Searching by Code: a New SearchBySnippet Dataset and SnippeR Retrieval Model for Searching by Code Snippets

Ivan Sedykh, Dmitry Abulkhanov, Nikita Sorokin,

Sergey Nikolenko, and Valentin Malykh

Key Contributions

- novel “search by snippet” code search task
- novel dataset
- novel model

Task Description

- Classic Code Search:
  - docstring -> function body
- Clone Detection:
  - function body -> function body
- Search by Snippet:
  - code snippet -> textual description

```
import math
def pythagorus():
    """
    This module takes an input of a, then b, to calculate c in the equation

    sqrt(a^2+b^2) = c
    """
    a,b = int(input("a?")),int(input("b?"))
    c = math.sqrt(a**2+b**2)
    print(c)
```

Data

- StackOverflow **Python**-related questions
  - Answers
  - Comments
  - Other meta-data
- **2 000 000** question-answer pairs
- Test set
  - full-duplicate questions

Create ArrayList from array

Asked 13 years, 8 months ago Modified 9 days ago Viewed 1.7m times

3960

667

I have an array that is initialized like:

```
Element[] array = {new Element(1), new Element(2), new Element(3)};
```

I would like to convert this array into an object of the `ArrayList` class.

```
ArrayList<Element> arraylist = ???;
```

java arrays arraylist type-conversion

Share Follow

edited Jan 31, 2020 at 5:25
JuanMoreno
1,991 ● 1 ● 20 ● 30

asked Oct 1, 2008 at 14:38
Ron Tuffin
52.2k ● 24 ● 63 ● 76

Add a comment

41 Answers

Sorted by: Highest score (default)

1 2 Next

4969

```
new ArrayList<>(Arrays.asList(array));
```

Share Follow

edited Feb 18, 2020 at 1:53
Unmitigated
46.1k ● 7 ● 44 ● 60

answered Oct 1, 2008 at 14:39
Tom
54.5k ● 3 ● 26 ● 35

Train set

- code + traceback is treated like query to find relevant StackOverflow post

```
{
    "post_id": 12560975,
    "related_post_id": 1767934,
    "question": "import httplib
                IP = "local ip"
                v_params = ""
                conn = httplib.HTTPConnection(IP,8021)
                conn.request("GET", v_params)
                response = conn.getresponse()
                data_response = response.read()
                Traceback(most recent call last):
                File "C:\Python24\lib\httplib.py", line 866, in getresponse
                response.begin()
                File "C:\Python24\lib\httplib.py", line 336, in begin
                version, status, reason = self._read_status()
                File "C:\Python24\lib\httplib.py", line 300, in _read_status
                raise BadStatusLine(line)
                httplib.BadStatusLine"
}
```

code snippet

traceback

Figure 3: Sample query.

```
{
    "post_id": 1767934,
    "text": "Why am I getting this error in python ? (httplib)
            <pre><code>if theurl.startswith("http://"): theurl = theurl[7:]
            head = theurl[:theurl.find('/')]
            tail = theurl[theurl.find('/'):]
            response_code = 0
            import httplib
            conn = httplib.HTTPConnection(head)
            conn.request("HEAD",tail)
            res = conn.getresponse()
            response_code = int(res.status)

            http://www.garageband.com/mp3cat/.UZCKbS6N4qk/01_Saraenglish.mp3
            Traceback (most recent call last):
            File "check_data_404.py", line 51, in &lt;module&gt;
                run()
            File "check_data_404.py", line 35, in run
                res = conn.getresponse()
            File "/usr/lib/python2.6/httplib.py", line 950, in getresponse
                response.begin()
            File "/usr/lib/python2.6/httplib.py", line 390, in begin
                version, status, reason = self._read_status()
            File "/usr/lib/python2.6/httplib.py", line 354, in _read_status
                raise BadStatusLine(line)
            httplib.BadStatusLine
            </code></pre>

            <p>Does anyone know what "Bad Status Line" is?</p>

            <p>Edit: I tried this for many servers, and many URL's
            and I still get this error?</p>
            <p>The Python Standard Library:
            <a href="https://docs.python.org/2/library/httplib.html"
            rel="nofollow noreferrer">httplib (Python 2)</a>
            (called <a href="https://docs.python.org/3/library/http.client.html"
            rel="nofollow noreferrer">http.client in Python 3</a>):</p>

            <blockquote>
            <p><code>exception httplib.BadStatusLine</code><br>
            A subclass of HTTPException. Raised if a server responds with
            a HTTP status code that we dont understand.</p>
            </blockquote>"
}
```

post title

post body

best answer

Figure 4: Sample document.

SearchBySnippet Dataset

- 900k questions in train set
- 500 questions in test set


```
{
    "text": "<p>How to find out what week number is
    current year on June 16th (wk24) with Python? </p>\n",
    "score": None,
    "probability": None,
    "question": None,
    "meta": {
        "post_id": "2600775",
        "accepted_answer_id": "0",
        "answer_count": "14",
        "closed_date": None,
        "comment_count": "3",
        "community_owned_date": None,
        "creation_date": "2010-04-08 14:35:53",
        "favorite_count": "56",
        "last_ activity_ date": "2020-03-18 06:55:44",
        "last_edit_ date": "2017-12-13 07:05:52",
        "last_editor_display name": None,
        "last_editor_user_id": "6790377",
        "owner_user_id": "311996",
        "parent_id": "0",
        "post_type_id": "1",
        "score": "334",
        "tags": "<python><datetime><week-number>",
        "title": "How to get week number in Python?",
        "view_count": "265693",
        "python_tag": True
    },
    "embedding": None,
    "id": "13ca360cb9c37d421e70e0d09c9a3e7dfil"
}
```


Figure 2: Data sample from the public dump.

Model

- Single Encoder Retriever
- Based on PLM
  - We tried several, the best is GraphCodeBERT
- Hard-negative mining for self-training

$$\mathcal{L}(q, d^+, D^-) = -\log \frac{e^{\text{score}(q,d^+)}}{\sum_{j=1}^{n} e^{\text{score}(q,d_j^-)} + e^{\text{score}(q,d^+)}}$$

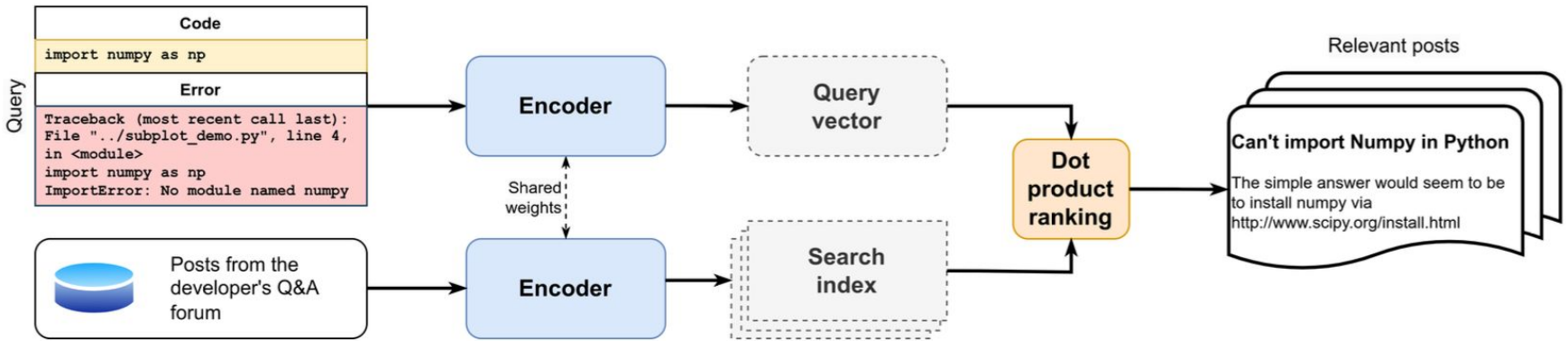
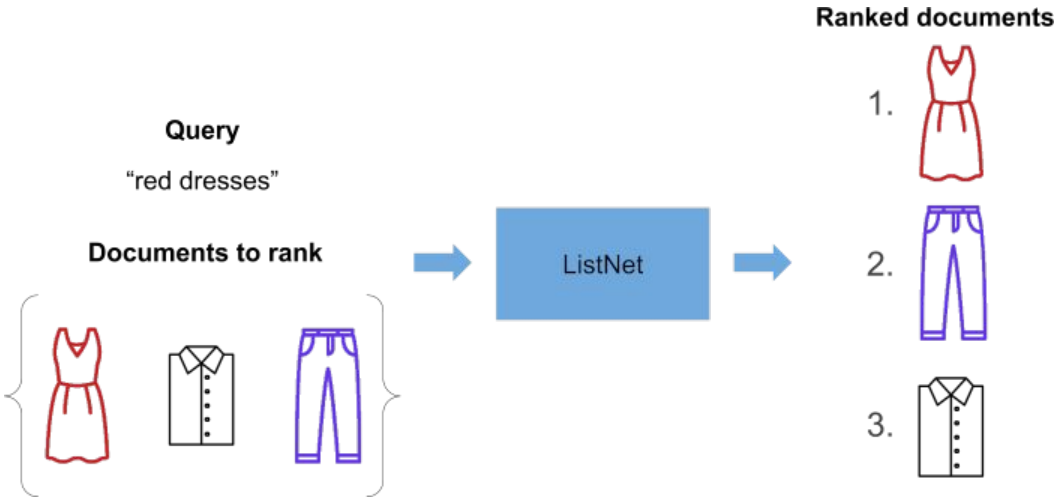


Figure 1: Overview of the problem setting and system design.

Negative
Hard-Negative



Self-training

- the model is used to obtain hard-negatives for the next training iteration

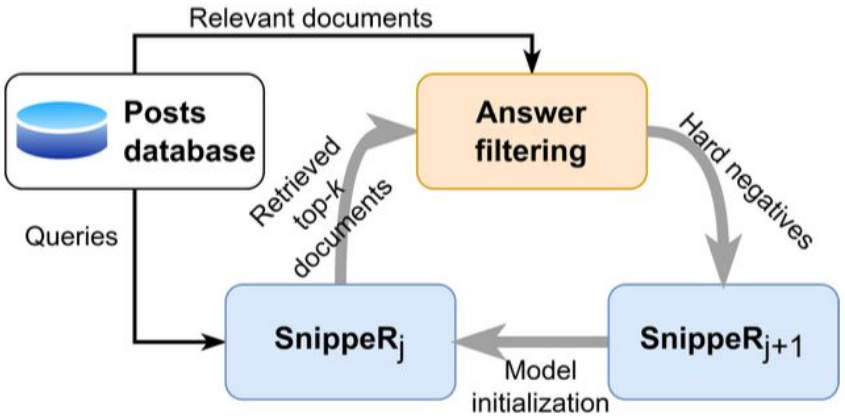


Figure 2: Self-training framework.

Results

- existing models lose to simple BM25
- SnippeR outperforms BM25

Model	Recall @5	@10	@20	@50
GraphCodeBERT (Guo et al., 2021)	0.001	0.001	0.002	0.009
CodeBERT (Feng et al., 2020)	0.001	0.006	0.010	0.013
SynCoBERT (Wang et al., 2021)	0.006	0.010	0.013	0.020
GraphCodeBERT (+CSN)	0.161	0.221	0.280	0.367
BM25 (Robertson et al., 1994)	0.311	0.406	0.474	0.562
SnippeR	**0.338**	**0.451**	**0.536**	**0.657**

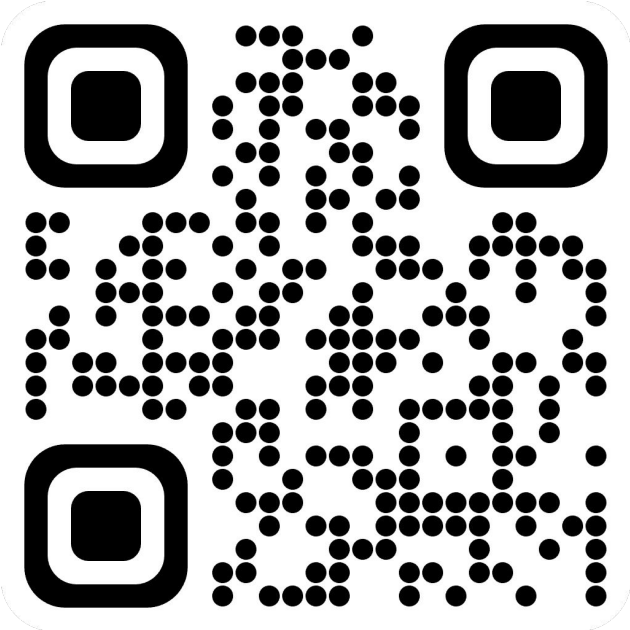
Table 3: Results on SearchBySnippet.

Conclusion

- Novel task is suggested to community.
- Novel significantly different dataset.
- Novel model effective for this task.

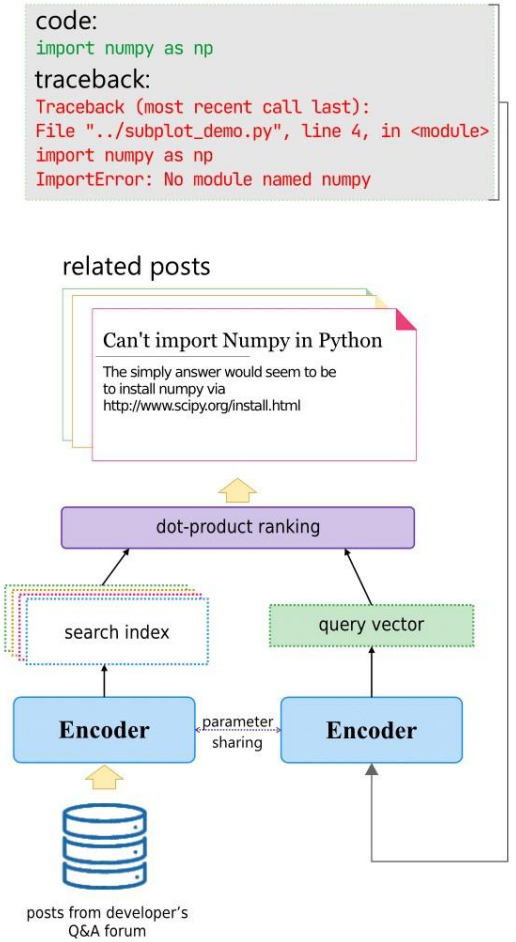
P.S. BM25 is still the best first choice!

My Contacts:



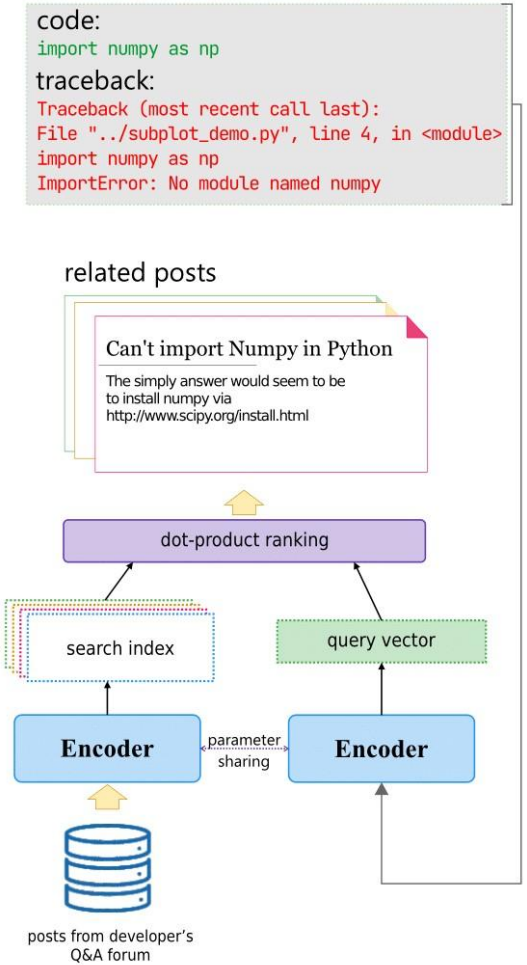
Problem Description

- Information retrieval
  - **Query**: code + traceback
  - **Documents**: answers from StackOverflow

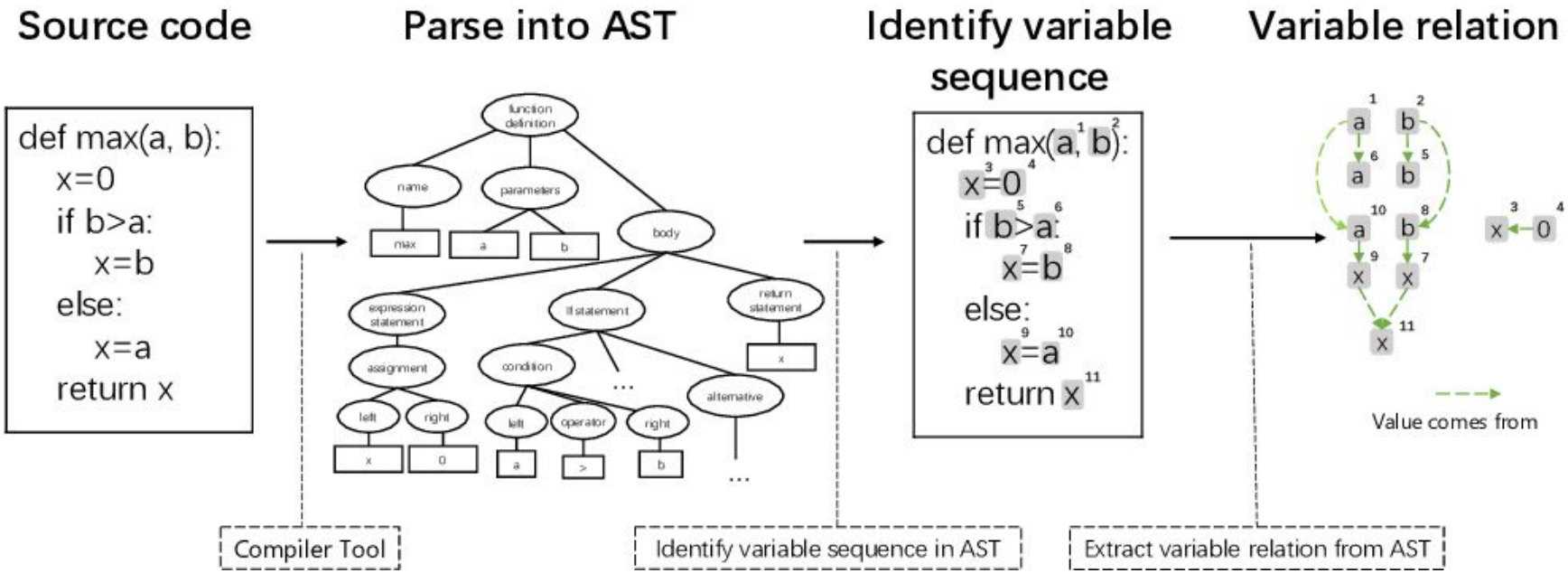


Model

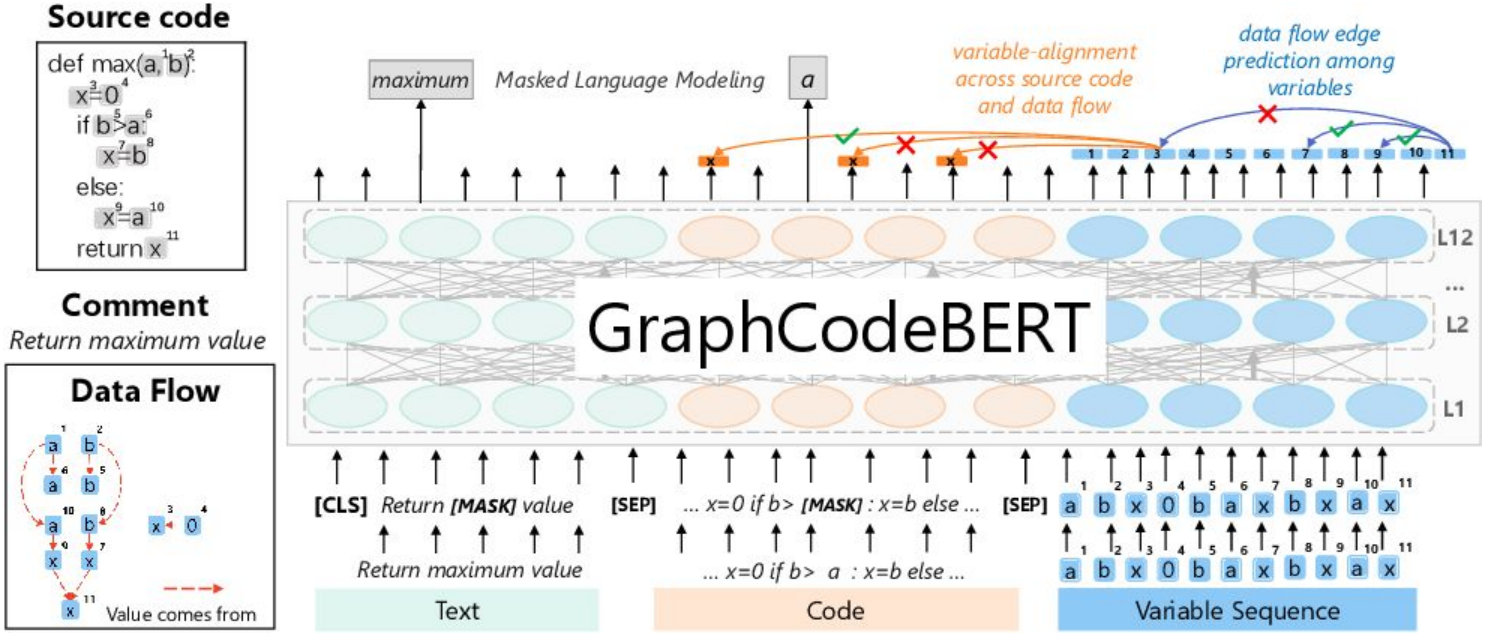
- Single Encoder Retriever
- Based on PLM
  - We tried several, the best is GraphCodeBERT
- Hard-negative mining for self-training



GraphCodeBERT

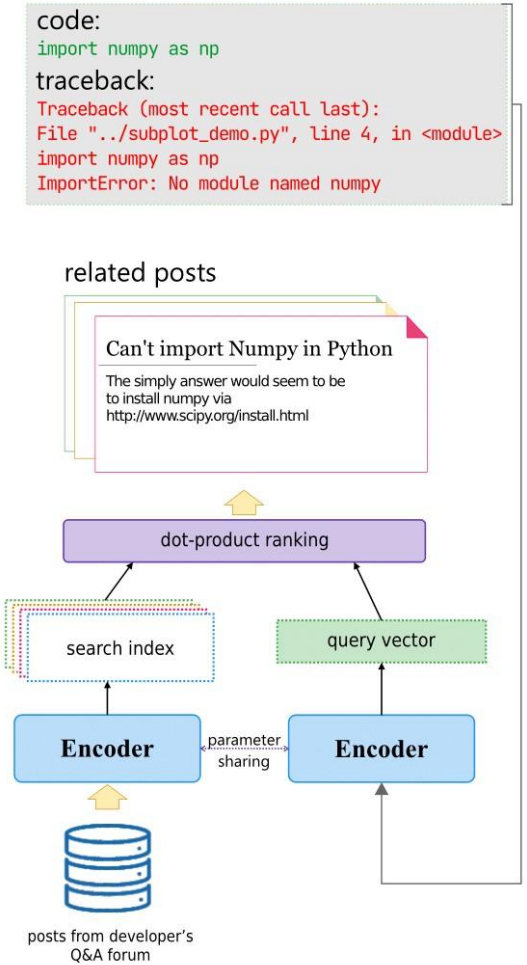


GraphCodeBERT

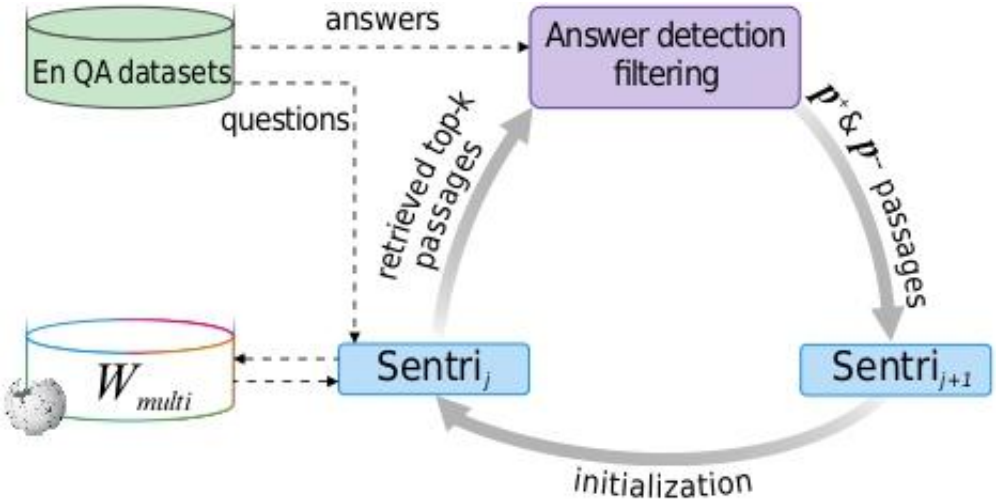


Model

- Single Encoder Retriever
- Based on PLM
  - We tried several, the best is GraphCodeBERT
- Hard-negative mining for self-training



Self-Training



Achieved Results

- We achieved 5% improvement over BM25
- We achieved more than 50% of Recall@10
- We transferred our model to Cloud IDE product

Model / Metric	Recall@5	Recall@10	Recall@20	Recall@50
GraphCodeBERT [12]	0.001	0.001	0.002	0.009
CodeBERT [8]	0.001	0.006	0.010	0.013
SynCoBERT [28]	0.006	0.010	0.013	0.020
GraphCodeBERT (CodeSearchNet)	0.161	0.221	0.280	0.367
BM25 [27]	0.311	0.406	0.474	0.562
CodeSentri (*ours*)	**0.338**	**0.451**	**0.536**	**0.657**

TABLE V

EVALUATION OF CODESENTRI AND THE BASELINE MODELS ON SEARCHBYSNIPPET DATASET.