Taxonomy-enhanced Document Retrieval with Dense Representations

Victor Mireles, **Artem Revenko**, Ioanna Lytra, Anna Breit, Julia Klezl

Semantic Web Company

May 2 @ iDSC 2023

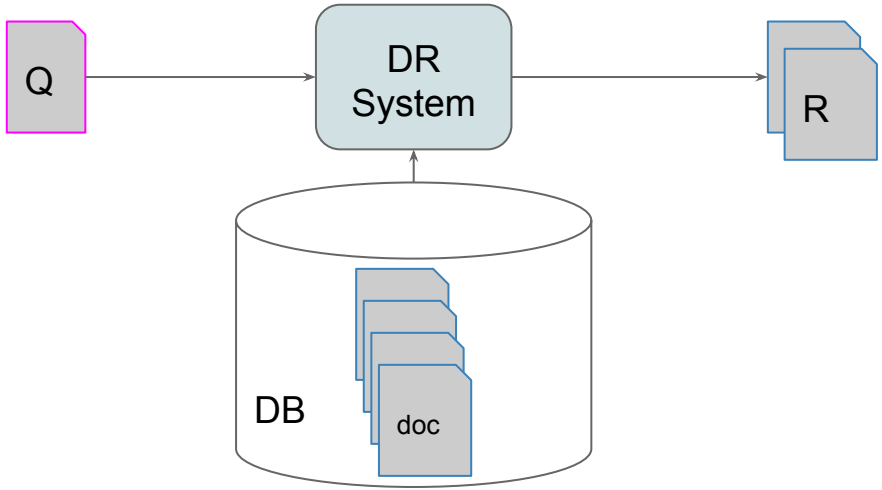
Motivation

Task Statement

SEMANTIC WEB COMPANY

Given a corpus and a query in natural language

Find the corpus documents that best match the query



Why retrieving documents?

SEMANTIC WEB COMPANY

*Document retrieval … is essential in various applications, such as **search engines**, **recommendation systems**, **question answering** systems, and more … and its **accuracy** and **effectiveness** are crucial in ensuring a good user experience.*

Why retrieving documents?

SEMANTIC WEB COMPANY

*Document retrieval … is essential in various applications, such as **search engines**, **recommendation systems**, **question answering** systems, and more … and its **accuracy** and **effectiveness** are crucial in ensuring a good user experience.*

ChatGPT

Experimental Setup

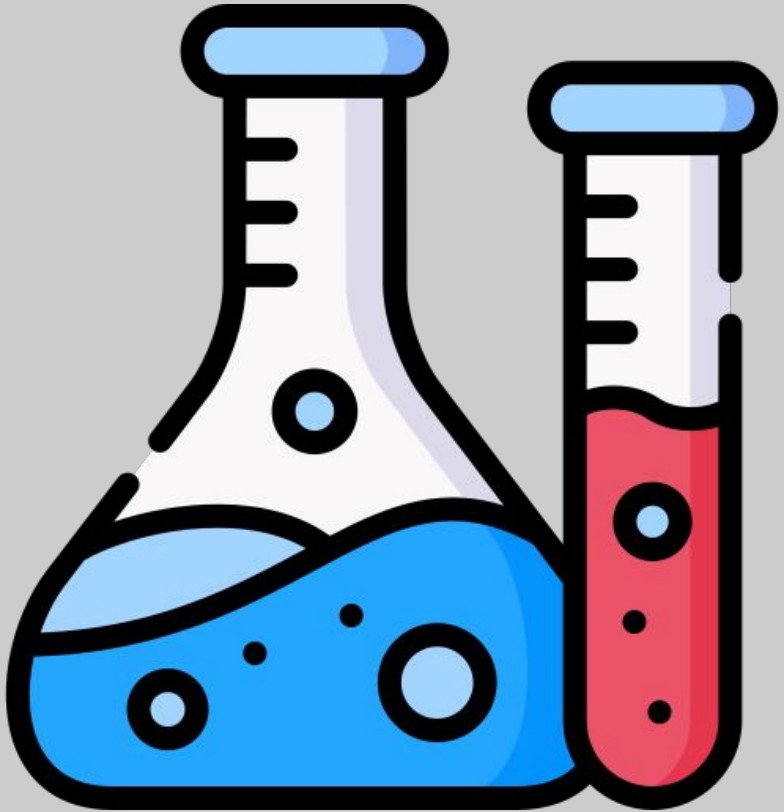


Image: Flaticon.com

PoolParty Help

PoolParty – semantic middleware by SWC. www.poolparty.biz

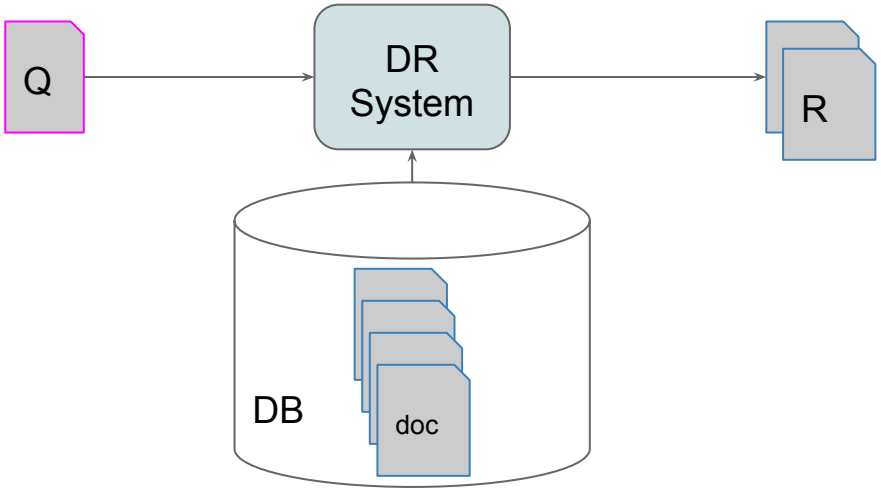
Dedicated help portal: help.poolparty.biz .

Dataset: from ~400 real search queries we manually select clean answers to 50 queries, including alternatives pages.

Challenges:

1. PP is an IT platform rich with many innovative functionalities
2. Various personas (roles) require various depth

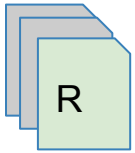
Hits @ k metrics

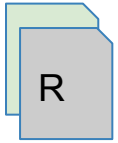


Hits@k =

|found docs in positions <= k| /

|all relevant documents|

1 

2 

Hits@1 = ½ = 0.5

Hits@3 = 1

Our Approach

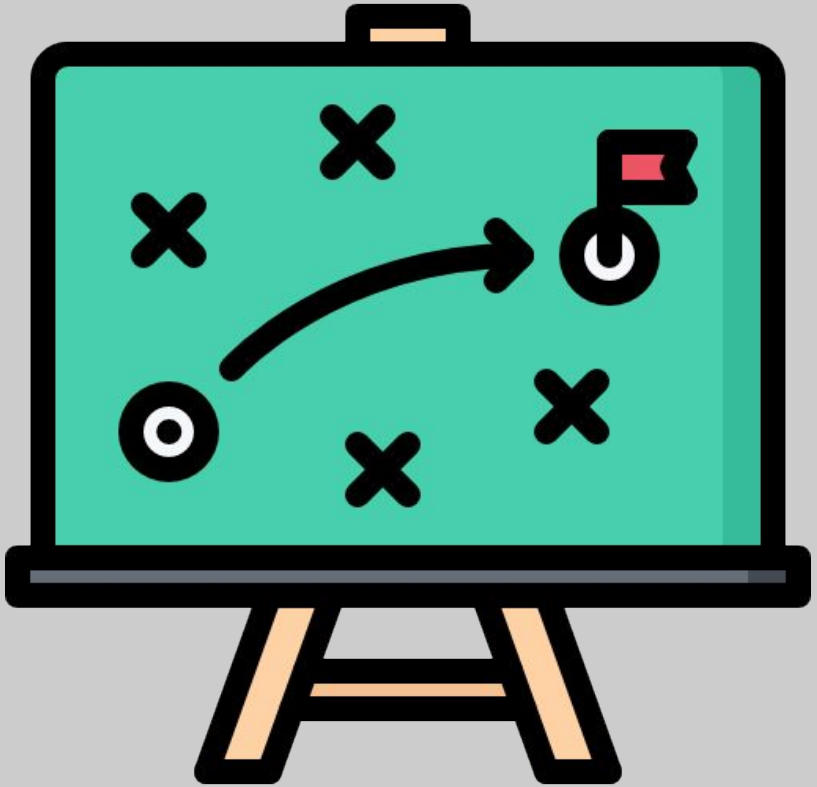
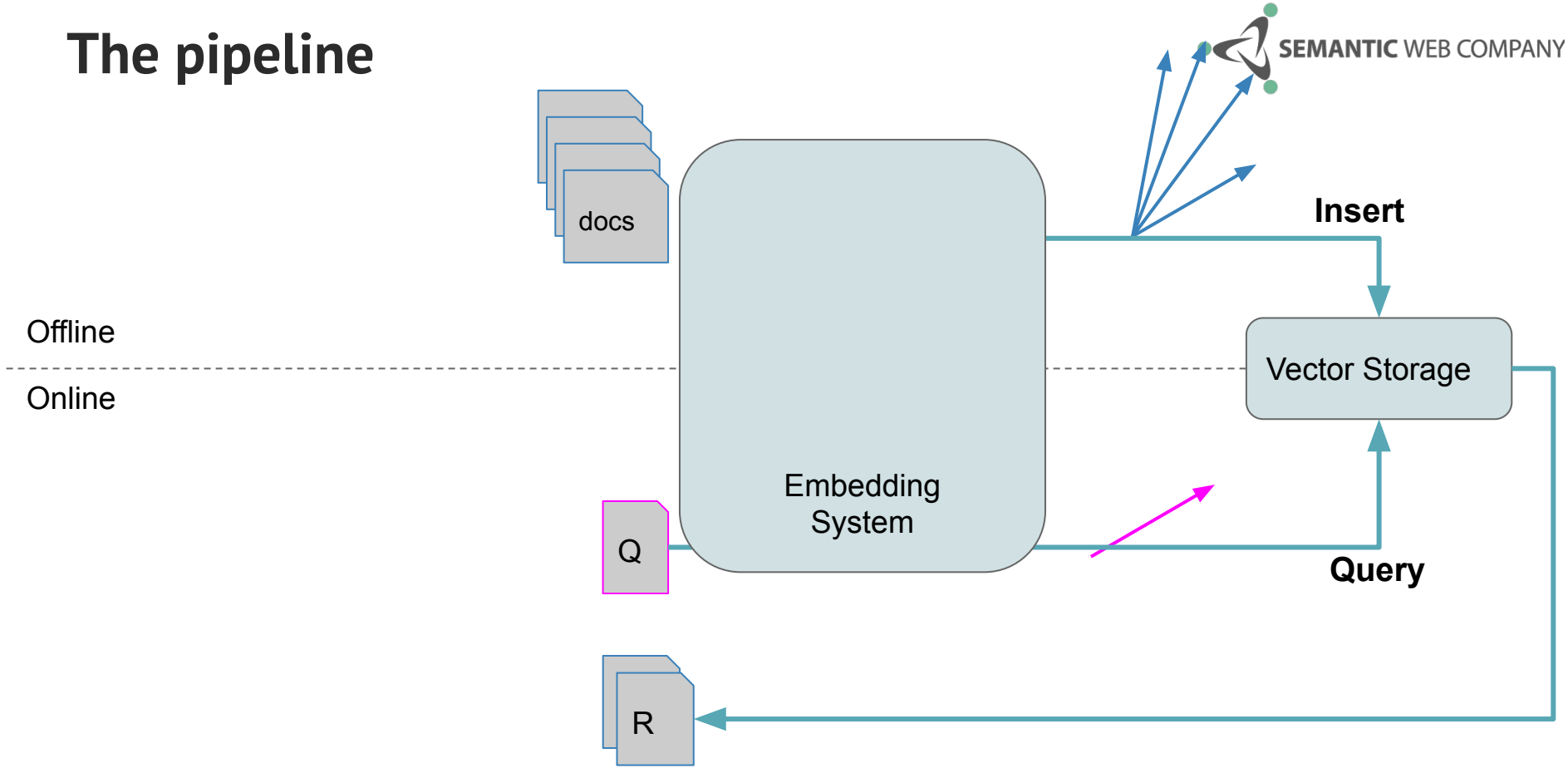


Image: Flaticon.com

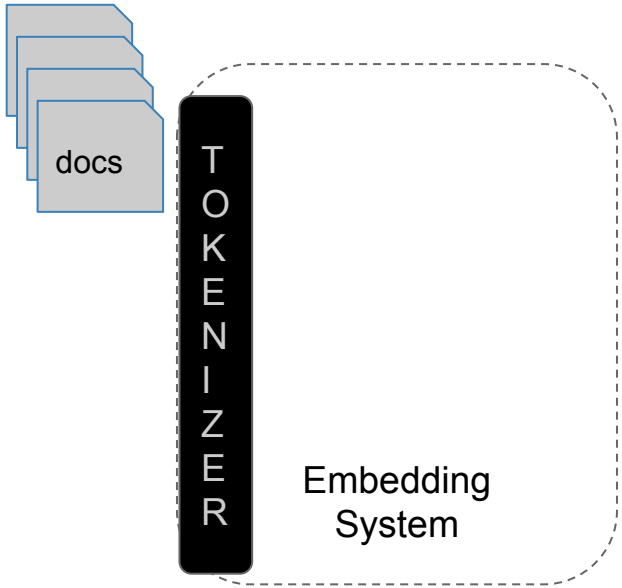
The pipeline



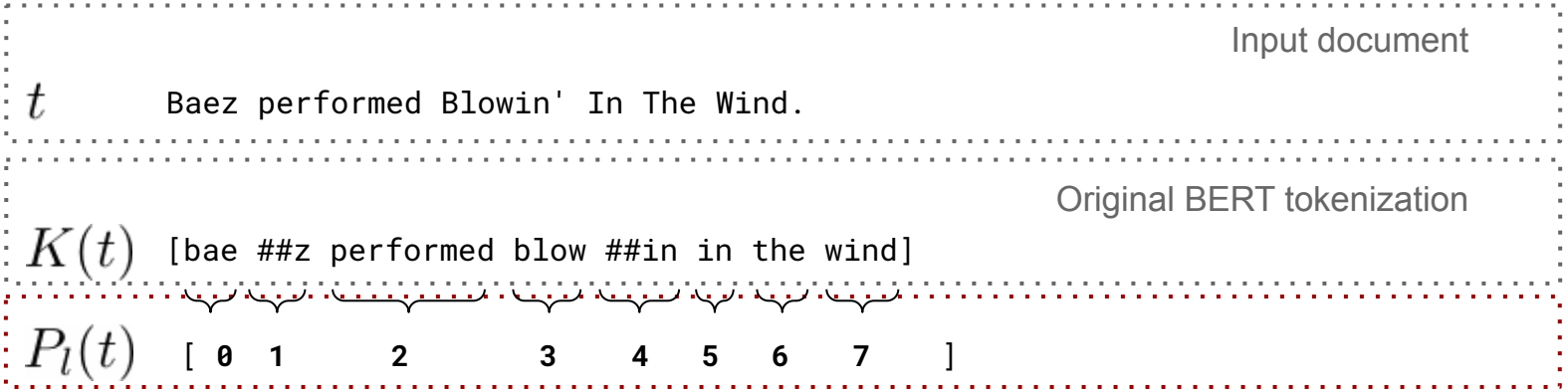
SEMANTIC WEB COMPANY

Offline

Document embeddings

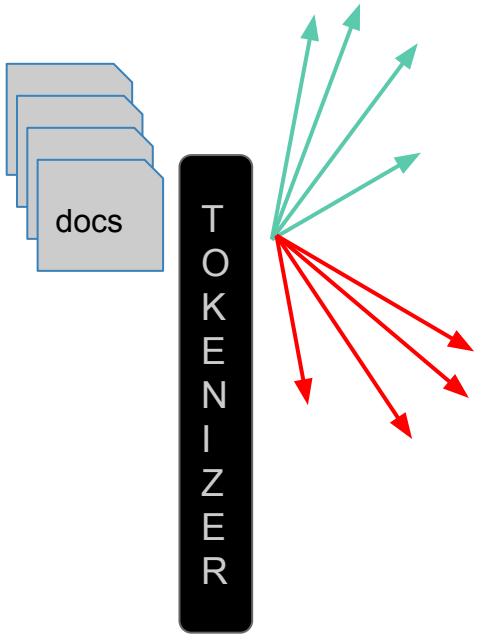


Document embeddings

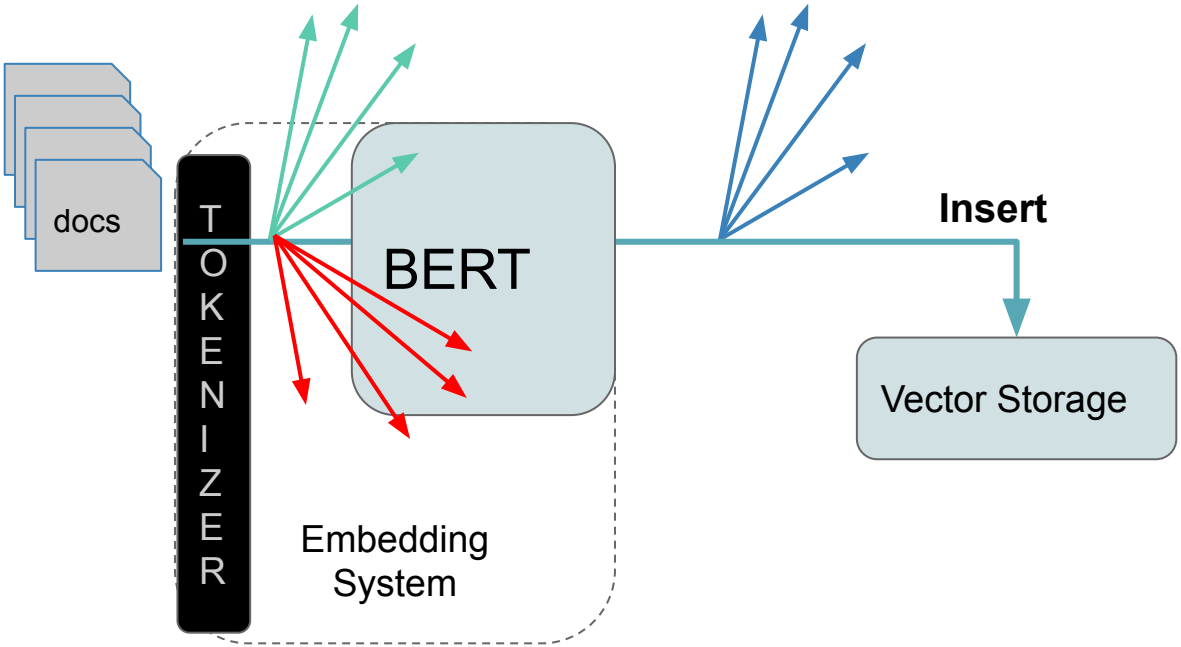


Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Document embeddings



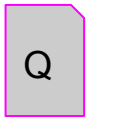
Document embeddings

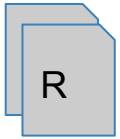


SEMANTIC WEB COMPANY

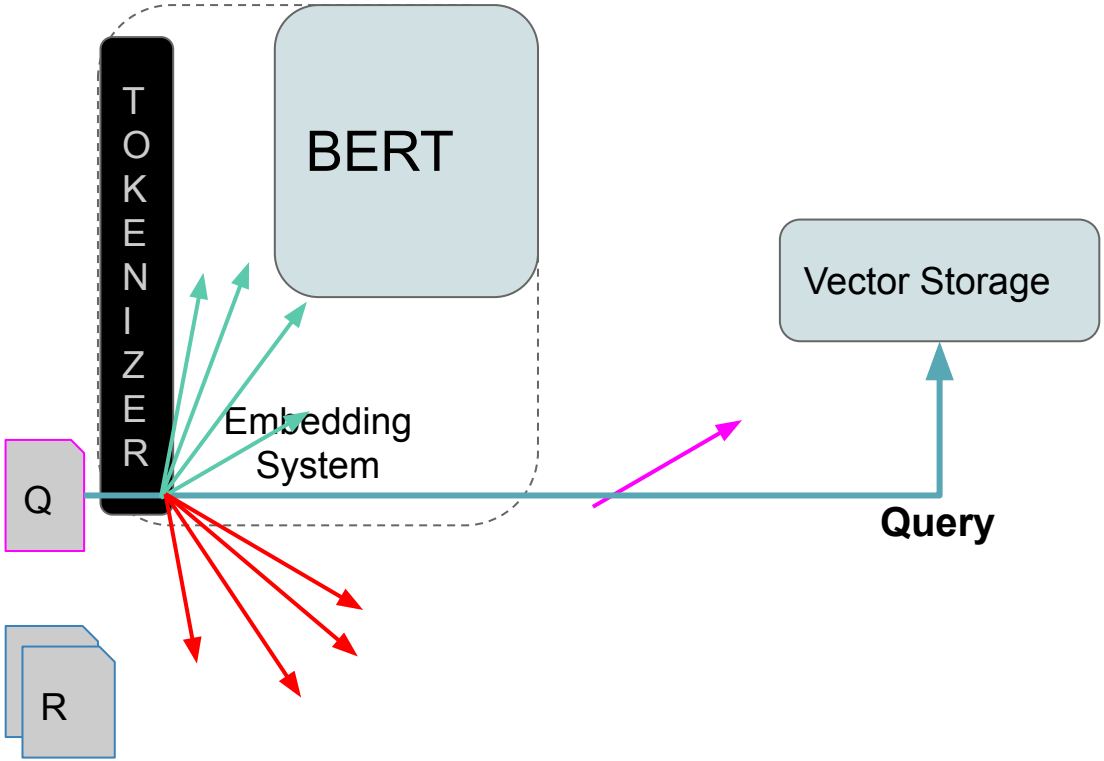
Online

Passage Retrieval





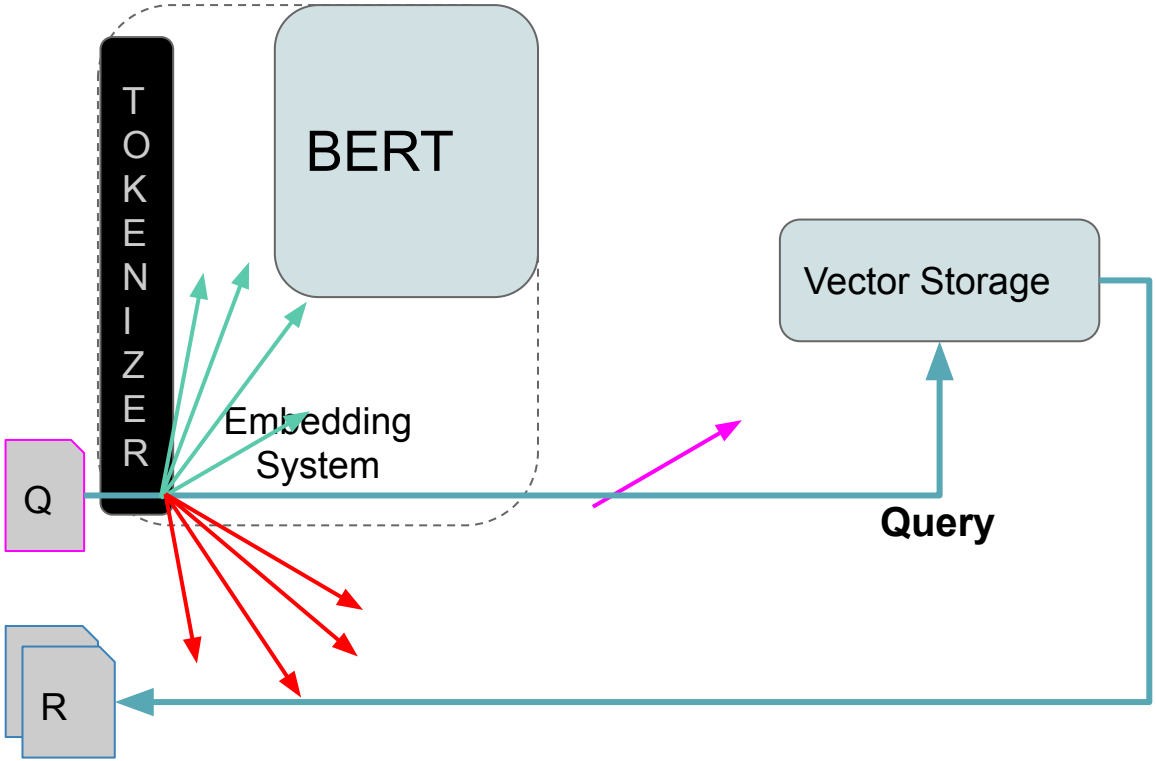
Passage Retrieval



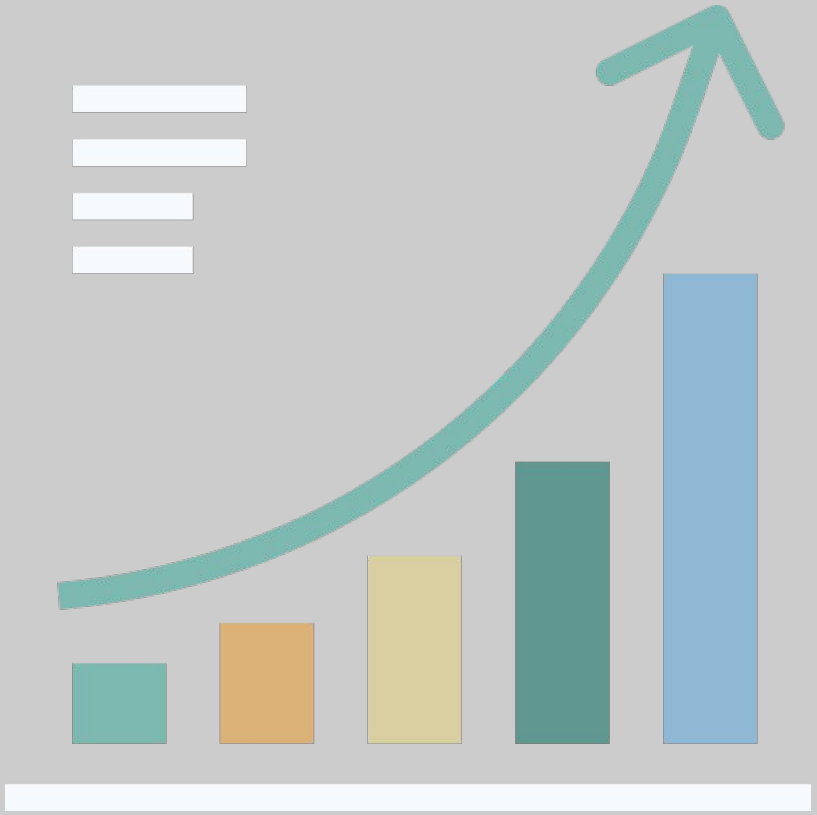
Passage Retrieval

SEMANTIC WEB COMPANY

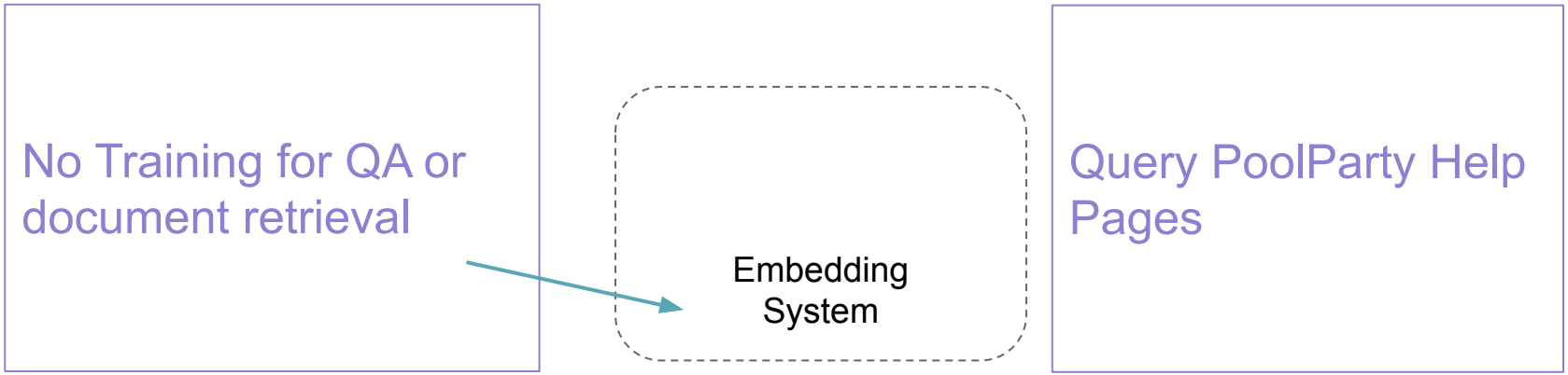
Using Cosine Similarity



Evaluation

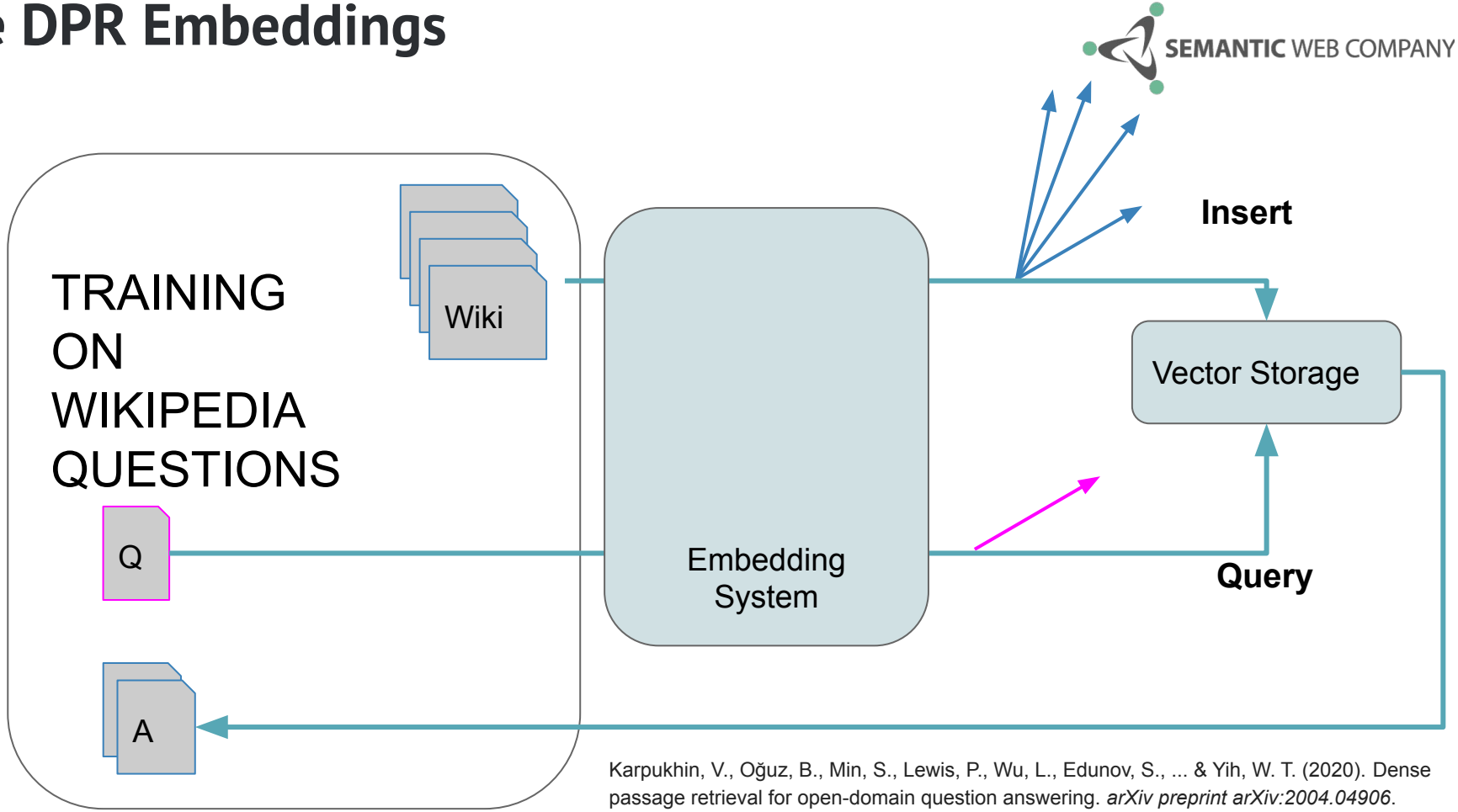


The BERT Embeddings



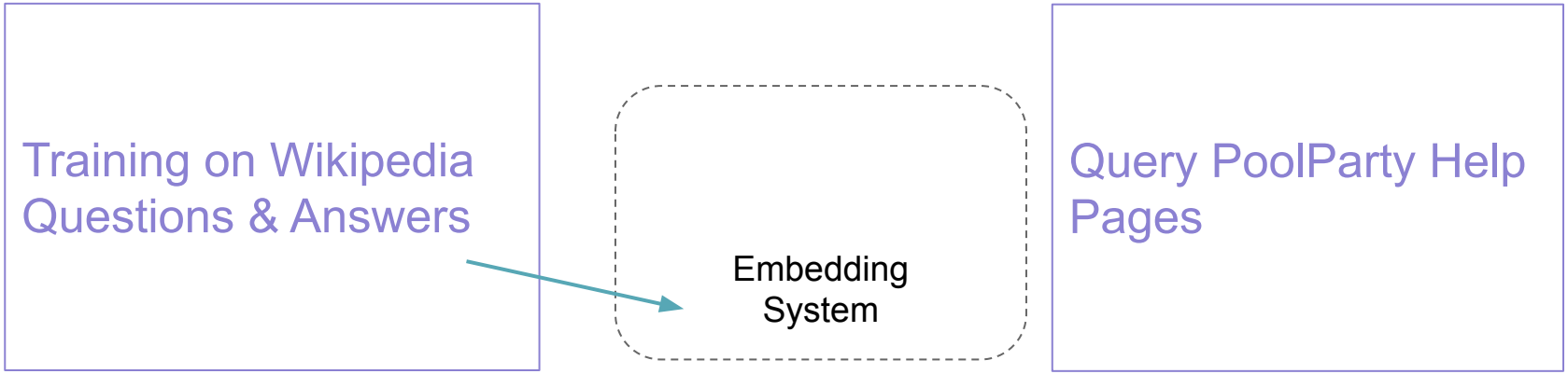
HF BERT without fine-tuning, only text	0.000	0.000	0.000	0.021

The DPR Embeddings



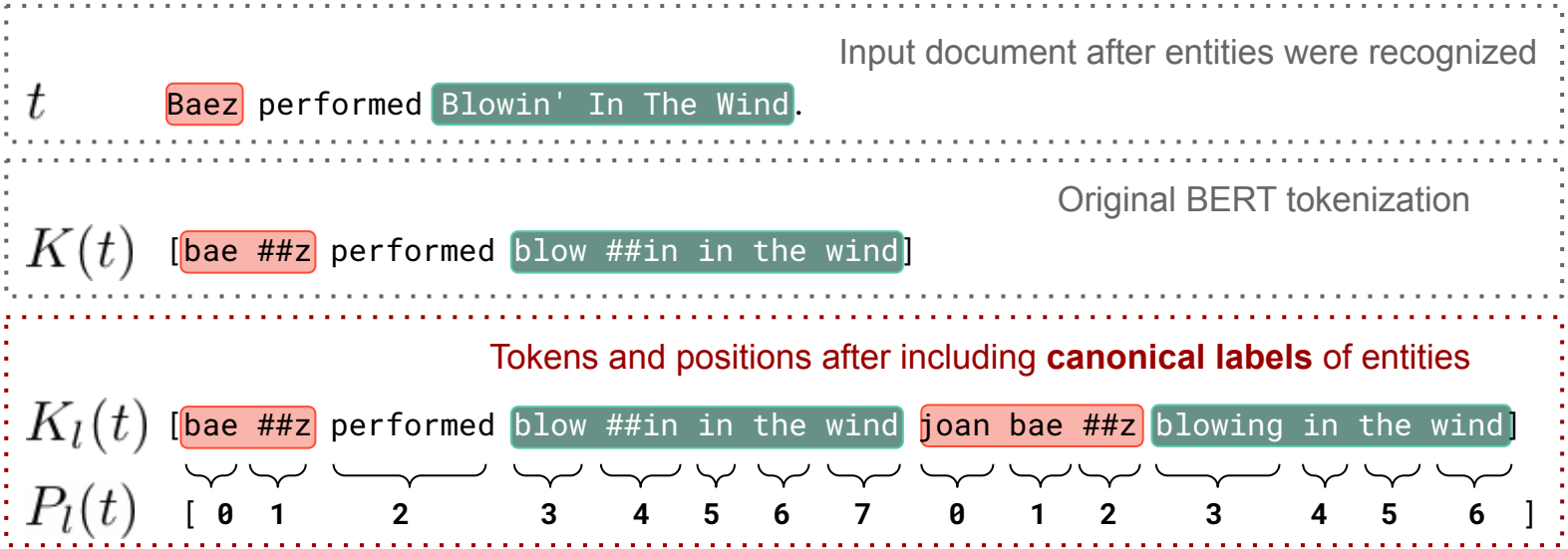
Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., ... & Yih, W. T. (2020). Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

The DPR Embeddings - On PP Help Pages



Vectorization	Hits @1	Hits @3	Hits @5	Hits @10
DPR only with text	**0.085**	0.191	0.255	0.532

Canonical Label Infusion

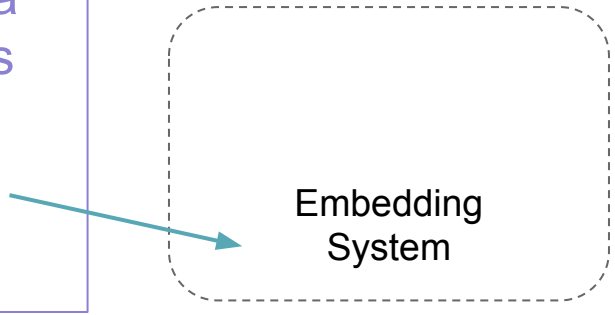


Canonical Label Infusion on PP Help Pages

SEMANTIC WEB COMPANY

Training on Wikipedia Questions & Answers

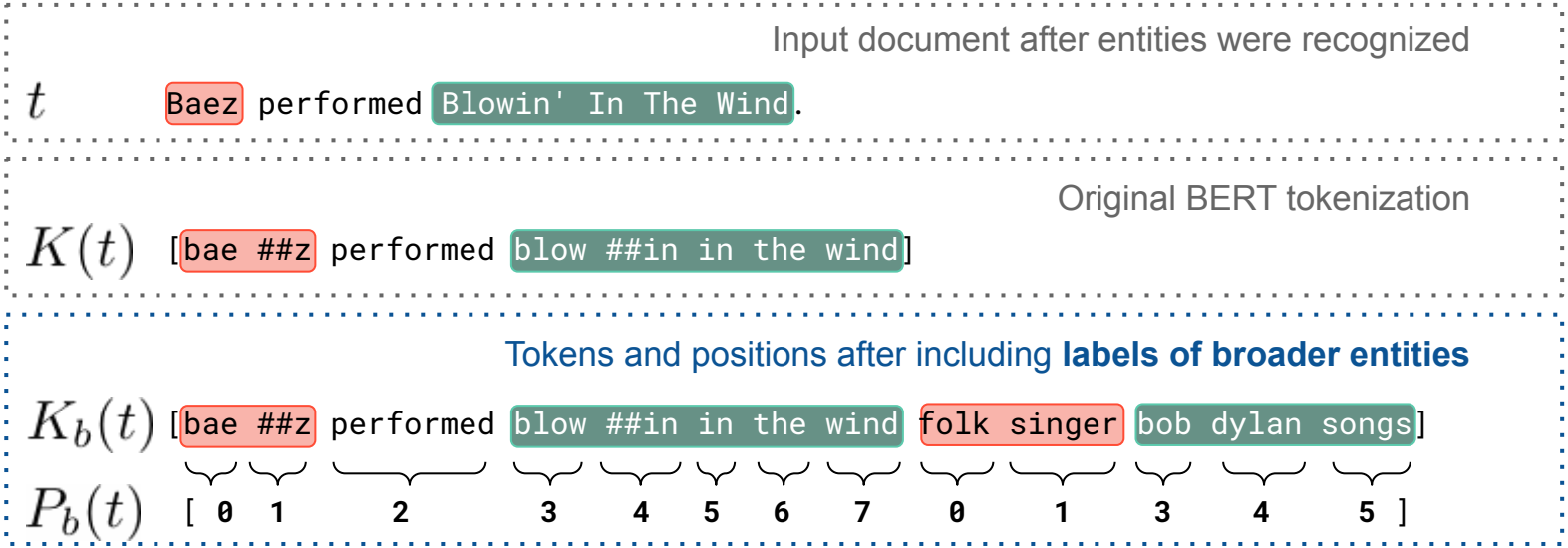
Use Concept's prefLabels



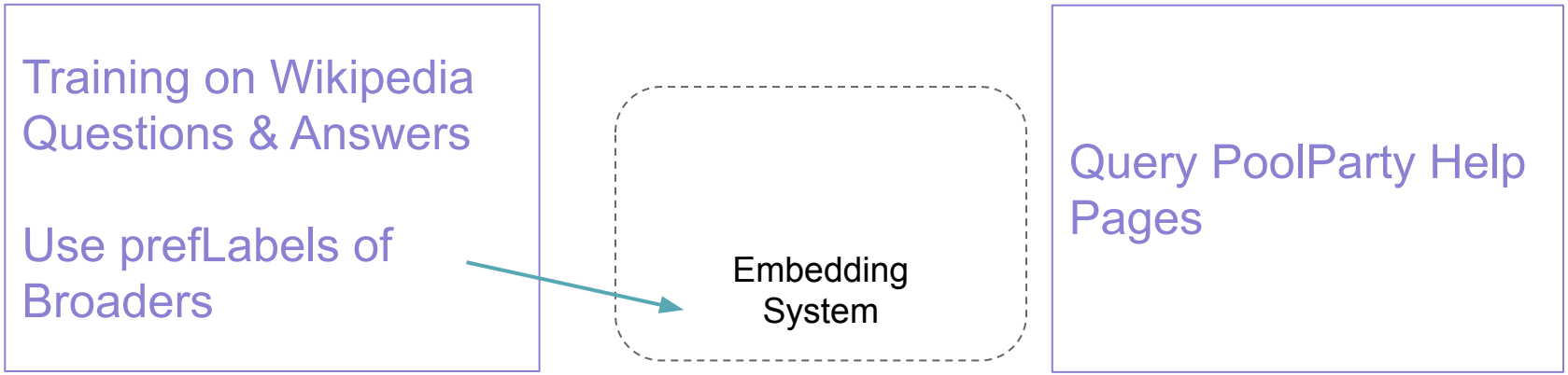
Query PoolParty Help Pages

Vectorization	Hits @1	Hits @3	Hits @5	Hits @10
DPR with canonical labels	**0.085**	0.255	**0.404**	**0.574**
DPR only with text	**0.085**	0.191	0.255	0.532

Broader Label Infusion



Broader Label Infusion on PP Help Pages



Vectorization	Hits @1	Hits @3	Hits @5	Hits @10
DPR with canonical labels	**0.085**	0.255	**0.404**	**0.574**
DPR with labels of broader entities	0.064	0.191	0.319	0.426
DPR only with text	**0.085**	0.191	0.255	0.532

Final Results

SEMANTIC WEB COMPANY

Vectorization	Hits @1	Hits @3	Hits @5	Hits @10
DPR with canonical labels	**0.085**	0.255	**0.404**	**0.574**
DPR with labels of broader entities	0.064	0.191	0.319	0.426
DPR only with text	**0.085**	0.191	0.255	0.532
Baseline TF-IDF	**0.234**	**0.319**	0.340	0.426

Examples

Using concept labels can weight words in better ways than TFIDF

Answer with TF-IDF:
“Import Assistant - Results and repair functions”

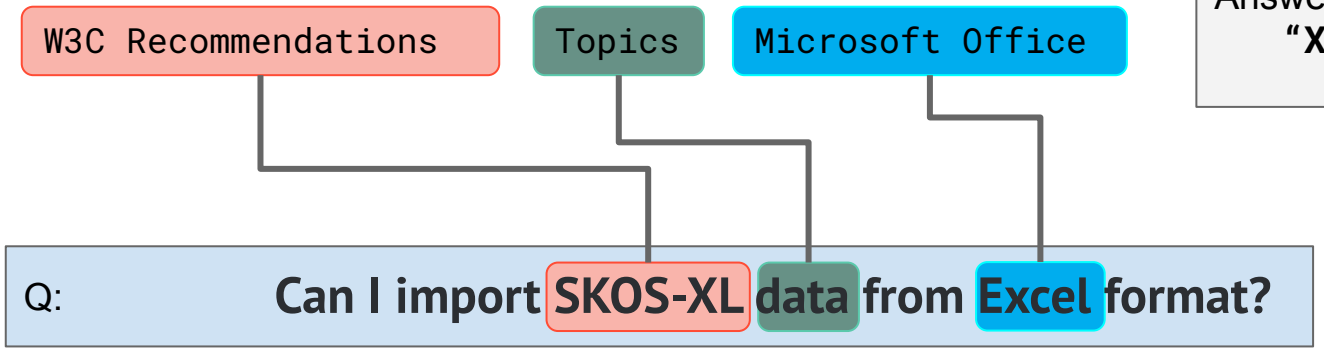
Mentions “import” and “data” very of often

Q: **Can I import SKOS-XL data from Excel format?**

Answer with Concepts:
“The PoolParty Excel Format”

Examples

Labels of broader concepts might be misleading



Answer with Broaders:
“XSLT”

Answer with Concepts:
“The PoolParty Excel Format”

Examples

TF-IDF is better when there are not enough concepts

Answer with Concepts:
“Access the user roles in PoolParty”

Q: **How can I create a new user?**

Answer with TF-IDF:
“Create a new user”


CONCLUSIONS

Conclusion & Open Topics

SEMANTIC WEB COMPANY

1. We demonstrate that knowledge infusion can improve state of the art DR methods
   a. We tried different strategies: canonical label, broaders.
2. Term-based baselines still perform better in certain scenarios
   a. In general DPR (and similar methods like REALM) has shown to be superior to term-based methods. But in case of complex domain-specific dataset concept semantics might be lost due to word-piece embedding strategy
      i. Other strategies to produce embeddings?
3. Small dataset is not conclusive
   a. We are currently extending the current dataset (challenging, but ~200 Qs)
   b. Further datasets, for example, chemical StackExchange

Thank You!

The work presented in this article has received funding from the Eureka Eurostars programme, Grant Number E114154 in frames of the PORQUE project

https://porque-project.eu/

Artem Revenko (speaker) contacts:

artem.revenko@semantic-web.com

https://twitter.com/revenkoartem

https://medium.com/@revenkoartem

For more fascinating research visit our page: https://semantic-web.com/research/