Real-Time Open-Domain QA with Dense-Sparse Phrase Index

Minjoon Seo*, Jinhyuk Lee*, Tom Kwiatkowski, Ankur Parikh, Ali Farhadi, Hannaneh Hajishirzi

UNIVERSITY of WASHINGTON NAVER Google AI AI2 XNOR.AI KOREA UNIVERSITY

* denotes equal contribution

Open-domain QA?

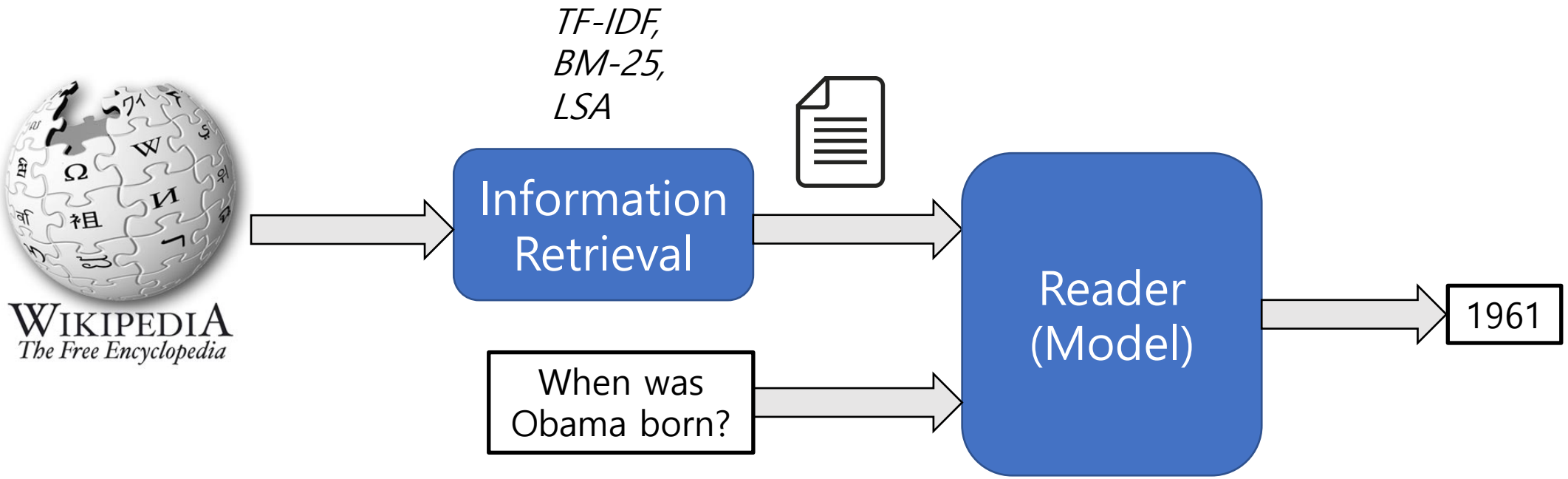
5 Million documents
3 Billion tokens

When was Obama born?

Some Model

1961

Retrieve & Read



1. Error propagation: reading only **5-10** docs
2. Query-dependent encoding: **30s+ per query**

Chen et al., 2017

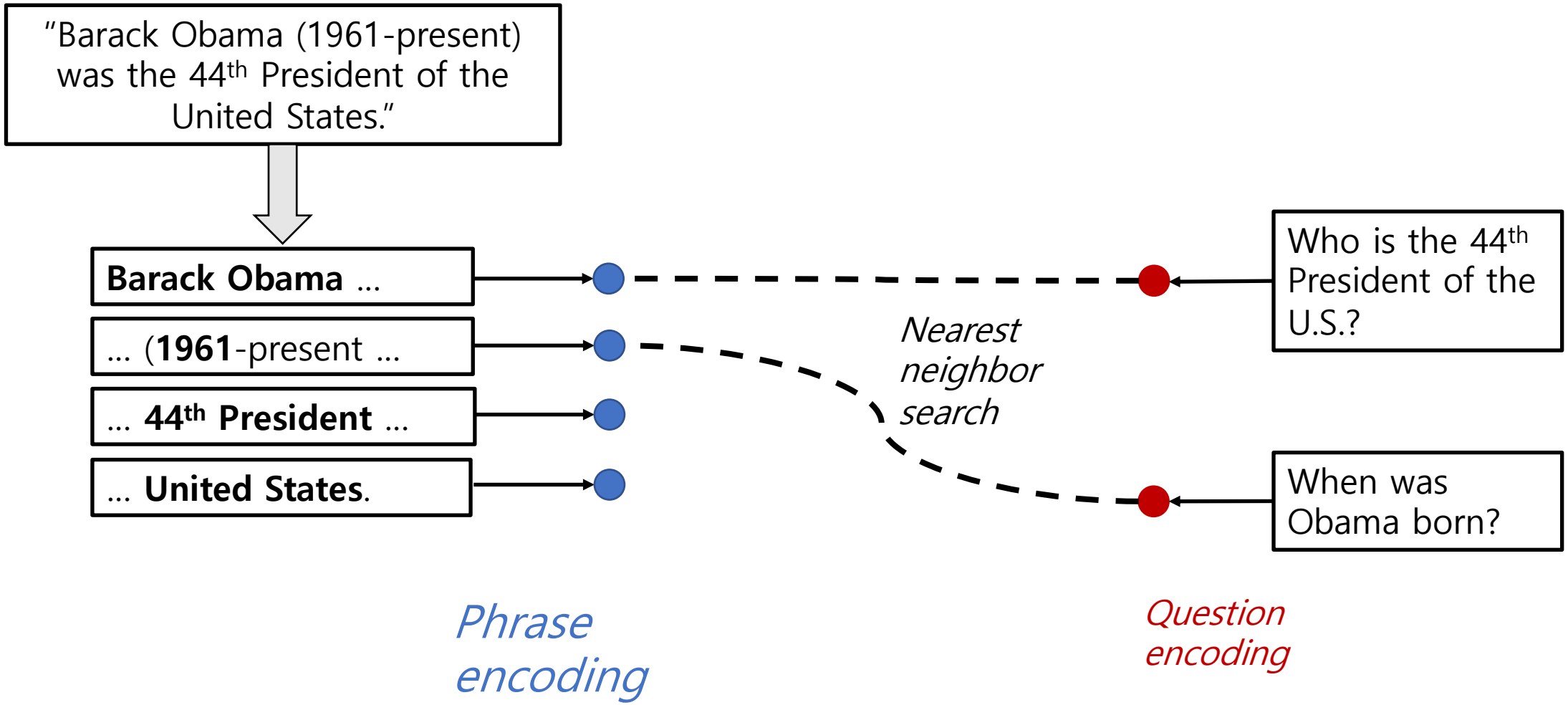
We want…

- To ”read” entire Wikipedia
  - 5-10 docs → 5 Million docs
  - Reach long-tail answers
- Fast inference on CPUs
  - 35s → 0.5s
  - Maintain high accuracy

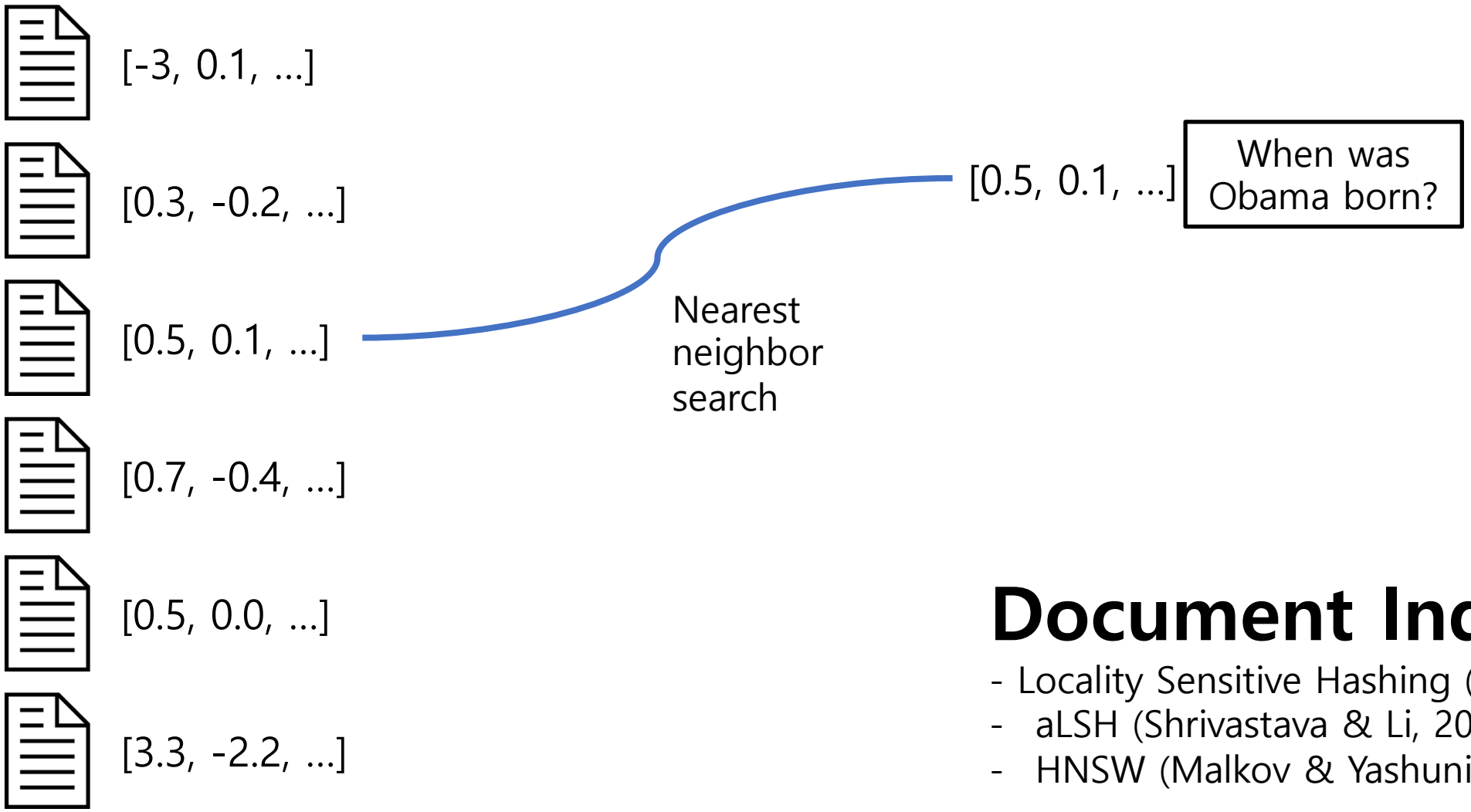
HOW?

Our approach: index phrases!

Phrase Indexing



Seo et al., 2018



Document Indexing

- Locality Sensitive Hashing (LSH)
- aLSH (Shrivastava & Li, 2014)
- HNSW (Malkov & Yashunin, 2018)

Model phrase question document

$$\hat{a} = \underset{a}{\operatorname{argmax}}\, F_\theta(a, q, d)$$

Query-Agnostic Decomposition

$$\hat{a} = \underset{a}{\operatorname{argmax}}\, G_\theta(q) \cdot H_\theta(a, d)$$

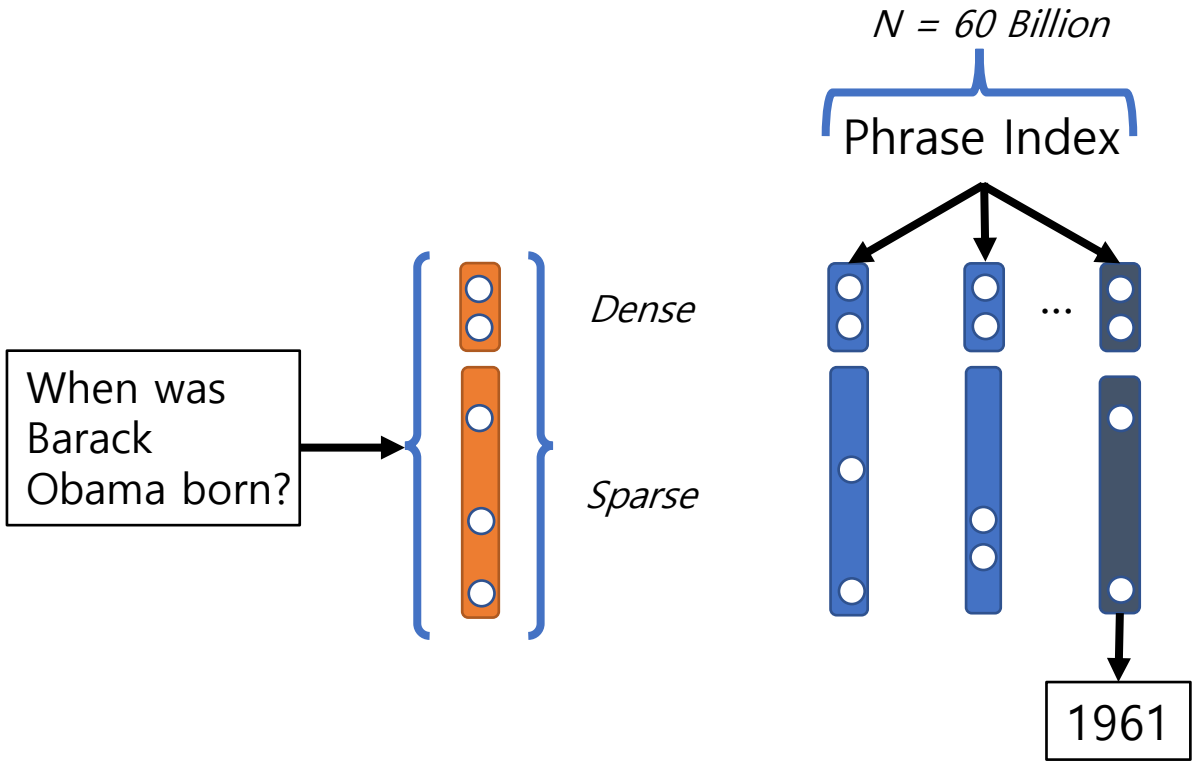
Question encoder

Phrase encoder

Phrase (and question) Representation

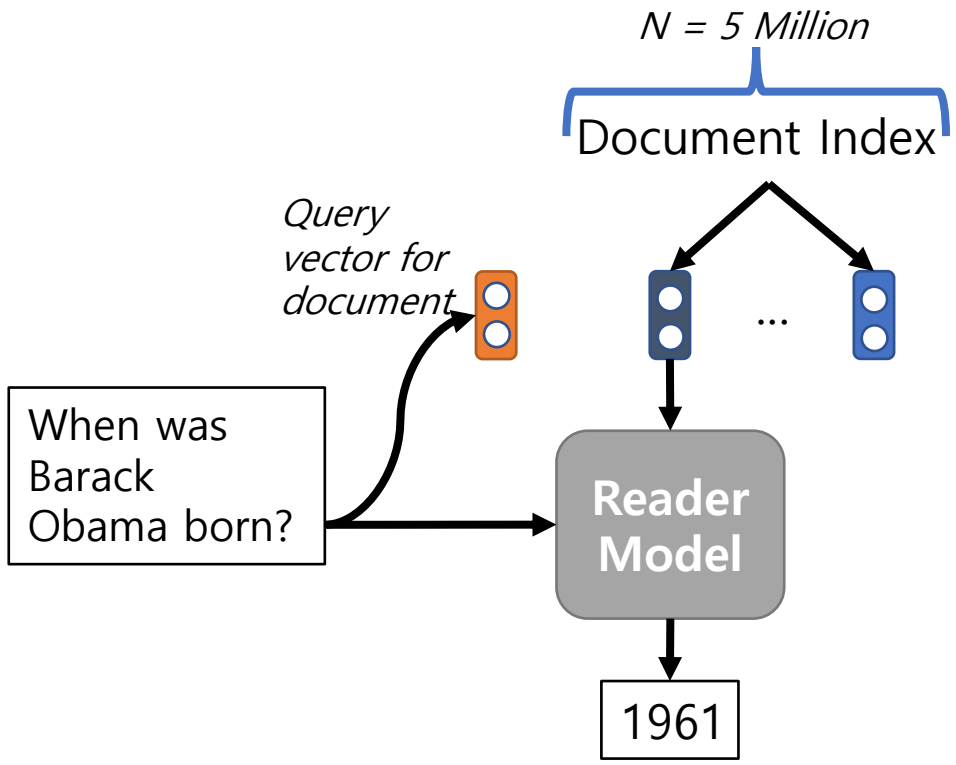
- Dense representation
  - Can utilize deep neural networks
  - great for capturing semantic and syntactic information
  - Not great for disambiguating "Einstein" vs "Tesla"
- Sparse representation (bag-of-word)
  - Great for capturing lexical information
- Represent each phrase with a *concatenation of both*

Dense-Sparse Phrase Index (DenSPI)



DenSPI

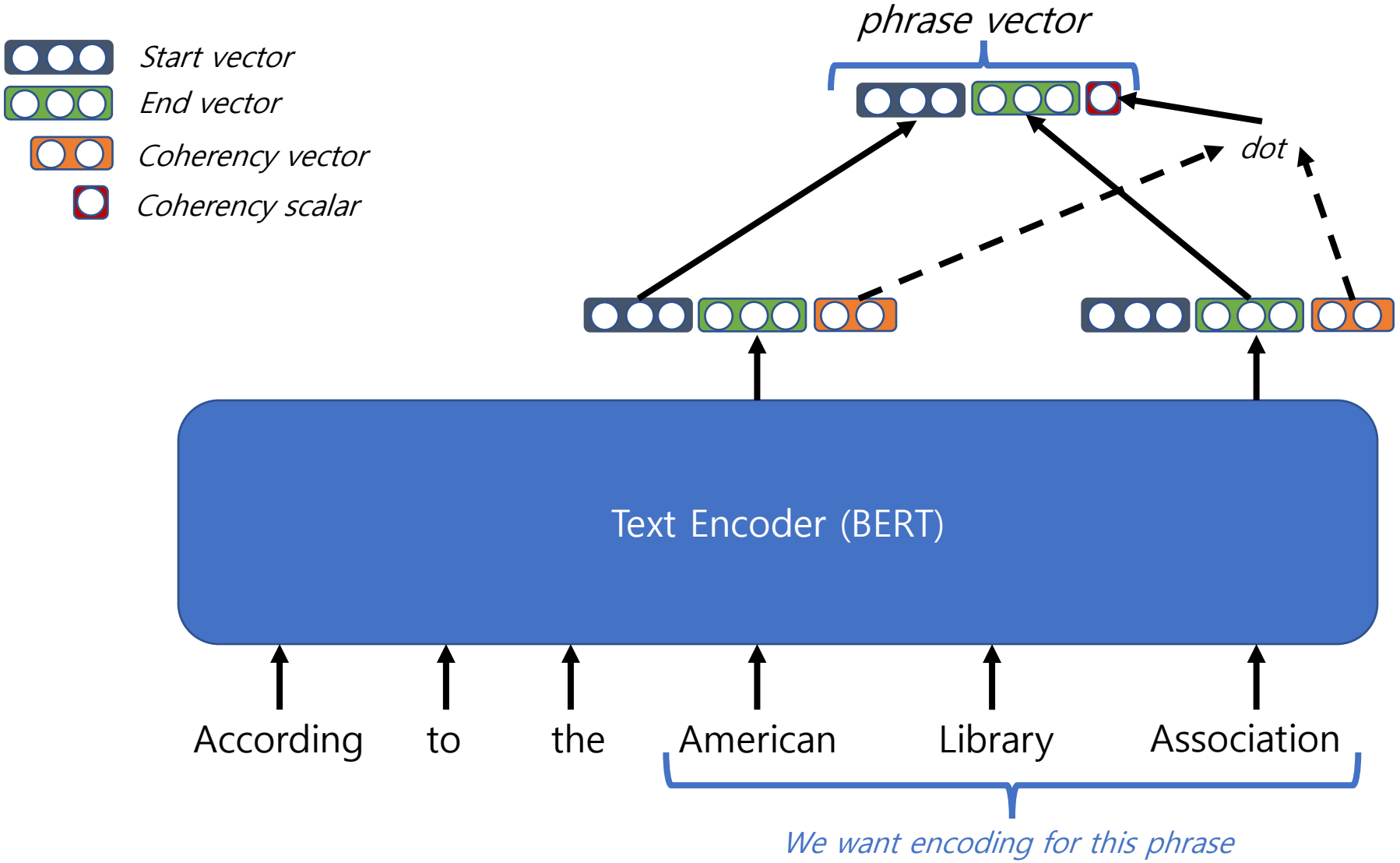
Ours



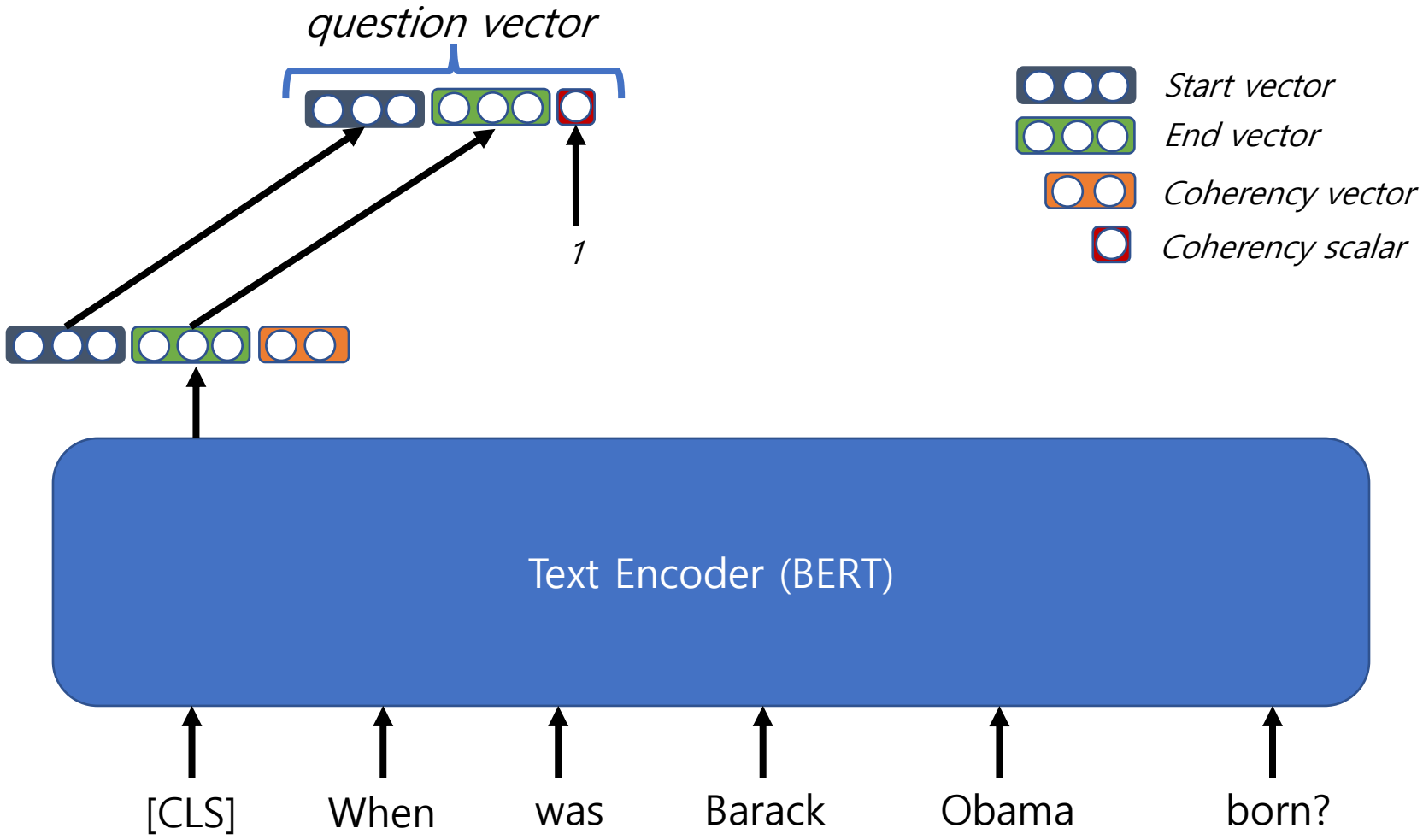
Retrieve & Read

(Chen et al., 2017)

Dense Representation for Phrases



Dense Representation for Questions



Sparse Representation

- TF-IDF document & paragraph vector, computed over Wikipedia
- Unigram & Bigram (vocab size = 17 Million)
- Adopted DrQA's vocab/TF-IDF (Chen et al., 2017)

Beware of the scale…

- **60 Billion** phrases in Wikipedia!
- Training
  - Softmax on 60 Billion phrases?
- Storage
  - 60 Billion phrases x 4 KB per phrase = 240 TB?
- Search
  - Exact search on 60 Billion phrases?

We want to be open-research-friendly

Training

- Close-domain QA dataset: the model can easily overfit
  - e.g. ”who” question when only one named entity in the context
- Negative sampling and concatenation
  - Sampling strategy is crucial
  - Use query encoder to associate similar questions in training set
  - Concatenate the context that the similar question belongs to

Storage

- 60 Billion phrases x 4 KB per phrase = **240 TB**!

1. **Pointer**: share start and end vectors
   - 240 TB → 12 TB
2. **Filter**: 1-layer classifier on phrase vectors
   - 12 TB → 4.5 TB
3. **Scalar Quantization**: 4 bytes → 1 byte per dim
   - 4.5 TB → **1.5 TB**

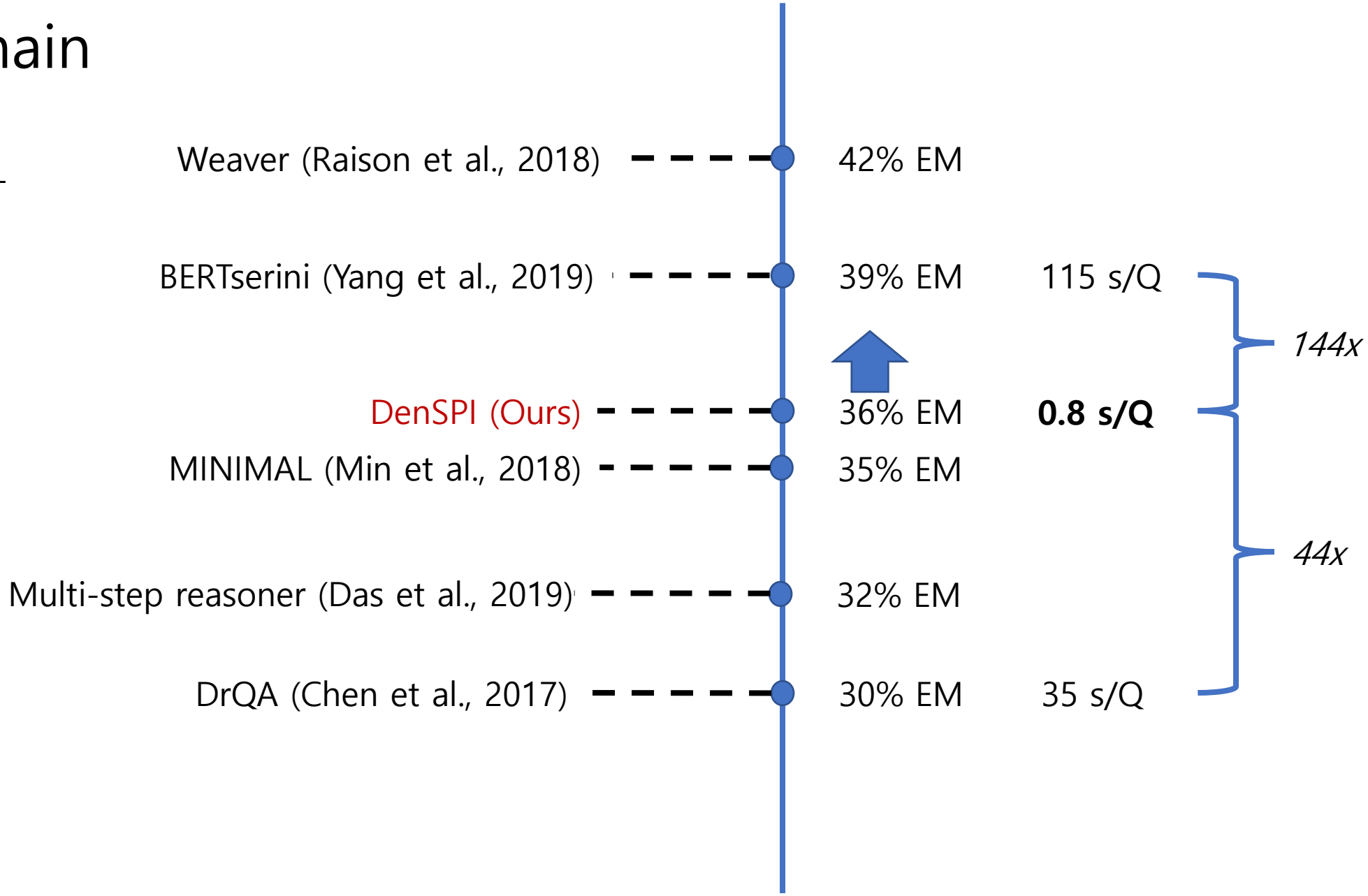
Search

- An open-source library for large-scale dense+sparse nearest neighbor search is non-existent
- Dense-first search (DFS)
- Sparse-first search (SFS)
- Hybrid

Experiments

Open-Domain SQuAD

Red color is query-agnostic.



Qualitative Comparisons

Q: What can hurt a teacher's mental and physical health?

Mental health

… and poor mental health can lead to problems such as **substance abuse**.

Teacher

Teachers face several occupational hazards in their line of work, including **occupational stress**…

Retrieve & Read (Chen et al., 2017)

DenSPI (Ours)

Q: Who was Kennedy's science adviser that opposed manned spacecraft flights?

Apollo program

Kennedy's science advisor **Jerome Wiesner**, … his opposition to manned spaceflight …

Apollo program

… and the sun by NASA manager **Abe Silverstein**, who later said that …

Apollo program

Although Grumman wanted a second unmanned test, **George Low** decided … be manned.

Apollo program

Kennedy's science advisor **Jerome Wiesner**, … his opposition to manned spaceflight …

Space Race

Jerome Wiesner of MIT, who served as a … advisor to … Kennedy, … opponent of manned

John F. Kennedy

… science advisor **Jerome Wiesner** … strongly opposed to manned space exploration, …

Q: What is the best thing to do when bored?

Bored to Death (song)

I'm nearly bored to death

Big Brother 2

When bored, she enjoys **drawing**.

Waterview Connection

The twin tunnels were bored by … tunnel **boring** machine (TBM) …

Angry Kid

he can think of a much more fun thing he can do while on his back: **painting**.

Bored to Death (song)

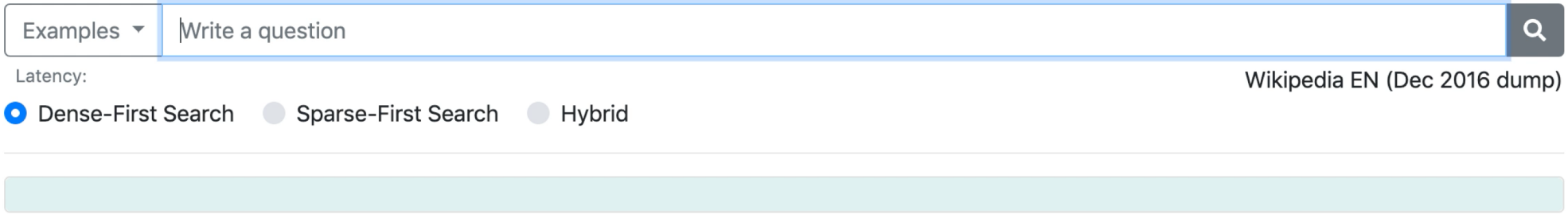
It's easier to say you're bored, or to be angry, than it is to be **sad**.

Pearls Before Swine

She is a live music goer, and her hobby is **watching movies**.

Demo

- http://nlp.cs.washington.edu/denspi



http://nlp.cs.washington.edu/denspi

Conclusion

- "Read" entire Wikipedia in 0.5s with CPUs
- Query-agnostic, *indexable* phrase representations
- Utilize both *dense* (BERT-based) and *sparse* (bag-of-word) representations for encoding lexical, syntactic, and semantic information
- **6,000x lower computational cost** with higher accuracy for exact search
- At least **44x faster open-domain QA** with higher accuracy
- (query-agnostic) decomposability gap still exists (6-10%); we hope future research can close the gap