Many Languages with Cross-lingual Dense Passage Retrieval. Asai et al. 2021.

Multilingual Open-Retrieval QA

➔ Approach 3: **multilingual retriever-generator** (Asai et al. 2021)

Models	Target Language L_i F1							Macro Average		
	Ar	Bn	Fi	Ja	Ko	Ru	Te	**F1**	**EM**	**BLEU**
CORA	**59.8**	**40.4**	**42.2**	**44.5**	**27.1**	**45.9**	**44.7**	**43.5**	**33.5**	**31.1**
SER	32.0	23.1	23.6	14.4	13.6	11.8	22.0	20.1	13.5	20.1
GMT+GS	31.5	19.0	18.3	8.8	20.1	19.8	13.6	18.7	12.1	16.8
MT+Mono	25.1	12.7	20.4	12.9	10.5	15.7	0.8	14.0	10.5	11.4
MT+DPR	7.6	5.9	16.2	9.0	5.3	5.5	0.8	7.2	3.3	6.3
BM25	31.1	21.9	21.4	12.4	12.1	17.7	–	–	–	–

Evaluation on “XOR-QA” dataset

One Question Answering Model for Many Languages with Cross-lingual Dense Passage Retrieval. Asai et al. 2021.
XOR QA: Cross-lingual Open-Retrieval Question Answering. Asai et al. 2020

Multilingual Open-Retrieval QA

➔ Aside: some QA metrics

F1 **EM** **BLEU**

- F1: treat generated and ground-truth answers as bags of tokens
  - Compute precision and recall of token matches

Speech and Language Processing, 3rd edition (draft). Jurafsky and Martin 2021.

Multilingual Open-Retrieval QA

➔ Aside: some QA metrics

Generated: "Cavalier King Charles Spaniel"
Actual: "King Charles"

F1 **EM** **BLEU**

Precision = $\frac{2}{4}$ Recall = 1 **F1** = $2\frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$ = $\frac{2}{3}$

- F1: treat generated and ground-truth answers as bags of tokens
  - Compute precision and recall of token matches

Speech and Language Processing, 3rd edition (draft). Jurafsky and Martin 2021.

Multilingual Open-Retrieval QA

➔ Aside: some QA metrics

Generated: “Cavalier King Charles Spaniel”
Actual: “King Charles”

F1 **EM** **BLEU**

Exact Match? False

- F1: treat generated and ground-truth answers as bags of tokens
  - Compute precision and recall of token matches
- EM: how often does generated answer *exactly match* ground-truth answer?

Speech and Language Processing, 3rd edition (draft). Jurafsky and Martin 2021.

Multilingual Open-Retrieval QA

➔ Aside: some QA metrics

Generated: “Cavalier King Charles Spaniel”
Actual: “King Charles”

F1 **EM** **BLEU**

- F1: treat generated and ground-truth answers as bags of tokens
  - Compute precision and recall of token matches
- EM: how often does generated answer *exactly match* ground-truth answer?
- BLEU: n-gram overlap

Speech and Language Processing, 3rd edition (draft). Jurafsky and Martin 2021.

Multilingual Open-Retrieval QA

➔ Approach 3: **multilingual retriever-generator** (Asai et al. 2021)

Models	Target Language L_i F1							Macro Average		
	Ar	Bn	Fi	Ja	Ko	Ru	Te	**F1**	**EM**	**BLEU**
CORA	**59.8**	**40.4**	**42.2**	**44.5**	**27.1**	**45.9**	**44.7**	**43.5**	**33.5**	**31.1**
SER	32.0	23.1	23.6	14.4	13.6	11.8	22.0	20.1	13.5	20.1
GMT+GS	31.5	19.0	18.3	8.8	20.1	19.8	13.6	18.7	12.1	16.8
MT+Mono	25.1	12.7	20.4	12.9	10.5	15.7	0.8	14.0	10.5	11.4
MT+DPR	7.6	5.9	16.2	9.0	5.3	5.5	0.8	7.2	3.3	6.3
BM25	31.1	21.9	21.4	12.4	12.1	17.7	–	–	–	–

Evaluation on “XOR-QA” dataset

One Question Answering Model for Many Languages with Cross-lingual Dense Passage Retrieval. Asai et al. 2021.
XOR QA: Cross-lingual Open-Retrieval Question Answering. Asai et al. 2020

Multilingual Open-Retrieval QA

➔ Approach 3: **multilingual retriever-generator** (Asai et al. 2021)

Models	Target Language L_i F1							Macro Average		
	Ar	**Bn**	**Fi**	**Ja**	**Ko**	**Ru**	**Te**	**F1**	**EM**	**BLEU**
CORA	**59.8**	**40.4**	**42.2**	**44.5**	**27.1**	**45.9**	**44.7**	**43.5**	**33.5**	**31.1**
SER	32.0	23.1	23.6	14.4	13.6	11.8	22.0	20.1	13.5	20.1
GMT+GS	31.5	19.0	18.3	8.8	20.1	19.8	13.6	18.7	12.1	16.8
MT+Mono	25.1	12.7	20.4	12.9	10.5	15.7	0.8	14.0	10.5	11.4
MT+DPR	7.6	5.9	16.2	9.0	5.3	5.5	0.8	7.2	3.3	6.3
BM25	31.1	21.9	21.4	12.4	12.1	17.7	–	–	–	–

their model

Evaluation on “XOR-QA” dataset

One Question Answering Model for Many Languages with Cross-lingual Dense Passage Retrieval. Asai et al. 2021.
XOR QA: Cross-lingual Open-Retrieval Question Answering. Asai et al. 2020

Multilingual Open-Retrieval QA

➔ Approach 3: **multilingual retriever-generator** (Asai et al. 2021)

	Models	Target Language L_i F1							Macro Average		
		Ar	**Bn**	**Fi**	**Ja**	**Ko**	**Ru**	**Te**	**F1**	**EM**	**BLEU**
their model	CORA	**59.8**	**40.4**	**42.2**	**44.5**	**27.1**	**45.9**	**44.7**	**43.5**	**33.5**	**31.1**
	SER	32.0	23.1	23.6	14.4	13.6	11.8	22.0	20.1	13.5	20.1
	GMT+GS	31.5	19.0	18.3	8.8	20.1	19.8	13.6	18.7	12.1	16.8
	MT+Mono	25.1	12.7	20.4	12.9	10.5	15.7	0.8	14.0	10.5	11.4
"Translate-Test"	MT+DPR	7.6	5.9	16.2	9.0	5.3	5.5	0.8	7.2	3.3	6.3
	BM25	31.1	21.9	21.4	12.4	12.1	17.7	–	–	–	–

Evaluation on "XOR-QA" dataset

One Question Answering Model for Many Languages with Cross-lingual Dense Passage Retrieval. Asai et al. 2021.
XOR QA: Cross-lingual Open-Retrieval Question Answering. Asai et al. 2020

Multilingual Open-Retrieval QA

➔ Approach 3: **multilingual retriever-generator** (Asai et al. 2021)

Models	Target Language L_i F1							Macro Average		
	Ar	**Bn**	**Fi**	**Ja**	**Ko**	**Ru**	**Te**	**F1**	**EM**	**BLEU**
CORA	**59.8**	**40.4**	**42.2**	**44.5**	**27.1**	**45.9**	**44.7**	**43.5**	**33.5**	**31.1**

their model

Note: the F1 and EM scores for the English *Natural Questions* dataset are 79.6 and 71.9

Evaluation on "XOR-QA" dataset

One Question Answering Model for Many Languages with Cross-lingual Dense Passage Retrieval. Asai et al. 2021.
XOR QA: Cross-lingual Open-Retrieval Question Answering. Asai et al. 2020

Multilingual QA Datasets

➔ Machine Reading Comprehension

On the Cross-lingual Transferability of Monolingual Representations. Artetxe et al. 2019.
MLQA: Evaluating Cross-lingual Extractive Question Answering. Lewis et al. 2019.

Multilingual QA Datasets

- ➔ Multilingual Machine Reading Comprehension
  - ◆ XQuAD (Artetxe et al. 2020)
    - Based on 1.1K SQuAD question-answer-passage triples
    - Each professionally translated into 10 languages
  - ◆ MLQA (Lewis et al. 2020)
    - ~5K samples in each of 6 languages + English

On the Cross-lingual Transferability of Monolingual Representations. Artetxe et al. 2019.
MLQA: Evaluating Cross-lingual Extractive Question Answering. Lewis et al. 2019.

Multilingual QA Datasets

- Multilingual Open-Retrieval QA
  - MKQA (Longpre et al. 2020)
    - 10K QA pairs from *Natural Questions* (Kwiatkowski et al. 2020) are translated into 26 languages
    - Assumes answer can be found from English Wikipedia

Natural Questions: a Benchmark for Question Answering Research. Kwiatkowski et al. 2020.
MKQA: A Linguistically Diverse Benchmark for Multilingual Open Domain Question Answering. Longpre et al. 2020.
TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages. Clark et al. 2020..

Multilingual QA Datasets

- ➔ Multilingual Open-Retrieval QA
  - ◆ MKQA (Longpre et al. 2020)
    - 10K QA pairs from *Natural Questions* (Kwiatkowski et al. 2020) are translated into 26 languages
    - Assumes answer can be found from English Wikipedia
  - ◆ TyDi QA (Clark et al. 2020)
    - 200K QA pairs are collected *naturally* in 11 languages
    - Text corpus is each language's native Wikipedia

Natural Questions: a Benchmark for Question Answering Research. Kwiatkowski et al. 2020.
MKQA: A Linguistically Diverse Benchmark for Multilingual Open Domain Question Answering. Longpre et al. 2020.
TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages. Clark et al. 2020..

Translationese

- ➔ Every dataset until TyDi QA translated English QA pairs into a target language
- ➔ Problems:

TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages. Clark et al. 2020..

Translationese

- ➔ Every dataset until TyDi QA translated English QA pairs into a target language
- ➔ Problems:
  1) Translated questions may not be natural
     - Translations into free word-order languages are more likely to adopt English word order

TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages. Clark et al. 2020..

Translationese

- ➔ Every dataset until TyDi QA translated English QA pairs into a target language
- ➔ Problems:
  1) Translated questions may not be natural
     - Translations into free word-order languages are more likely to adopt English word order
  2) Answers are assumed to be Anglocentric
     - e.g. The question “Which Indian lawyer advocated the preservation of Marina Beach in Chennai?” cannot be answered from English Wikipedia
       - Answer: V. Krishnaswamy Iyer (கிருஷ்ணசுவாமி ஐயர்)

TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages. Clark et al. 2020..

Multilingual QA Datasets

- Crosslingual Open-Retrieval QA
  - XOR-QA (Asai et al. 2020)
    - Simply TyDi QA with an added feature
    - If the question has no answer in the original language, a translated answer in English wikipedia is provided

XOR QA: Cross-lingual Open-Retrieval Question Answering. Asai et al. 2020..

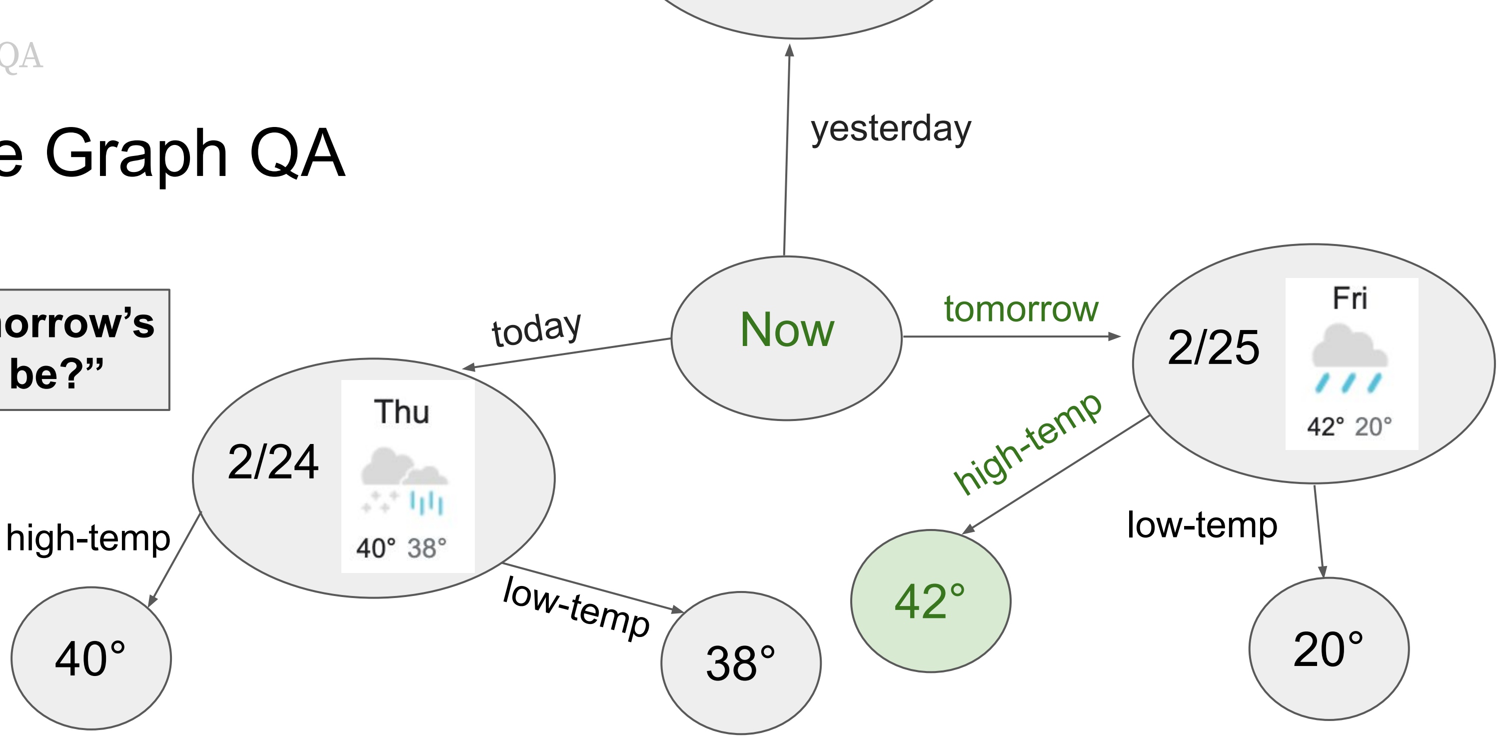
Multilingual QA Datasets

➔ XQuAD, MLQA, and TyDi-QA are all included in XTREME multilingual modeling benchmark (Hu et al 2020)

XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization. Hu, Ruder, et al. 2020.

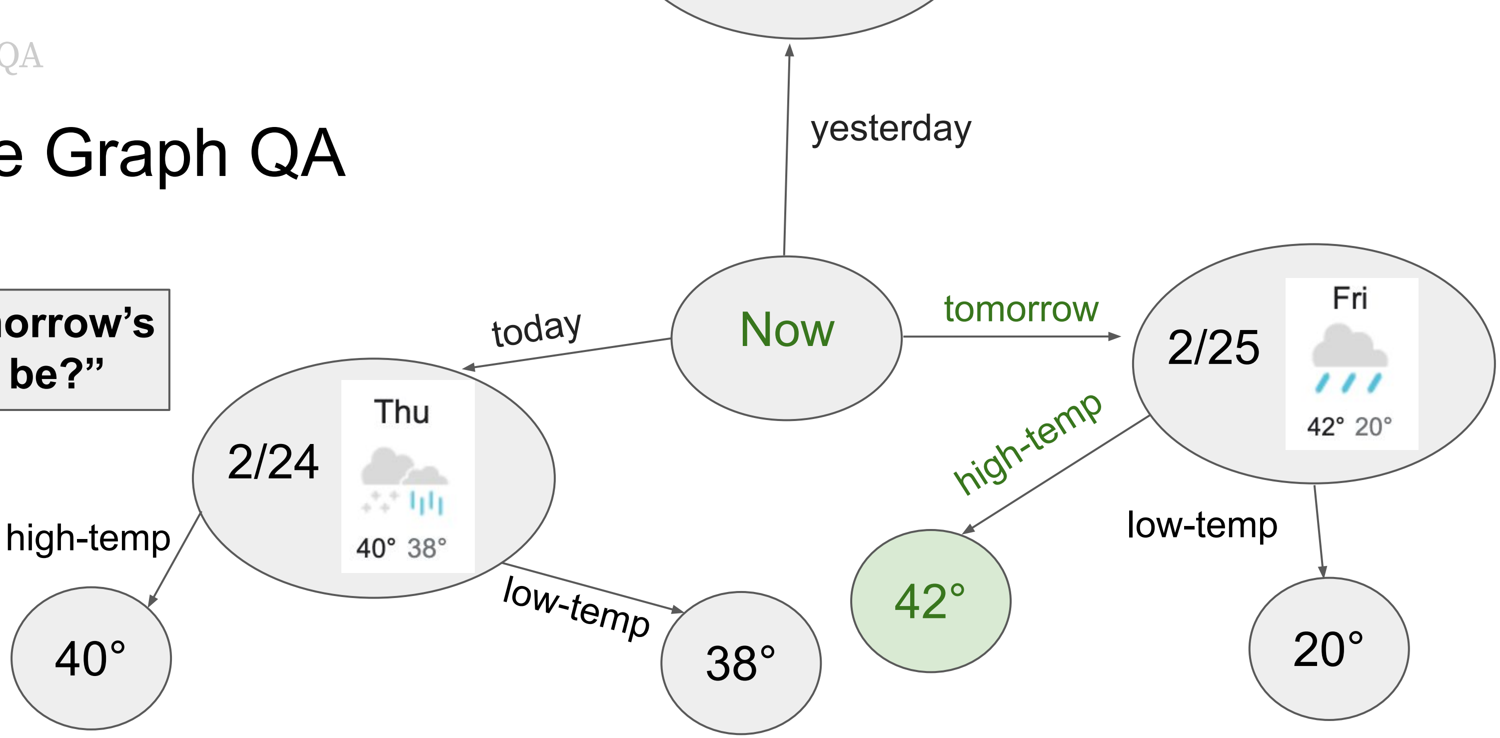
Knowledge Graph QA

Q: "What will tomorrow's high temperature be?"



Knowledge Graph QA

Q: “What will tomorrow’s high temperature be?”



➔ Goal: convert question into a query on a structured knowledge graph

- *Semantic parsing*

Semantic parsing on freebase from question-answer pairs. Berant et al. 2013.

Knowledge Graph QA Methods

➔ Little work has been done in this space

➔ Known methods:

1) Zero-Shot Transfer
2) Translation-Based Adaptation (e.g. “Translate-Train”)
   - Used in the paper given in the reading

Improving Zero-Shot Cross-lingual Transfer for Multilingual Question Answering over Knowledge Graph. Zhou et al. 2021.

Knowledge Graph QA Methods

➔ Little work has been done in this space

➔ Known methods:

1) Zero-Shot Transfer
2) Translation-Based Adaptation (e.g. “Translate-Train”)
    - Used in the paper given in the reading (Zhou et al 2021)
    - In the paper, they leverage word-level translation (“unsupervised bilingual lexicon induction” instead of true MT
        - Motivation: KGQA mainly is concerned with phrase-level semantics

Improving Zero-Shot Cross-lingual Transfer for Multilingual Question Answering over Knowledge Graph. Zhou et al. 2021.

Knowledge Graph QA Datasets

- Dataset for Multilingual KGQA
  - RuBQ 2.0 (Rybin et al. 2021)
    - 2910 questions in Russian to be answered using Wikidata (multilingual)
  - CWQ (Cui et al. 2021)
    - 10K questions in English, Hebrew, Kannada and Chinese
    - Each accompanied with a SPARQL query, which gives the correct answer when run on Wikidata

RuBQ 2.0: An innovated Russian question answering dataset. Rybin et al. 2021.
Multilingual Compositional Wikidata Questions. Cui et al. 2021.

Knowledge Graph QA Datasets

- Dataset for Multilingual KGQA
  - QALD-9-plus (Perevalov et al 2022)
    - Released in Feb. 2022
    - 558 QA pairs to be answered using DBPedia (multilingual KG)
    - Questions in 8 languages:
      - German, Russian, French, Armenian, Belarusian, Lithuanian, Bashkir, and Ukrainian

QALD-9-plus: A Multilingual Dataset for Question Answering over DBpedia and Wikidata Translated by Native Speakers. Perevalov et al. 2021.

Open Problems in Multilingual QA

- ➔ Multilingual KGQA
  - ◆ New-ish research area
- ➔ Crosslingual QA
  - ◆ Plenty of room for improvement
- ➔ Multimodal Multilingual QA
  - ◆ How to leverage non-linguistic cues?
- ➔ Code-Switching in QA
  - ◆ Answering code-mixed questions

Multi-domain Multilingual Question Answering [blog post]. Sebastian Ruder 2021.

Discussion Question

- Read either:
  - One Question Answering Model for Many Languages with Cross-lingual Dense Passage Retrieval (Asai et al 2021)
  - Improving Zero-Shot Cross-lingual Transfer for Multilingual Question Answering over Knowledge Graph (Zhou et al 2021)
  - Multi-domain Multilingual Question Answering (Ruder 2021 [blog post])

Discussion Question

- Think about a practical question-answering application for a language or domain of interest to you, with an eye towards low-resource or crosslingual QA.
  - What population would use this application in that language/domain, if any?
  - Do you think the methods in this paper would work for your language/domain?
  - What other resources or strategies might you consider to solve and evaluate this task? *(optional)*