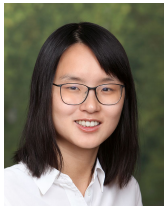
AUTOMATED KNOWLEDGE BASE 2021 | irvine CONSTRUCTION

Phrase Retrieval Learns Passage Retrieval, Too



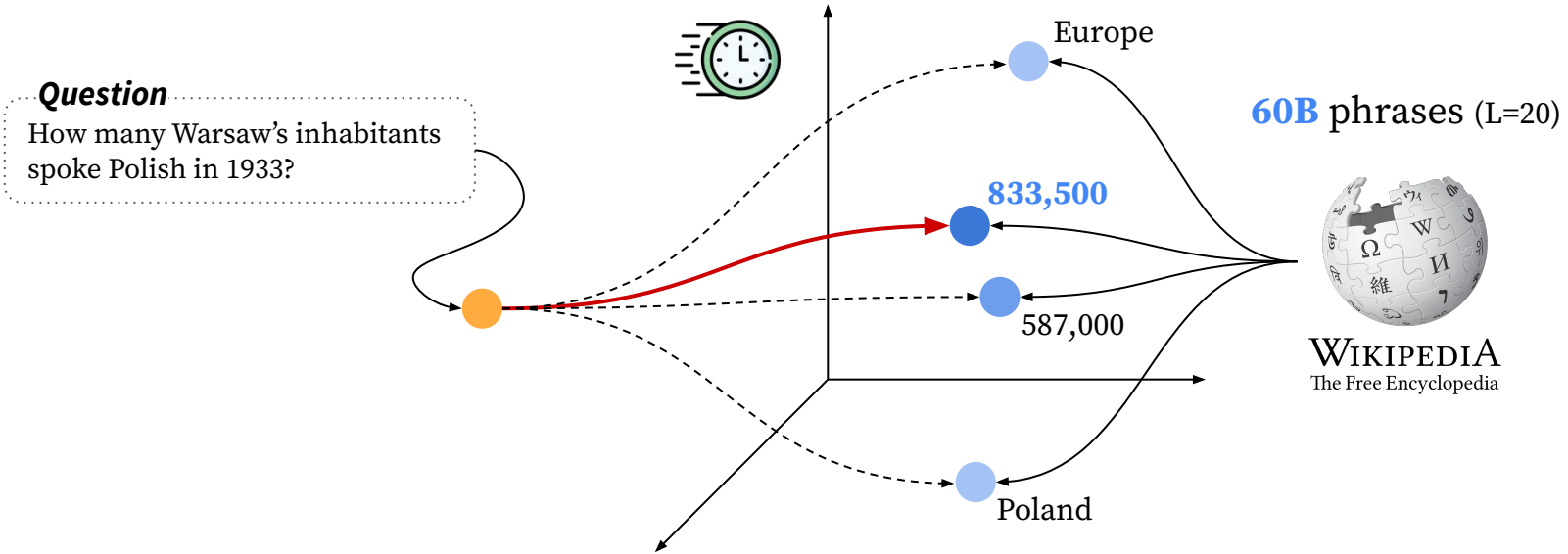


Jinhyuk Lee Alexander Wettig Danqi Chen

Department of Computer Science, Princeton University

Phrase Retrieval for Open-Domain QA

Phrase = any contiguous segment of text up to L words (Seo et al., 2019)



DensePhrases (Lee et al., 2021)

Fixed Granularities for Text Retrieval

Sentence Retrieval

SBERT (Reimers et a., 2019), SimCSE (Gao et al., 2021): **1 sentence** for each sentence vector

Passage Retrieval

ORQA (Lee et al., 2019): **288 subwords** for each passage vector

DPR (Karpukhin et al., 2020): **100 words** for each passage vector

...

Do we really need different methods for different granularities?

Phrase as a Basic Retrieval Unit

Retrieving **Phrases** $\Rightarrow$ Sentences $\Rightarrow$ Passages $\Rightarrow$ Documents $\Rightarrow$ ...

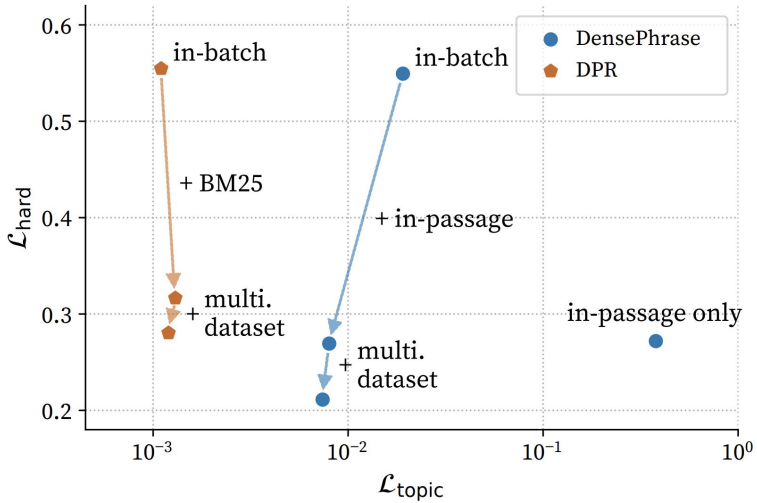
Using **DensePhrases** as a phrase retrieval model, we can retrieve:

```
In [3]: model.search('Who won the Nobel Prize in peace?', retrieval_unit='phrase', top_k=5)
Out[3]:
['Denis Mukwege,',
 'Theodore Roosevelt',
 'Denis Mukwege',
 'John Mott',
 'Mother Teresa']

In [4]: model.search('Why is the sky blue', retrieval_unit='sentence', top_k=1)
Out[4]: ['The blue color is sometimes wrongly attributed to Rayleigh scattering, which is respo
nsible for the color of the sky.']

In [5]: model.search('How to become a great researcher', retrieval_unit='paragraph', top_k=1)
Out[5]: ['Levine has also won awards in teaching and research, them most recent being the Provo
st\'s Award for University Outstanding Teacher of the Year, California State University, Fresno
 in 2007. He currently teaches social psychology, persuasion and mind control, supervised field
 experience in psychology, and other courses. In an interview with the Association for Psycholo
gical Science, Levine said he believes the key to being a great researcher is having passion fo
r research in and working on questions that the researcher is truly curious about. He said: "Ha
ve patience, persistence and enthusiasm and you'll be fine."']

In [6]: model.search('What is the history of internet', retrieval_unit='document', top_k=3)
Out[6]:
['Computer network',
 'History of the World Wide Web',
 'History of the Internet']
```

66.8 (DPR) vs. **69.9** (DensePhrases) Top-5 Passage Retrieval Accuracy

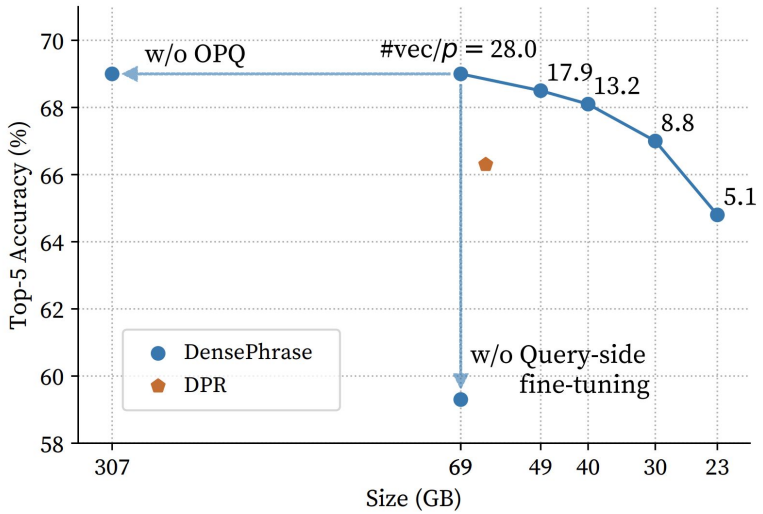
Works well for **document retrieval** (entity linking, dialogue), too

Why Does It Work?



Why does this happen?
Differentiates hard negatives better!

How Efficient is It?



300GB => **20GB** (vs. DPR: 62GB)

Interpretation as multi-vector encodings

Code & models available at https://github.com/princeton-nlp/DensePhrases