Code and Named Entity Recognition in StackOverflow




Jeniya Tabassum, Mounica Maddela, Alan Ritter, Wei Xu

THE OHIO STATE UNIVERSITY

StackOverflow Entity Recognition

BRAT

I am trying to do a search query on the SoundCloud API. (Library)

Using their JavaScript (Language) SDK, (Application) the following works:

NEW NER CORPUS

15,372 Sentences
Manually Annotated
20 Entity Types

StackOverflow Entity Recognition

BRAT

I am trying to do a search query on the SoundCloud API. (Library)

Using their JavaScript (Language) SDK, (Application) the following works:

Available at: https://github.com/jeniyat/StackOverflowNER

NEW NER CORPUS

15,372 Sentences
Manually Annotated
20 Entity Types

StackOverflow Entity Recognition

BRAT

I am trying to do a search query on the SoundCloud API. (Library)

Using their JavaScript (Language) SDK, (Application) the following works:

Available at: https://github.com/jeniyat/StackOverflowNER

NEW NER CORPUS

15,372 Sentences
Manually Annotated
20 Entity Types

SoftNER

extracts the software entities with **79.1%** $\mathbf{F_1}$

StackOverflow Entity Recognition

1st

BRAT

I am trying to do a search query on the SoundCloud API. (Library)

Using their JavaScript (Language) SDK, (Application) the following works:

Available at: https://github.com/jeniyat/StackOverflowNER

NEW NER CORPUS

15,372 Sentences
Manually Annotated
20 Entity Types

SoftNER

extracts the software entities with **79.1% F_1**

Fine-tuned BERT$_{off\text{-}the\text{-}self}$

extracts the software entities with **57.5% F_1**

NEW NER MODEL

21.6 F_1↑

StackOverflow Entity Recognition

1st

BRAT

I am trying to do a search query on the SoundCloud API. (Library)

Using their JavaScript (Language) SDK, (Application) the following works:

Available at: https://github.com/jeniyat/StackOverflowNER

NEW NER CORPUS

15,372 Sentences
Manually Annotated
20 Entity Types

SoftNER

extracts the software entities with **79.1% F_1**

Fine-tuned BERT*off-the-self*

extracts the software entities with **57.5% F_1**

NEW NER MODEL

21.6 F_1↑

BERTOverflow
Pre-trained on 152M Sentences

StackOverflow NER

Code Retrieval

NL Query:

Using mysql find the sorted list of 200 most popular links with date earlier than 2014/02/25

Code Snippet

```
SELECT url FROM links WHERE date =
    ( SELECT date FROM links WHERE
    date < "2014/02/25" ORDER BY date
    DESC LIMIT 1)
ORDER BY date DESC, clicks DESC LIMIT 200
```

[Yao et al., 2019]
[Iyer et al. 2016]
[Giorgi and Bader 2018]

StackOverflow NER

Code Retrieval

NL Query:

Using mysql find the sorted list of 200 most popular links with date earlier than 2014/02/25

Code Snippet

```
SELECT url FROM links WHERE date =
    ( SELECT date FROM links WHERE
    date < "2014/02/25" ORDER BY date
    DESC LIMIT 1)
ORDER BY date DESC, clicks DESC LIMIT 200
```

[Yao et al., 2019]
[Iyer et al. 2016]
[Giorgi and Bader 2018]

StackOverflow NER

Code Retrieval

NL Query:

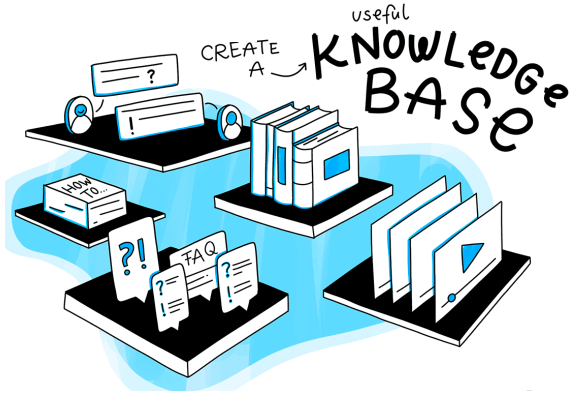
Using mysql find the sorted list of 200 most popular links with date earlier than 2014/02/25

Code Snippet

```
SELECT url FROM links WHERE date =
    ( SELECT date FROM links WHERE
    date < "2014/02/25" ORDER BY date
    DESC LIMIT 1)
ORDER BY date DESC, clicks DESC LIMIT 200
```

[Yao et al., 2019]
[Iyer et al. 2016]
[Giorgi and Bader 2018]

KB Creation



[Movshovitz-Attias and Cohen 2015]

StackOverflow NER

Code Retrieval

NL Query:

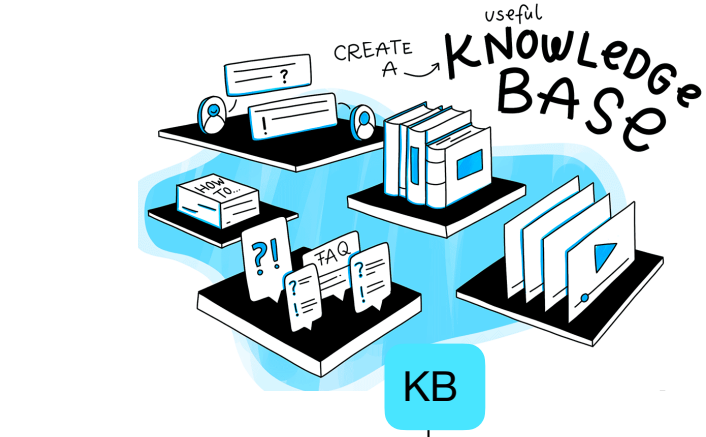
Using mysql find the sorted list of 200 most popular links with date earlier than 2014/02/25

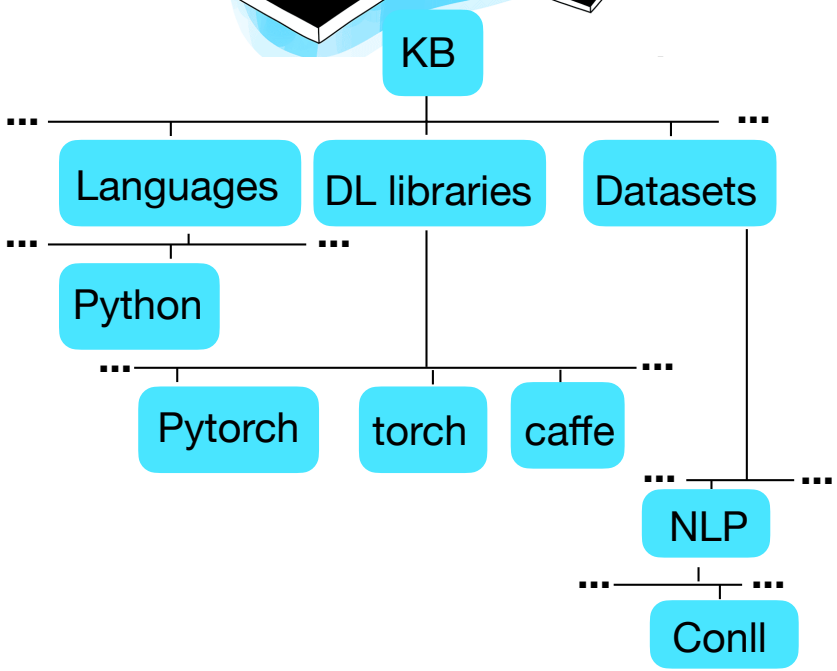
Code Snippet

```
SELECT url FROM links WHERE date =
    ( SELECT date FROM links WHERE
    date < "2014/02/25" ORDER BY date
    DESC LIMIT 1)
ORDER BY date DESC, clicks DESC LIMIT 200
```

[Yao et al., 2019]
[Iyer et al. 2016]
[Giorgi and Bader 2018]

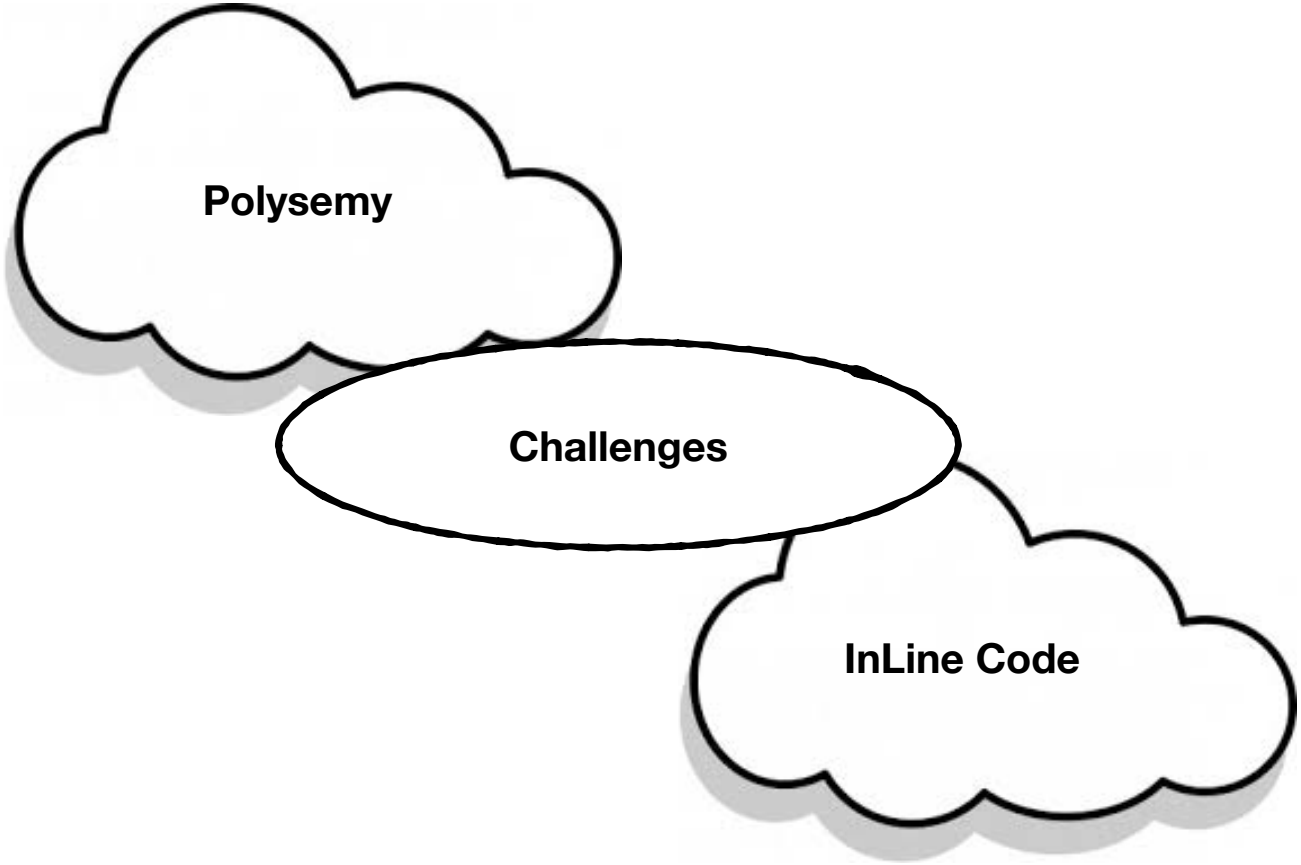
KB Creation



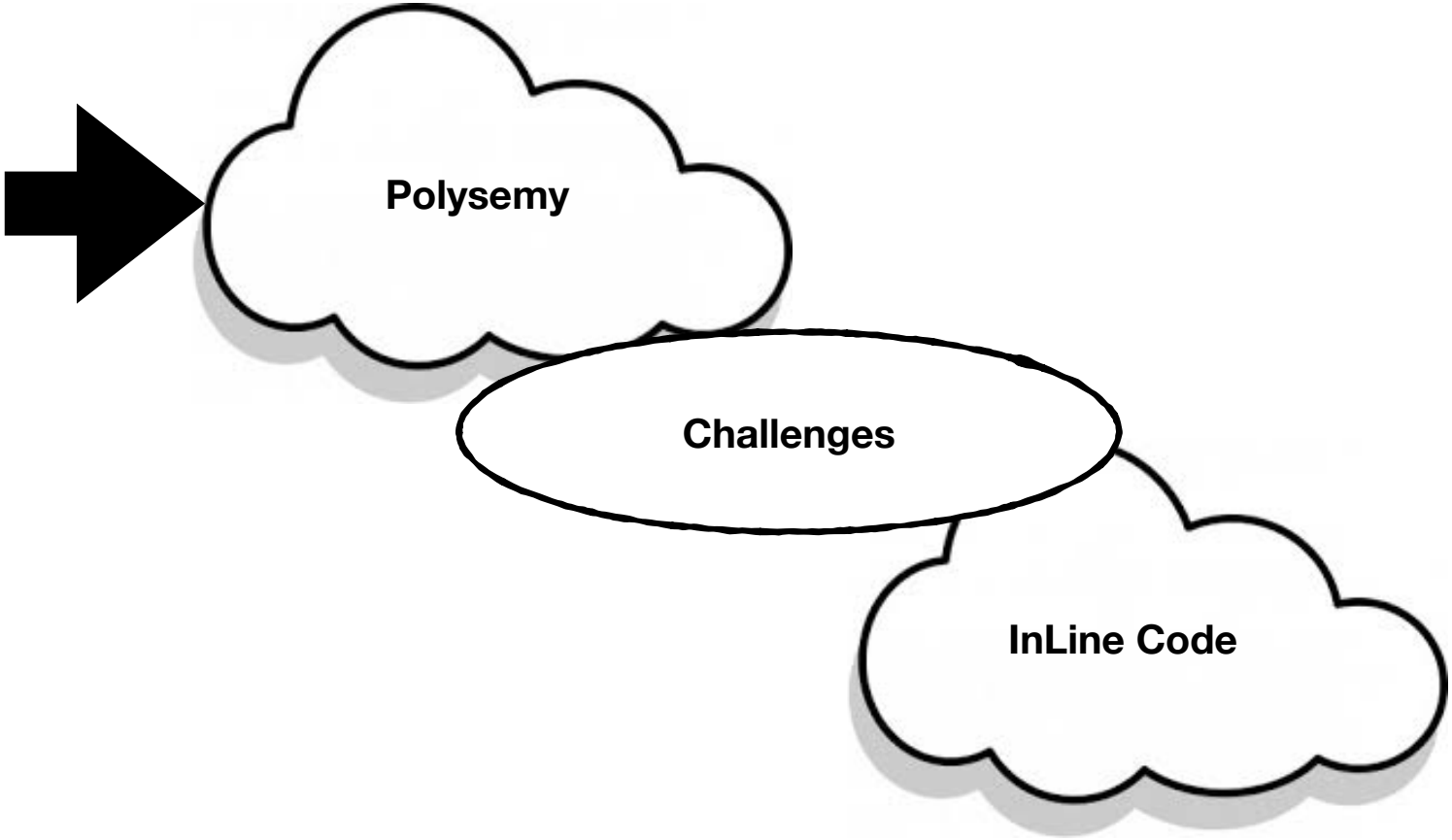


[Movshovitz-Attias and Cohen 2015]

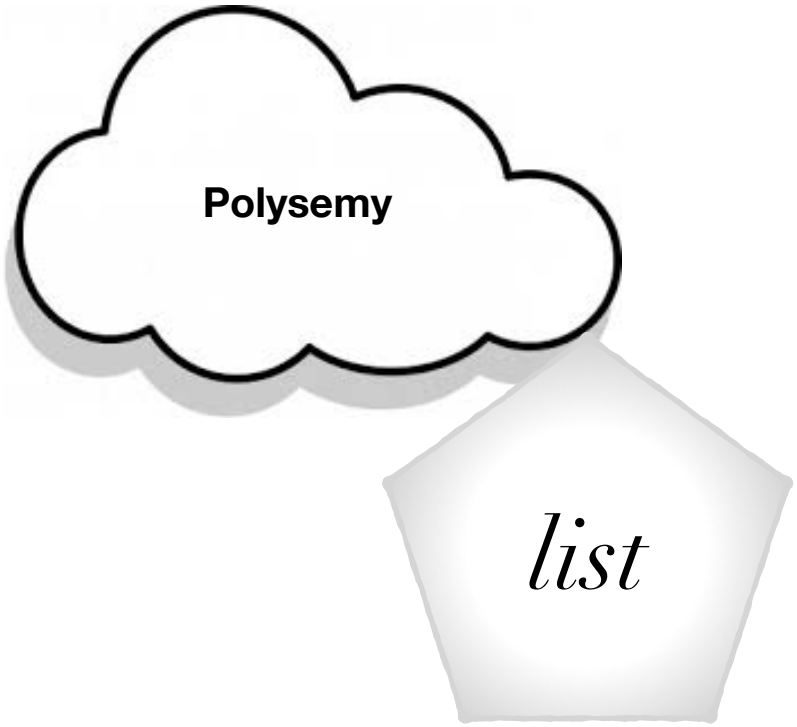
StackOverflow Entity Recognition



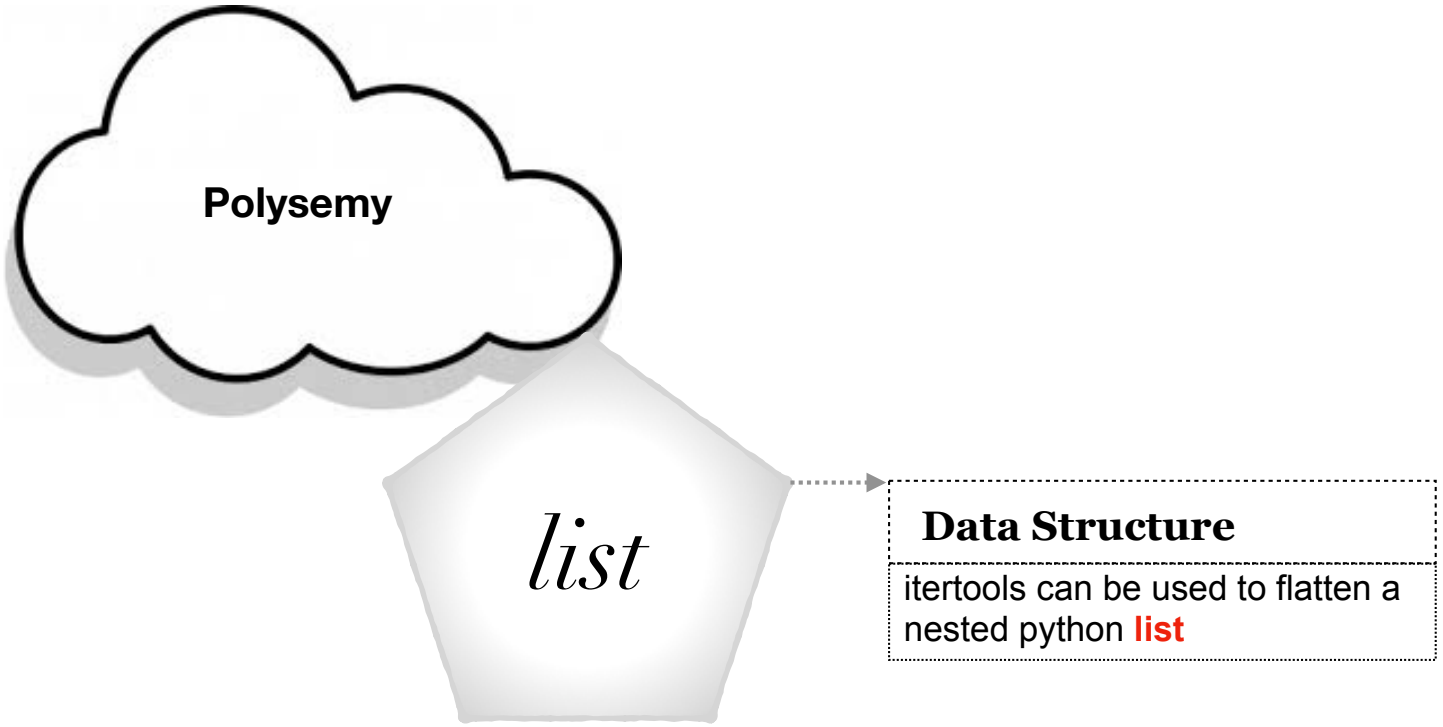
StackOverflow Entity Recognition



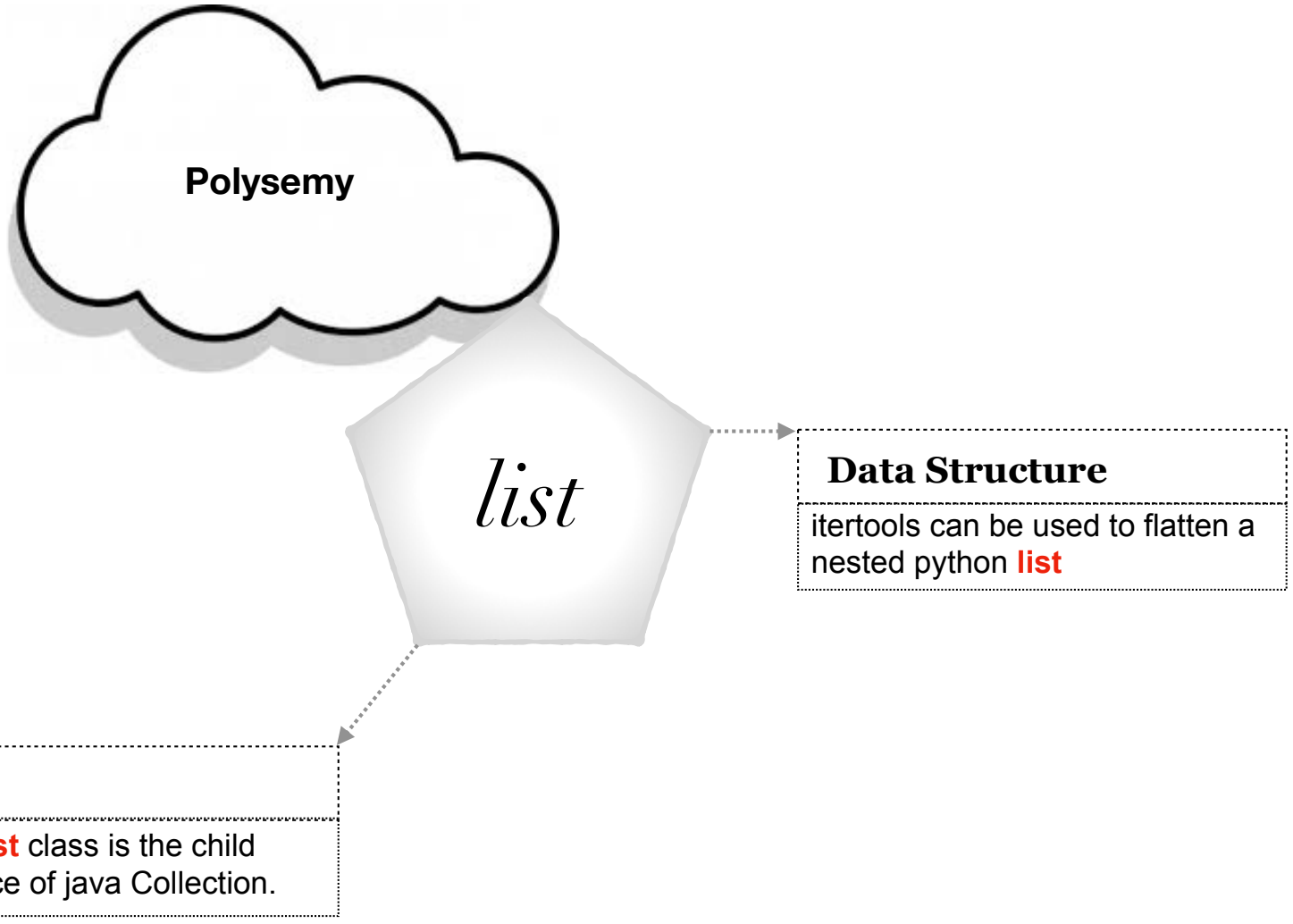
StackOverflow Entity Recognition



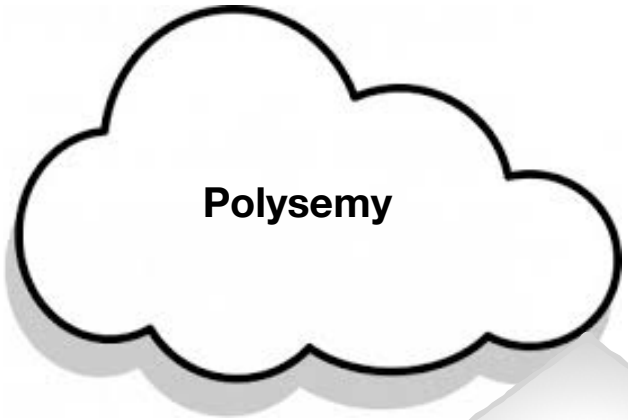
StackOverflow Entity Recognition



StackOverflow Entity Recognition



StackOverflow Entity Recognition



list

Data Structure

itertools can be used to flatten a nested python **list**

Class

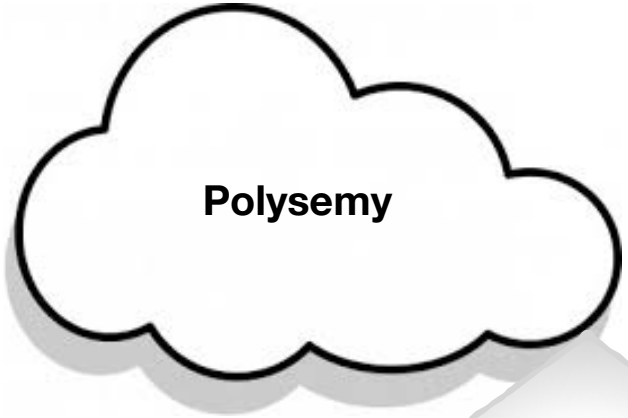
The **List** class is the child interface of java Collection.

Variable

I want to iterate through the following **list**

```
ArrayList<String> list=new ArrayList<String>();
                list.add("http://www.google.com");
                list.add("http://www.stackoverflow.com");
                  exchange.getOut().setHeader("endpoints",list);
```

StackOverflow Entity Recognition



list

Non Entity

Here is a **list** of most useful ajax frameworks

Data Structure

itertools can be used to flatten a nested python **list**

Class

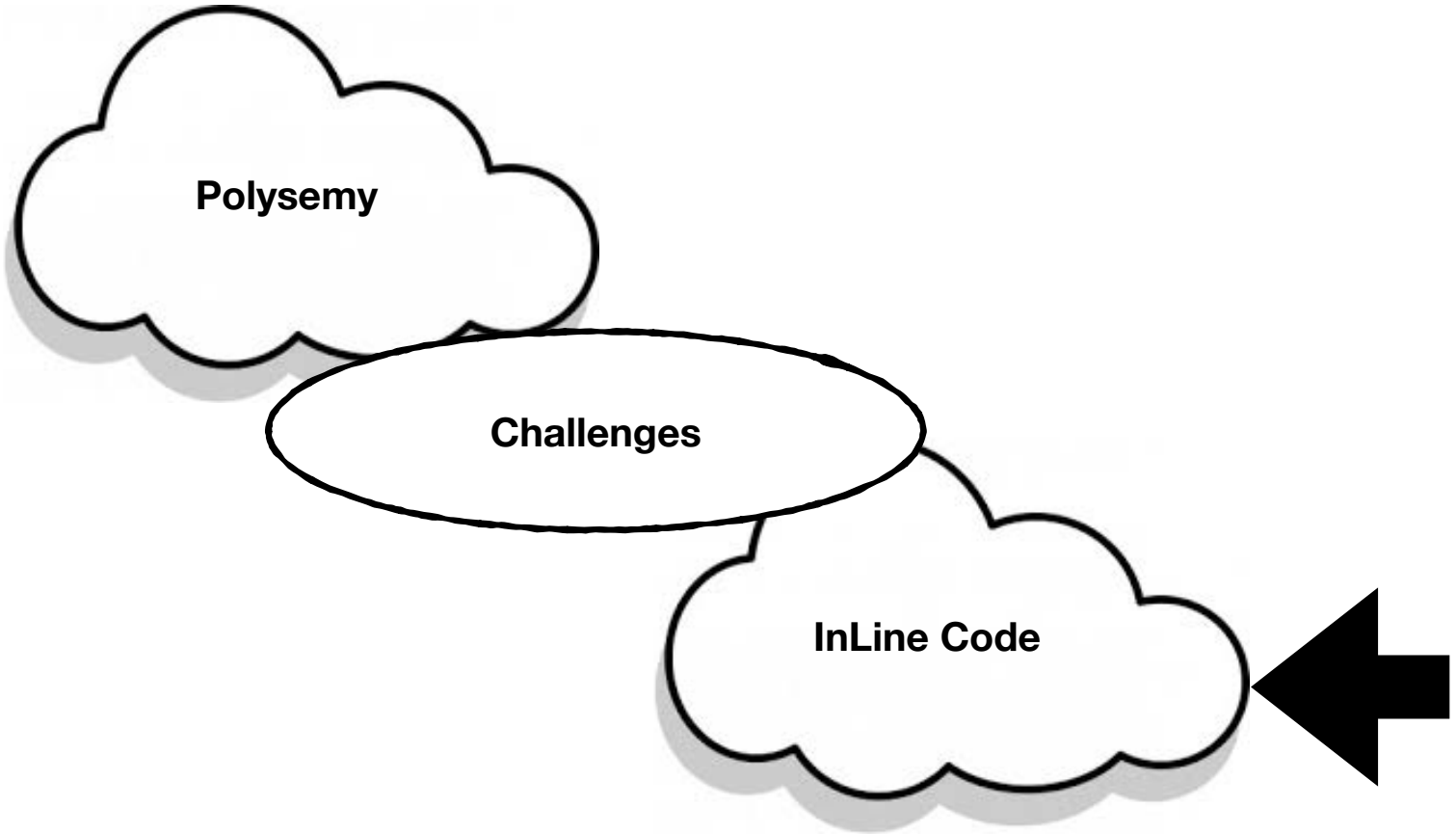
The **List** class is the child interface of java Collection.

Variable

I want to iterate through the following **list**

```
ArrayList<String> list=new ArrayList<String>();
                list.add("http://www.google.com");
                list.add("http://www.stackoverflow.com");
                  exchange.getOut().setHeader("endpoints",list);
```

StackOverflow Entity Recognition



StackOverflow Entity Recognition

check if key is numeric by `is_numeric($key)` function

StackOverflow Entity Recognition

check if key is numeric by is_numeric($key) function

StackOverflow Entity Recognition

980 Unique Function Names in **15k sentences**

check if key is numeric by is_numeric($key) function

StackOverflow Entity Recognition

980 Unique Function Names in **15k sentences**

Before adding element to array, check if key is numeric by is_numeric($key) function.
If it return false, then, covert key to integer using typecasting, (int)$key.

Now, the array will have numeric keys only and can be ordered.

share improve this answer follow

answered Oct 23 '15 at 9:16

300 ● 1 ● 5

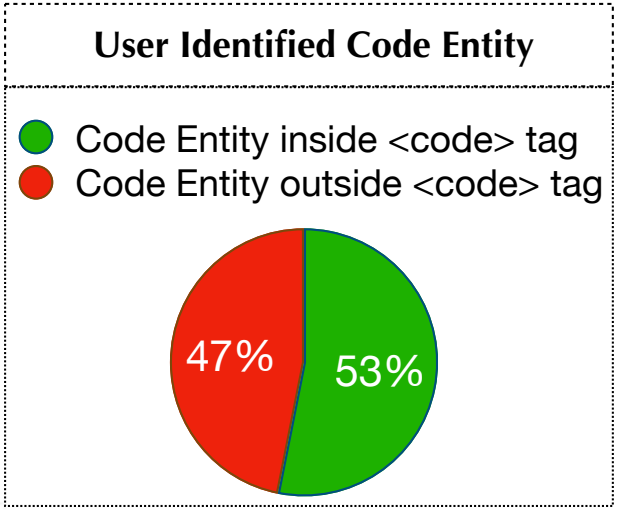
InLine Code

check if key is numeric by is_numeric($key) function

StackOverflow Entity Recognition

980 Unique Function Names in **15k sentences**

User Identified Code Entity



InLine Code

Before adding element to array, check if key is numeric by is_numeric($key) function.
If it return false, then, covert key to integer using typecasting, (int)$key.

Now, the array will have numeric keys only and can be ordered.

share improve this answer follow

answered Oct 23 '15 at 9:16

300 ● 1 ● 5

check if key is numeric by is_numeric($key) function

Contribution #1: Construct **new data** for Software Entities

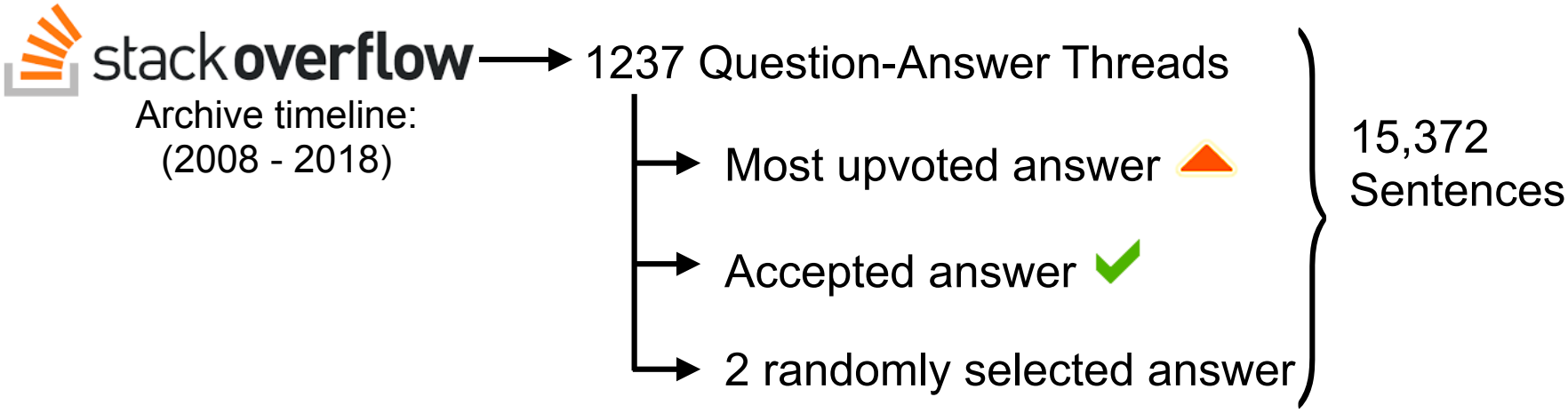
Contribution #2: Propose a **new model** for Software Entities

Contribution #1: Construct **new data** for Software Entities

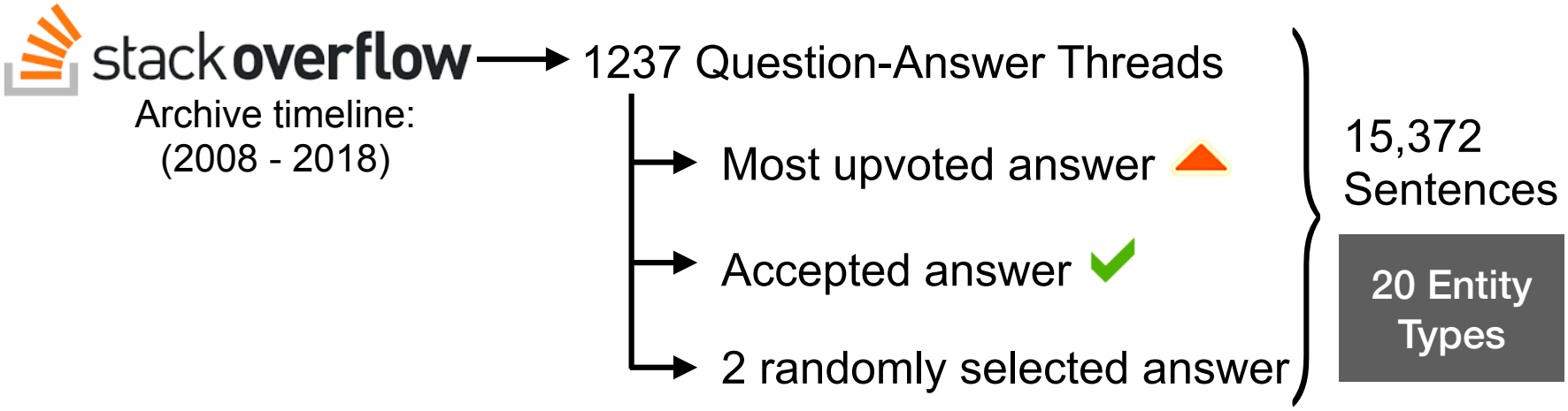
- Manually annotated sentence with **20 types of entities**
- **15k** StackOverflow sentence

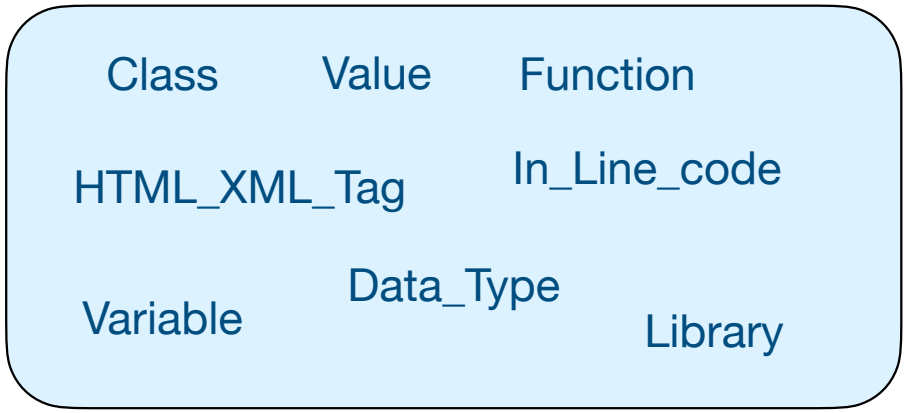
Contribution #2: Propose a **new model** for Software Entities

Annotated StackOverflow Corpus

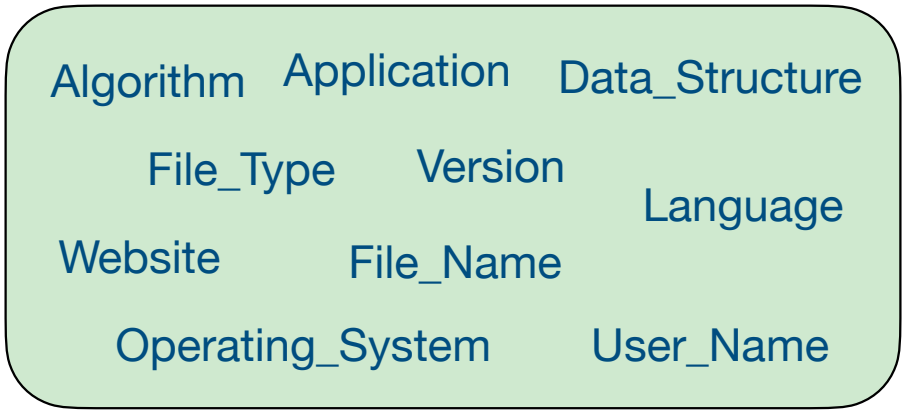


Annotated StackOverflow Corpus



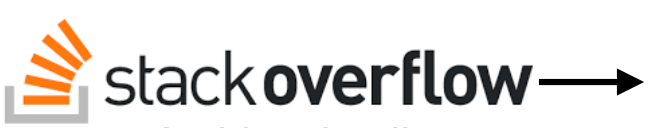


Code Entity Types



Natural Language Entity

Annotated StackOverflow Corpus



Archive timeline:
(2008 - 2018)

→ 1237 Question-Answer Threads

- Most upvoted answer
- Accepted answer
- 2 randomly selected answer

15,372 Sentences

20 Entity Types

Code Entity Types

Class, Value, Function, HTML_XML_Tag, In_Line_code, Variable, Data_Type, Library

Natural Language Entity

Algorithm, Application, Data_Structure, File_Type, Version, Language, Website, File_Name, Operating_System, User_Name

Annotated StackOverflow Corpus

stackoverflow → 1237 Question-Answer Threads

Archive timeline:
(2008 - 2018)

- Most upvoted answer
- Accepted answer
- 2 randomly selected answer

15,372 Sentences

20 Entity Types

1 Answer_to_Question_ID: 27926052
2 Question_URL: https://stackoverflow.com/questions/27926052/

4 You can now add customised selector(please refer 'IQUIView+IQKeyboardToolbar.h') for previous/next/done to get notify.
(Library_Class: 'IQUIView+IQKeyboardToolbar.h'; File_Name: 'IQUIView+IQKeyboardToolbar.h'; Value: previous/next/done)

5 Note that custom selector doesn't affect the native functionality of previous/next/done, it's just used for callback purpose only.
(Value: previous/next/done)

6 For detail documentation please refer 'IQUIView+IQKeyboardToolbar.h', for 'how to use?'
(Library_Class: 'IQUIView+IQKeyboardToolbar.h'; File_Name: 'IQUIView+IQKeyboardToolbar.h')

7 please refer 'TextFieldViewController.m'.
(File_Name: 'TextFieldViewController.m')

[annotators]

[**40%** double annotated]
[**62%** agreemnet]

Annotated StackOverflow Corpus

stackoverflow → 1237 Question-Answer Threads

Archive timeline:
(2008 - 2018)

- Most upvoted answer
- Accepted answer
- 2 randomly selected answer

15,372 Sentences

20 Entity Types


```
1 Answer_to_Question_ID: 27926052
2 Question_URL: https://stackoverflow.com/questions/27926052/

4 You can now add customised selector(please refer 'IQUIView+IQKeyboardToolbar.h') for
  previous/next/done to get notify.

5 Note that custom selector doesn't affect the native functionality of previous/next/done, it's just
  used for callback purpose only.

6 For detail documentation please refer 'IQUIView+IQKeyboardToolbar.h', for 'how to use?'

7 please refer 'TextFieldViewController.m'.
```

(Library_Class, File_Name, Value, Value, Library_Class, File_Name, File_Name)

[annotators]

[40% double annotated]
[62% agreemnet]

[adjudicator]

- Resolves disagreement in double annotated data
- Ensures the sanity of single annotated data

Contribution #1: Construct **new data** for Software Entities

- Manually annotated sentence with **20 types of entities**
  - **15k** StackOverflow sentence

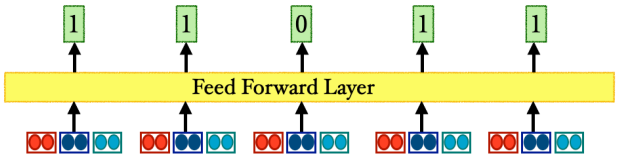
Contribution #2: Proposed a **new model** for Software Entities

- Attentive NER tagger
  - combined with in-domain contextual representation

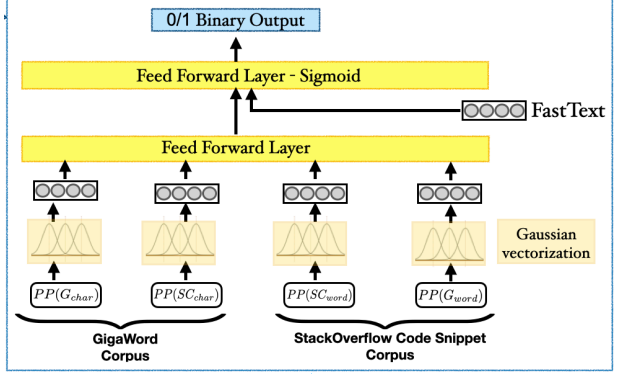
SoftNER

Contextual Word Representation

Entity Segmenter



Code Recognizer



Polysemy

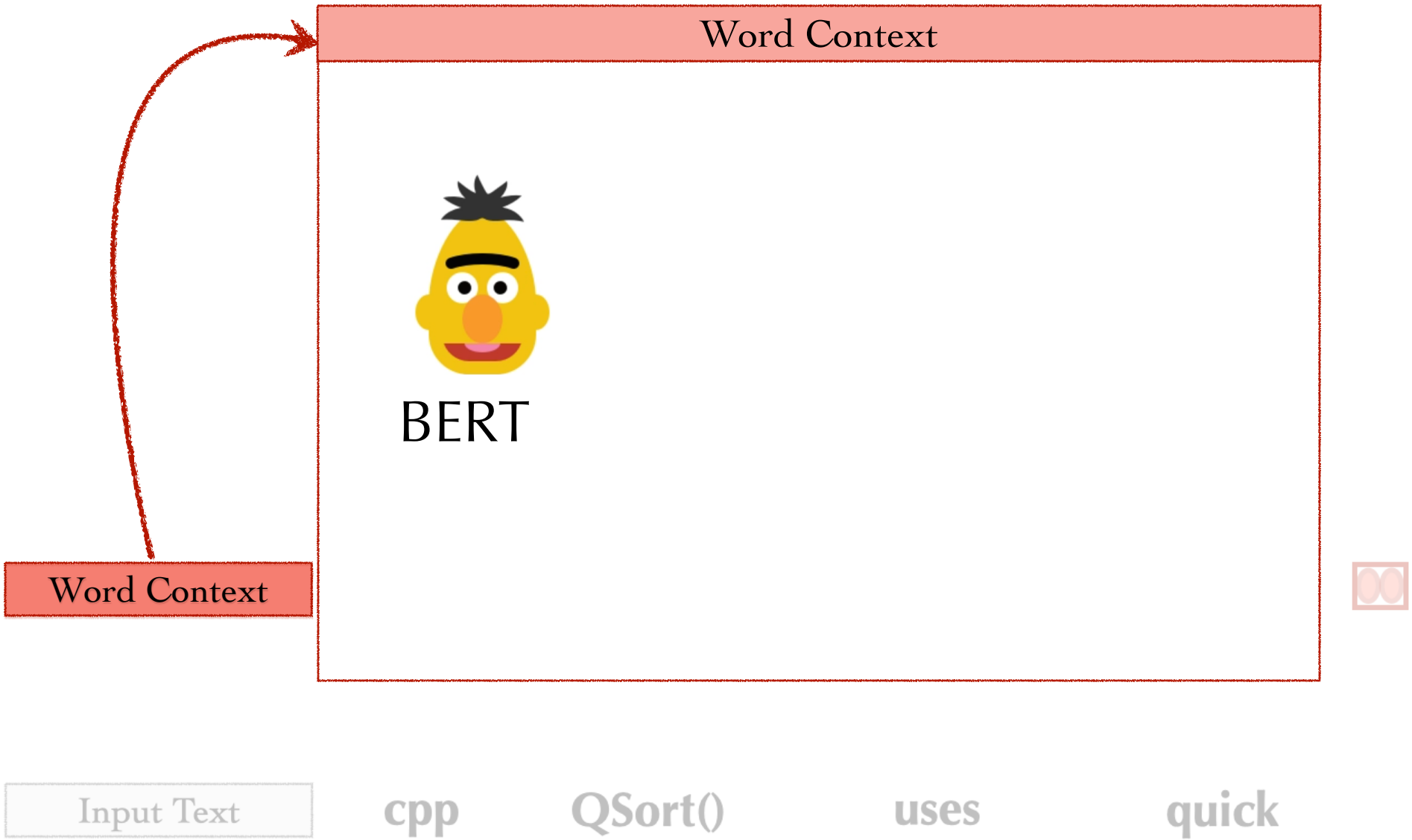
Challenges

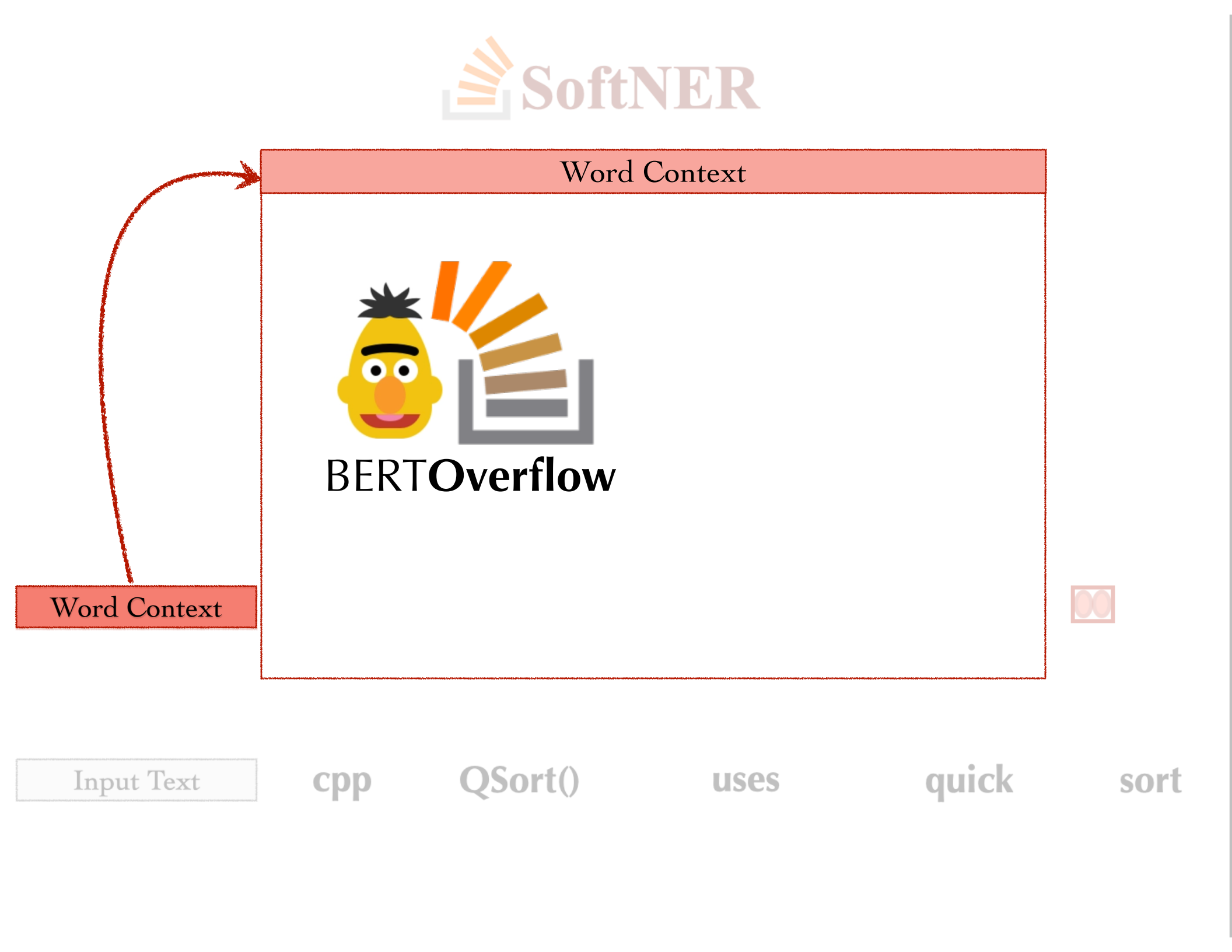
InLine Code

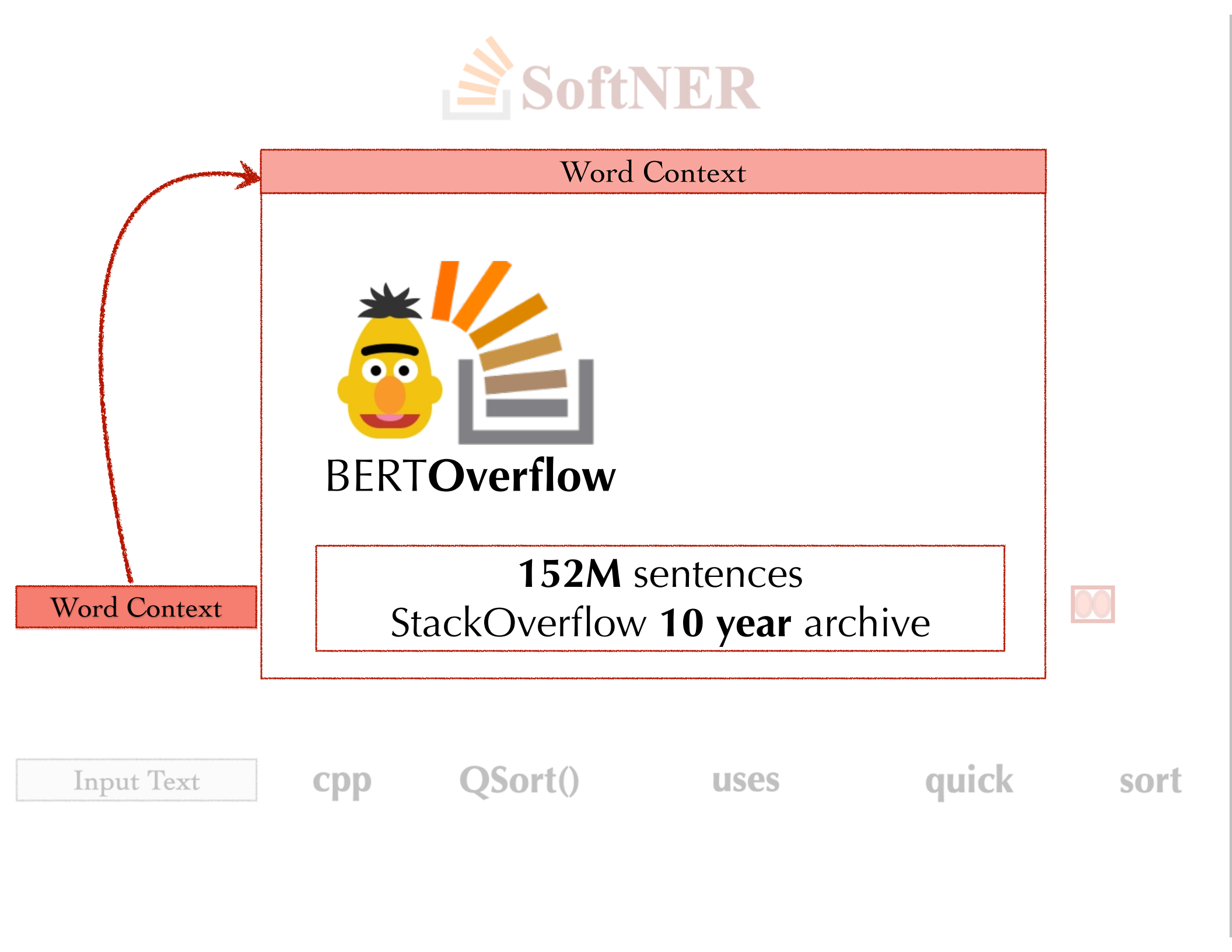
Input Text

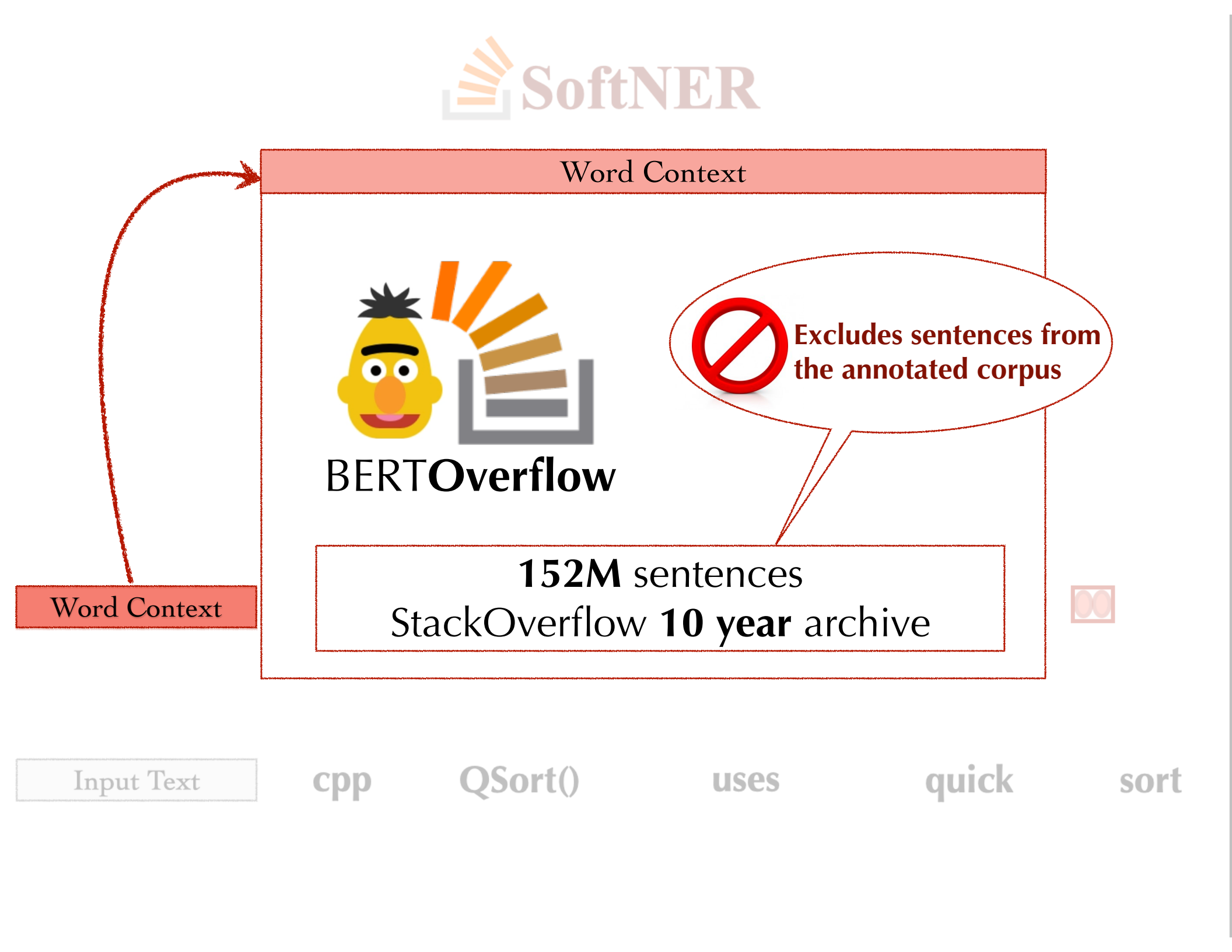
cpp **QSort()** **uses** **quick** **sort**

SoftNER

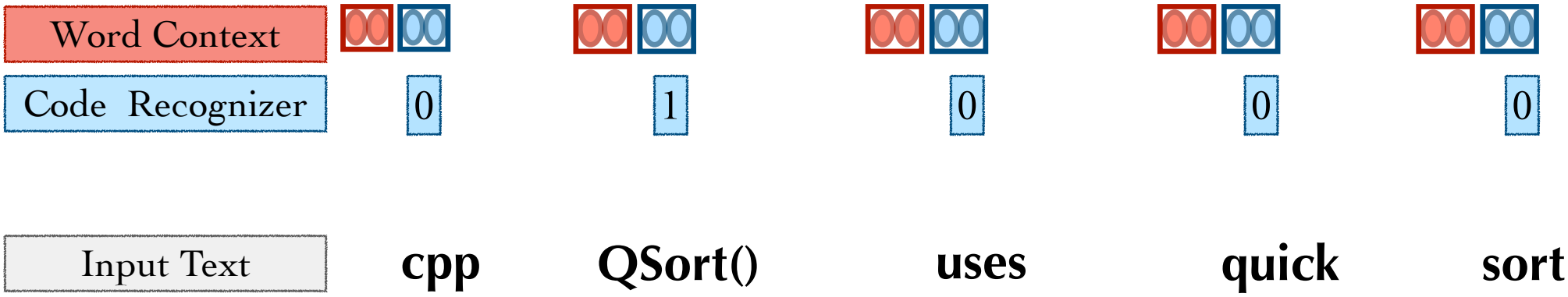

SoftNER
Word Context
BERT
Word Context
Input Text
cpp
QSort()
uses
quick
sort

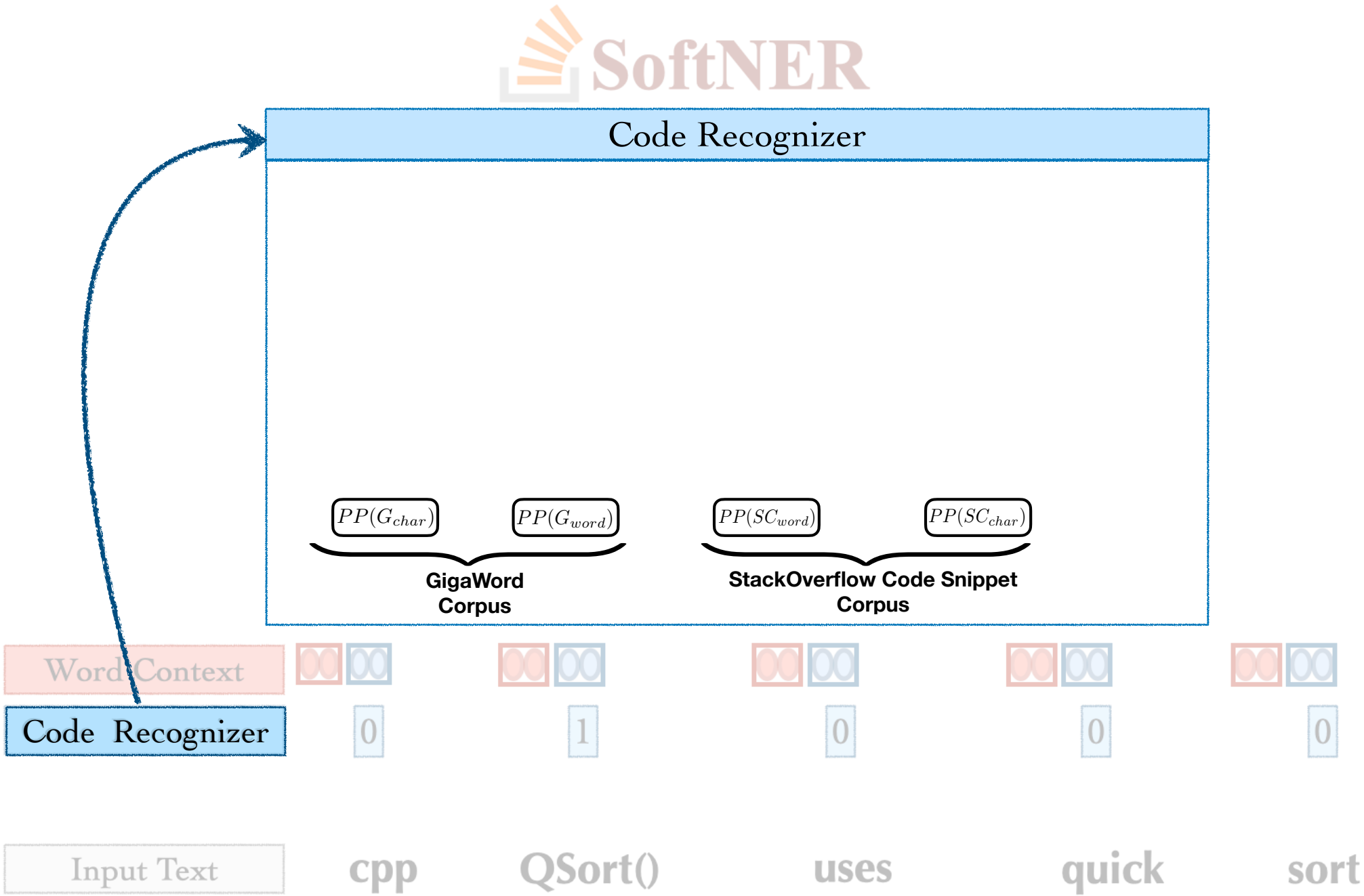

SoftNER
Word Context
BERTOverflow
Word Context
Input Text
cpp
QSort()
uses
quick
sort

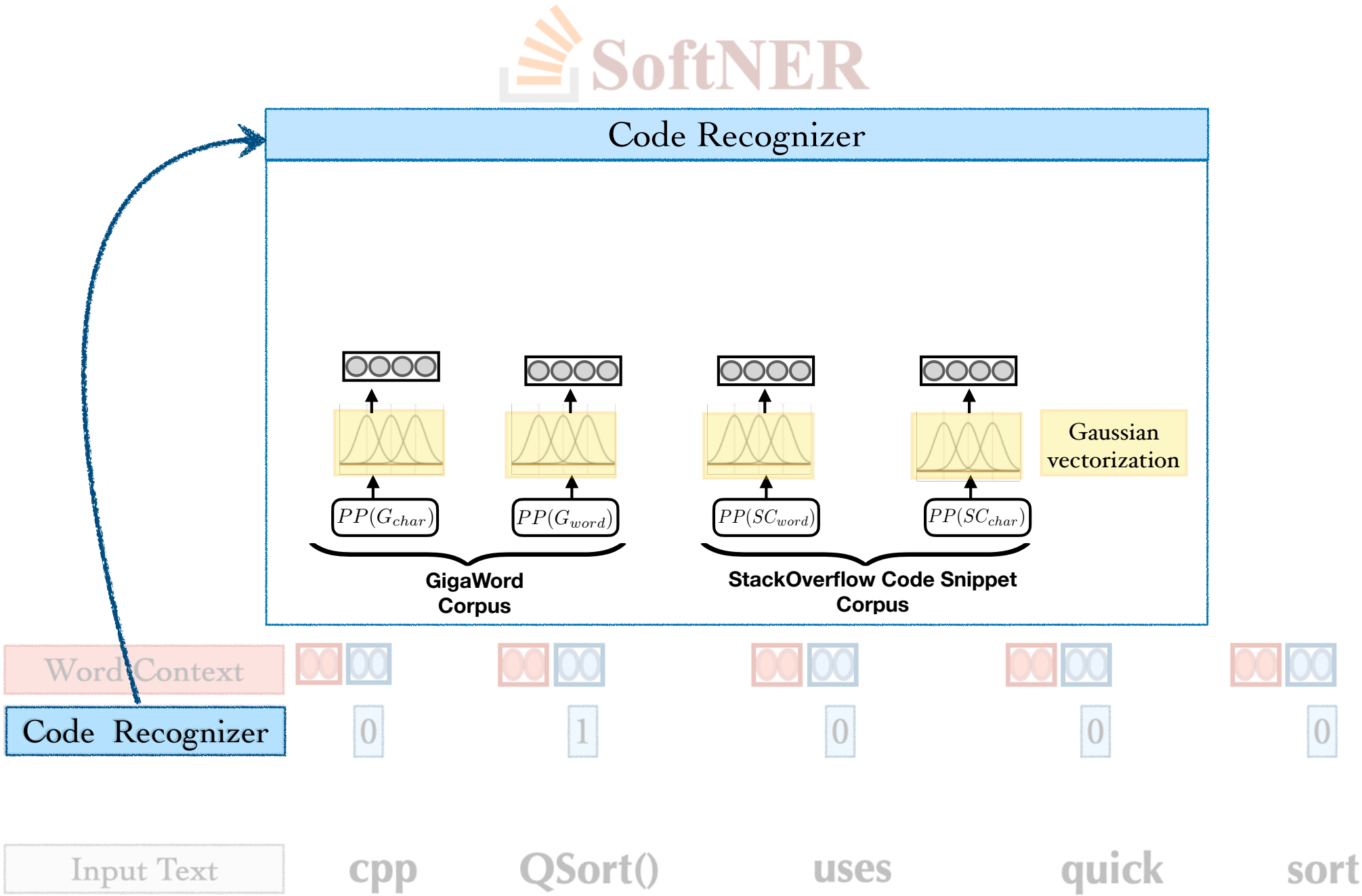

SoftNER
Word Context
BERTOverflow
152M sentences
StackOverflow 10 year archive
Word Context
Input Text
cpp
QSort()
uses
quick
sort

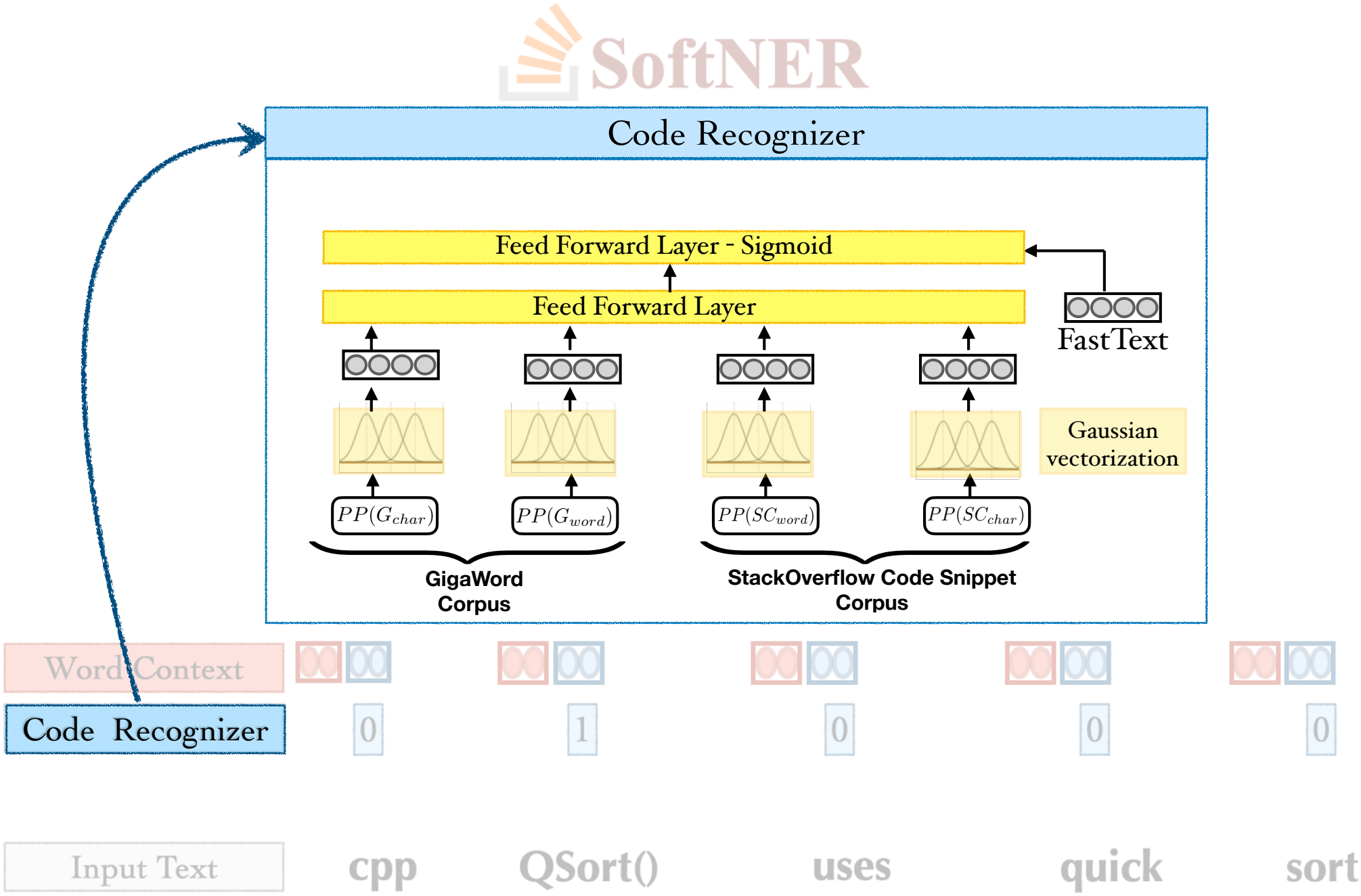

SoftNER
Word Context
BERTOverflow
Excludes sentences from the annotated corpus
152M sentences
StackOverflow 10 year archive
Word Context
Input Text
cpp
QSort()
uses
quick
sort

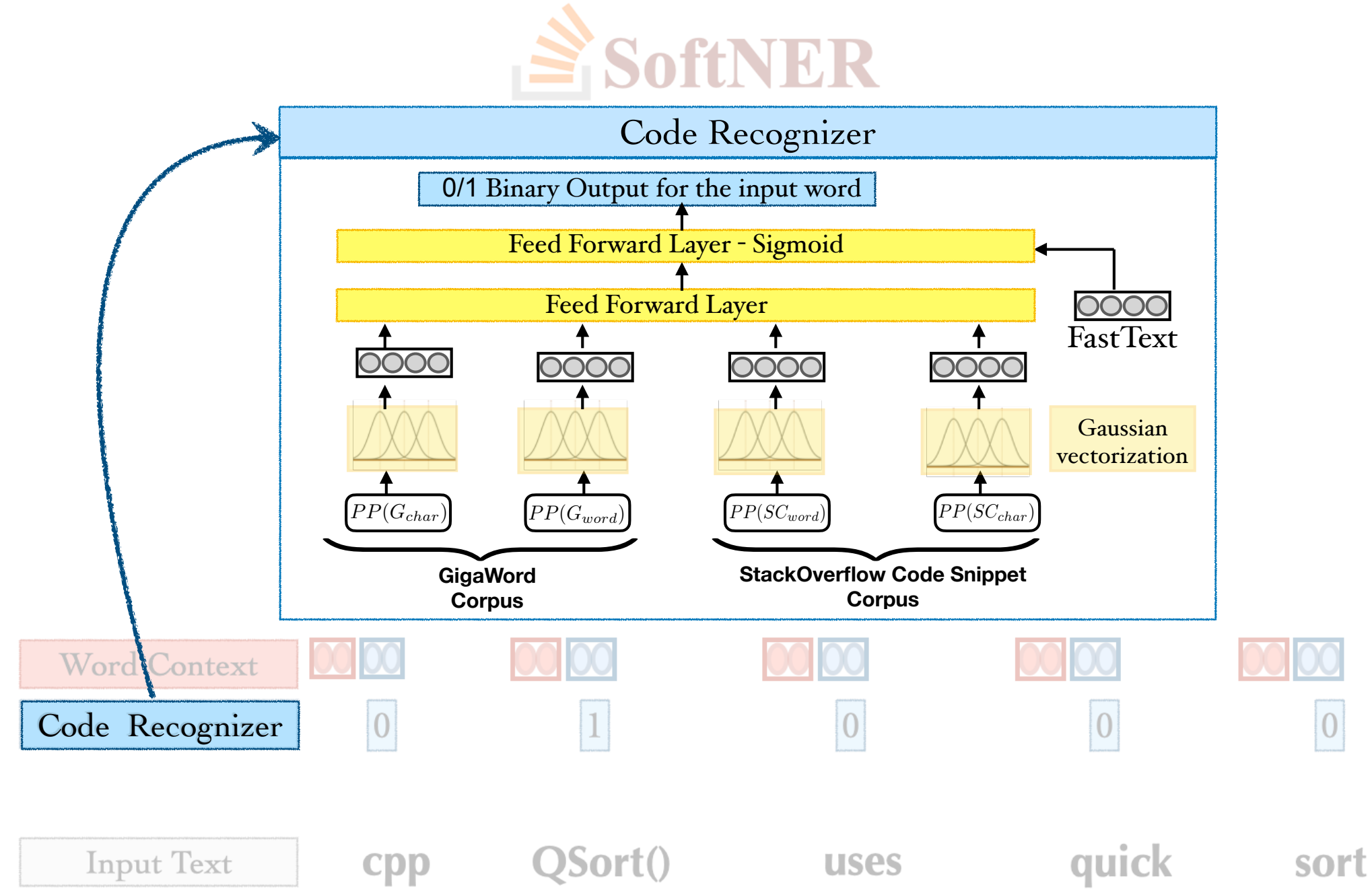
SoftNER

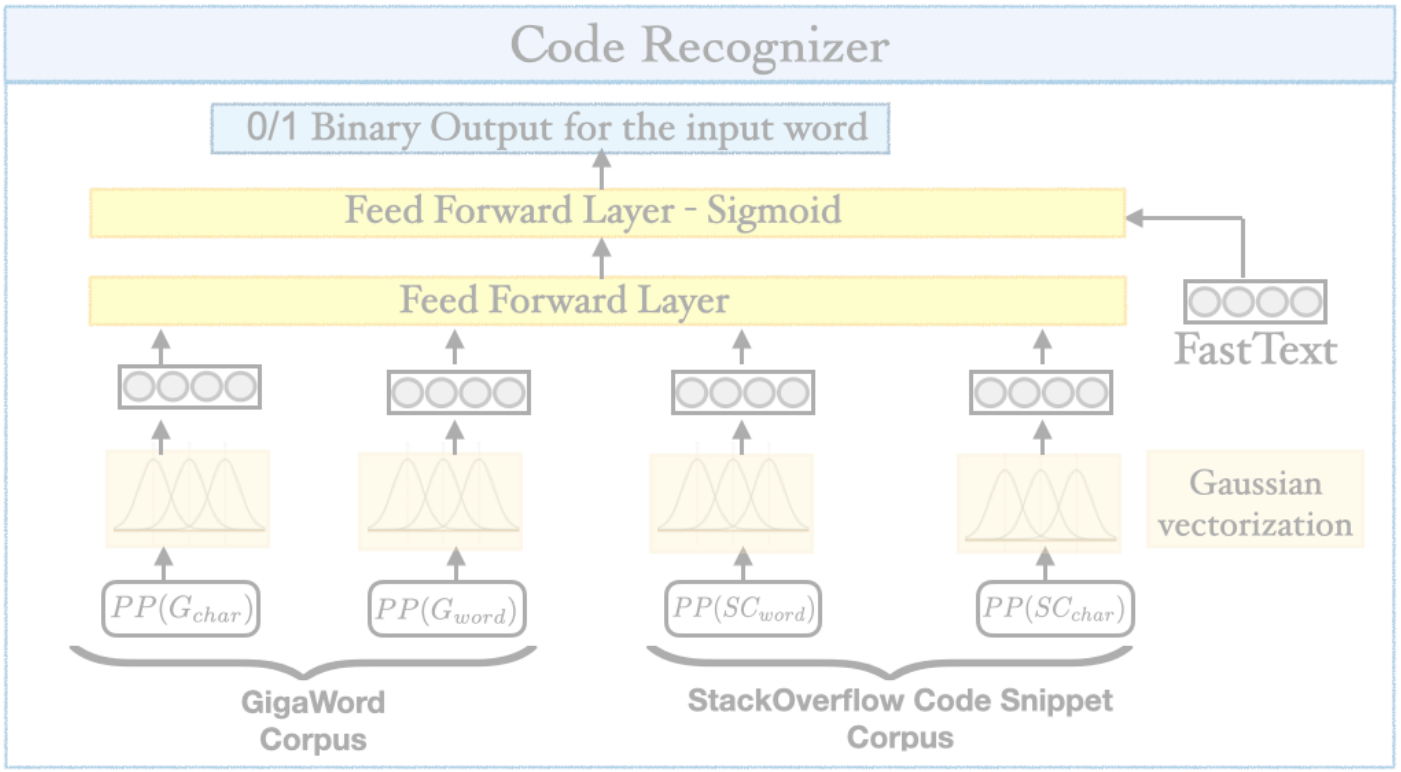

Word Context
Code Recognizer
0
1
0
0
0
Input Text
cpp
QSort()
uses
quick
sort


SoftNER
Code Recognizer
$PP(G_{char})$
$PP(G_{word})$
$PP(SC_{word})$
$PP(SC_{char})$
GigaWord Corpus
StackOverflow Code Snippet Corpus
Word Context
Code Recognizer
0
1
0
0
0
Input Text
cpp
QSort()
uses
quick
sort


SoftNER
Code Recognizer
$PP(G_{char})$
$PP(G_{word})$
$PP(SC_{word})$
$PP(SC_{char})$
Gaussian vectorization
GigaWord Corpus
StackOverflow Code Snippet Corpus
Word Context
Code Recognizer
0
1
0
0
0
Input Text
cpp
QSort()
uses
quick
sort

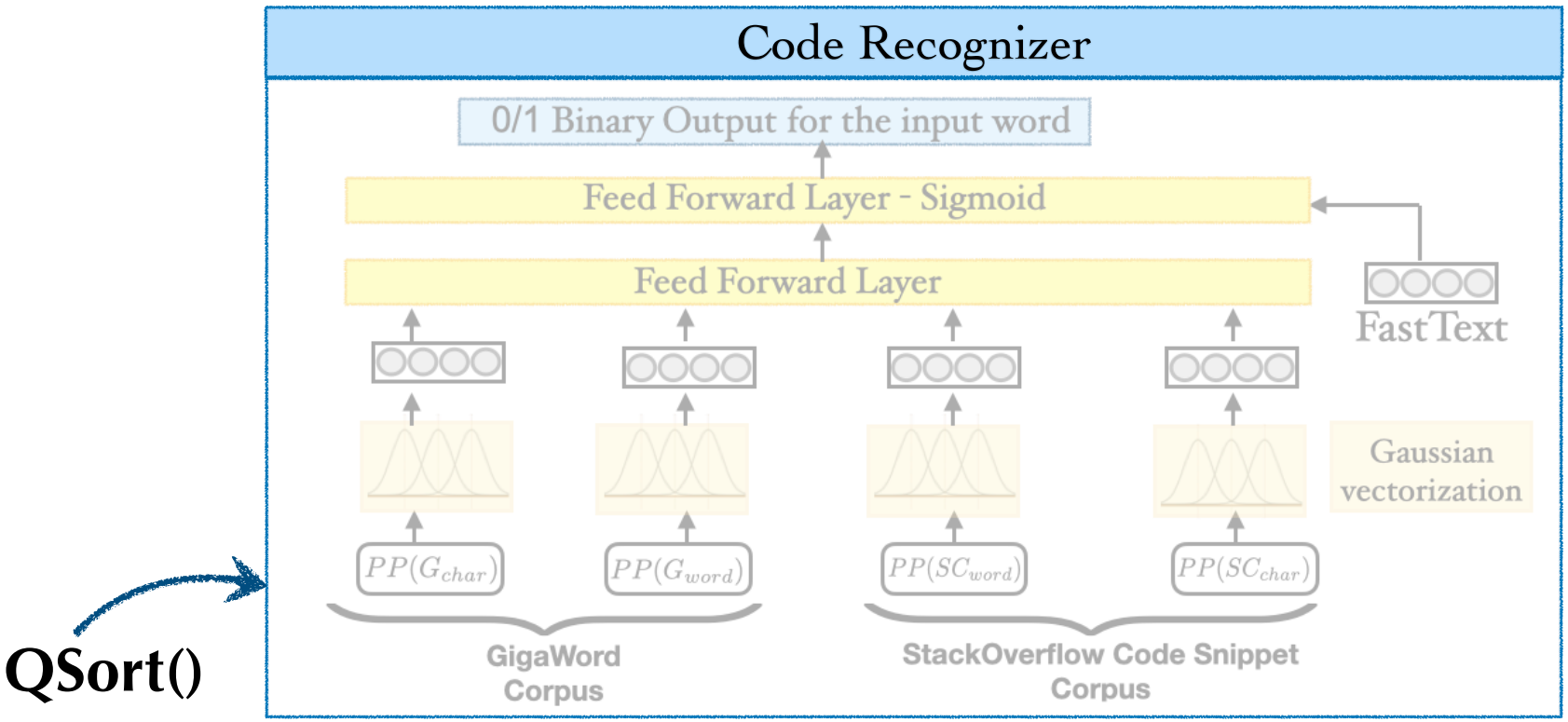

SoftNER
Code Recognizer
Feed Forward Layer - Sigmoid
Feed Forward Layer
FastText
Gaussian vectorization
$PP(G_{char})$
$PP(G_{word})$
$PP(SC_{word})$
$PP(SC_{char})$
GigaWord Corpus
StackOverflow Code Snippet Corpus
Word Context
Code Recognizer
0
1
0
0
0
Input Text
cpp
QSort()
uses
quick
sort

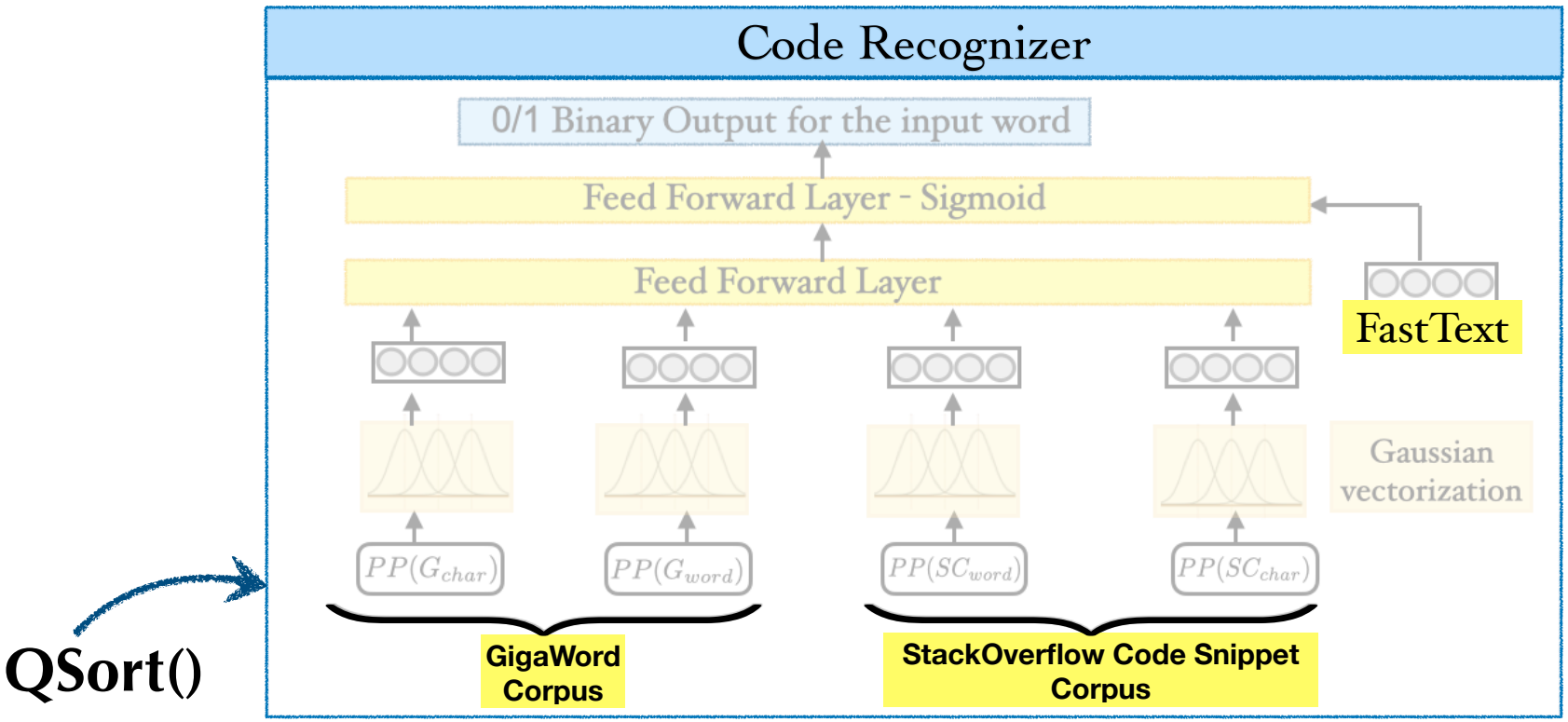

SoftNER
Code Recognizer
0/1 Binary Output for the input word
Feed Forward Layer - Sigmoid
Feed Forward Layer
FastText
Gaussian vectorization
$PP(G_{char})$
$PP(G_{word})$
$PP(SC_{word})$
$PP(SC_{char})$
GigaWord Corpus
StackOverflow Code Snippet Corpus
Word Context
Code Recognizer
0
1
0
0
0
Input Text
cpp
QSort()
uses
quick
sort

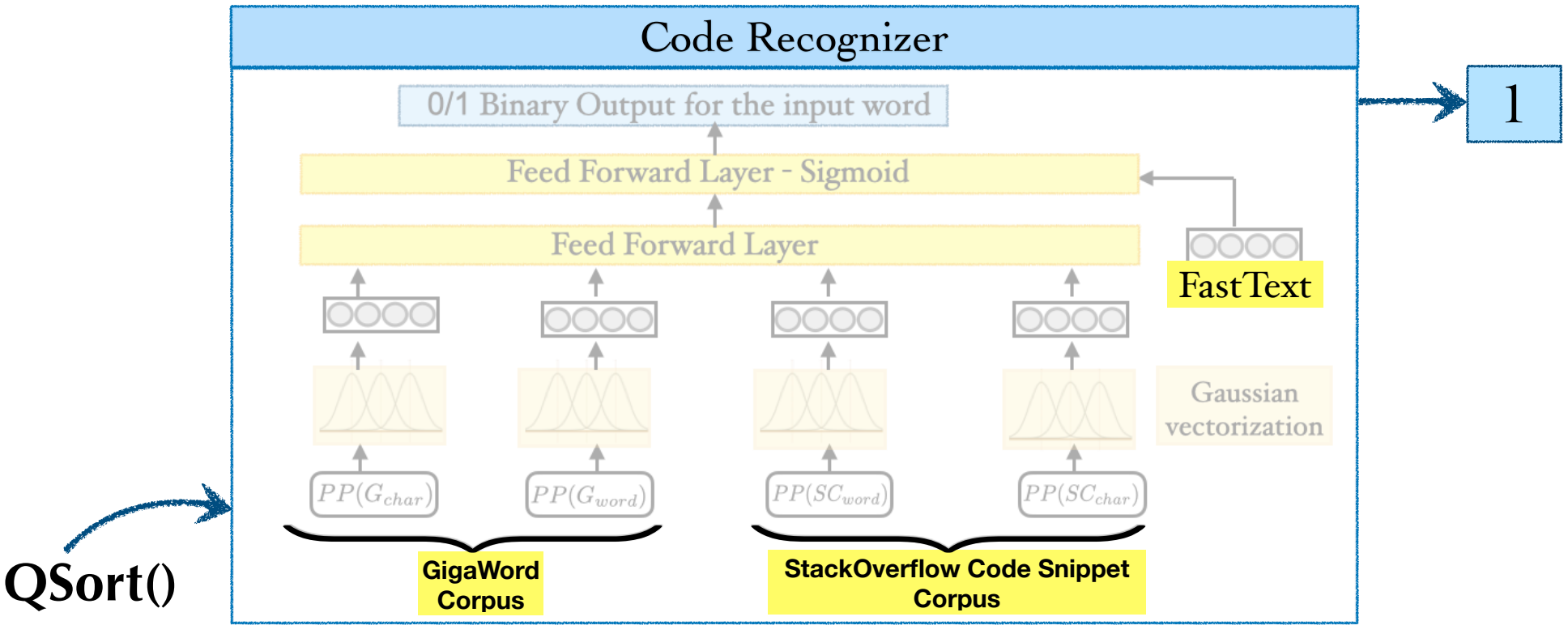


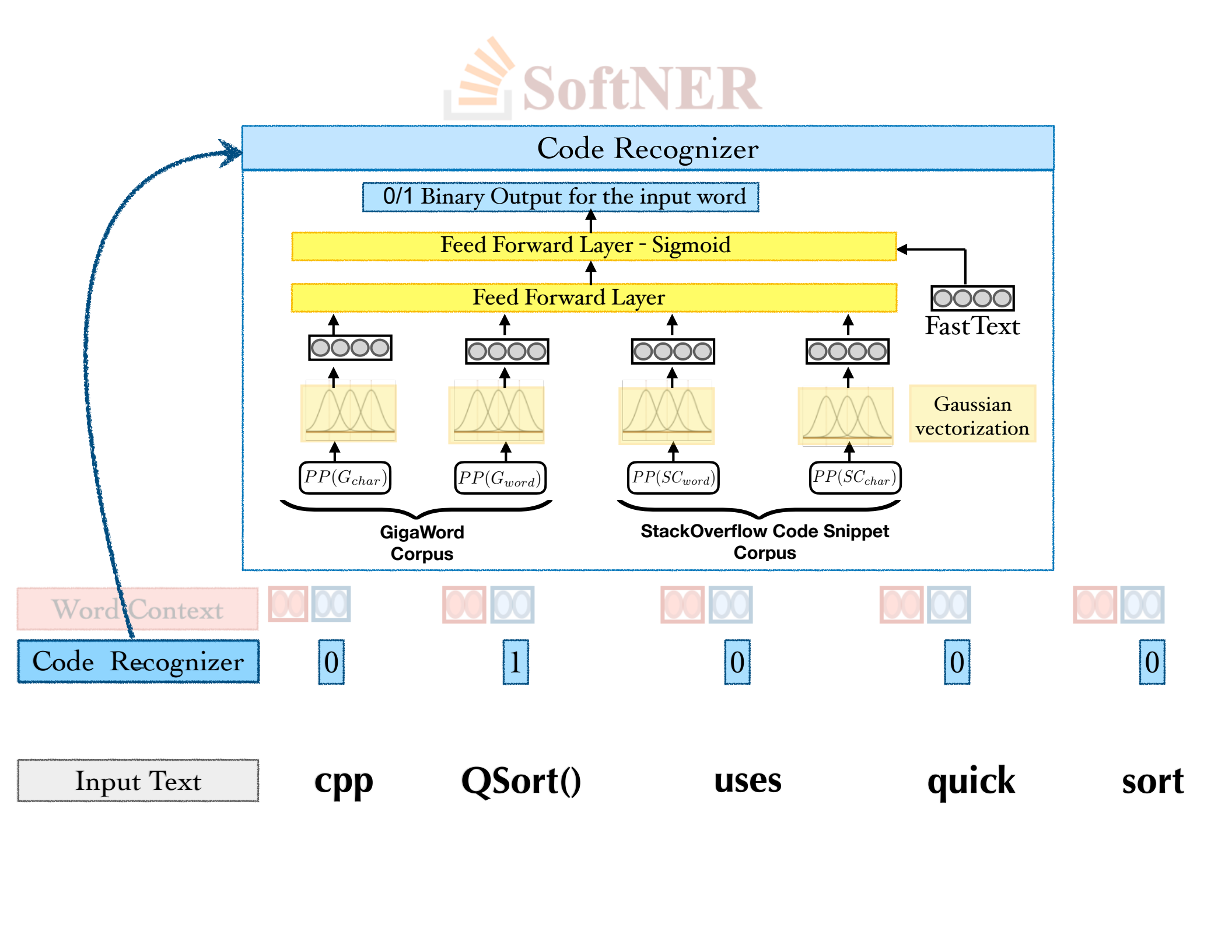
QSort()

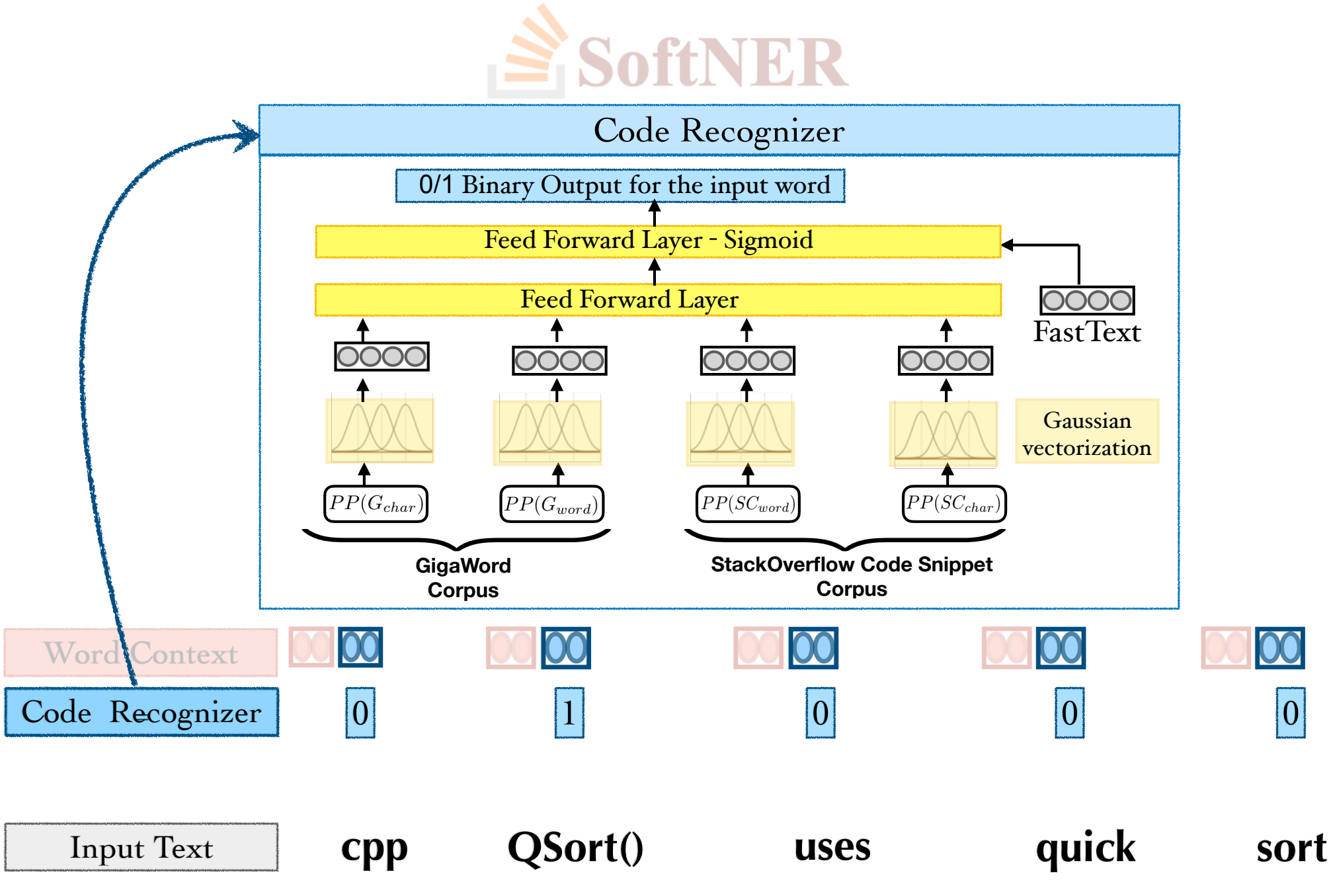
Word Context					
Code Recognizer	0	1	0	0	0
Input Text	cpp	QSort()	uses	quick	sort


Code Recognizer
0/1 Binary Output for the input word
Feed Forward Layer - Sigmoid
Feed Forward Layer
FastText
Gaussian vectorization
$PP(G_{char})$
$PP(G_{word})$
$PP(SC_{word})$
$PP(SC_{char})$
GigaWord Corpus
StackOverflow Code Snippet Corpus
QSort()

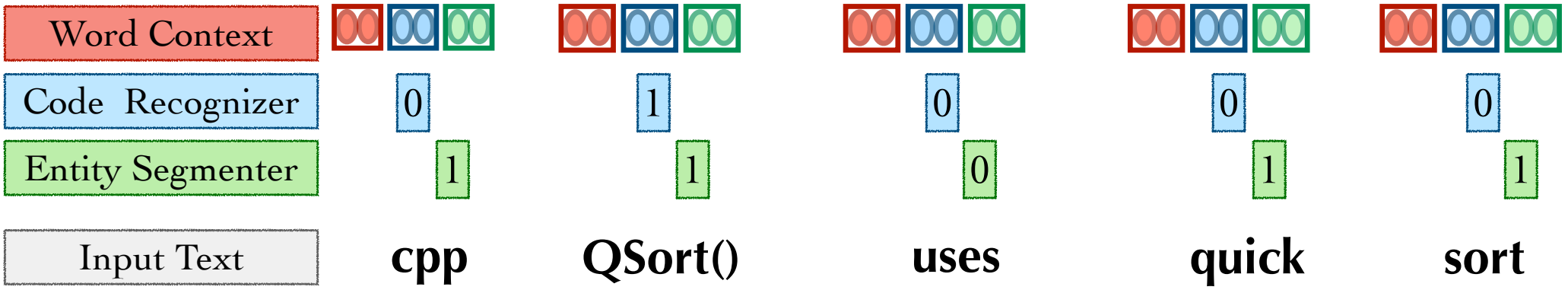

Code Recognizer
0/1 Binary Output for the input word
Feed Forward Layer - Sigmoid
Feed Forward Layer
FastText
Gaussian vectorization
$PP(G_{char})$
$PP(G_{word})$
$PP(SC_{word})$
$PP(SC_{char})$
GigaWord Corpus
StackOverflow Code Snippet Corpus
QSort()

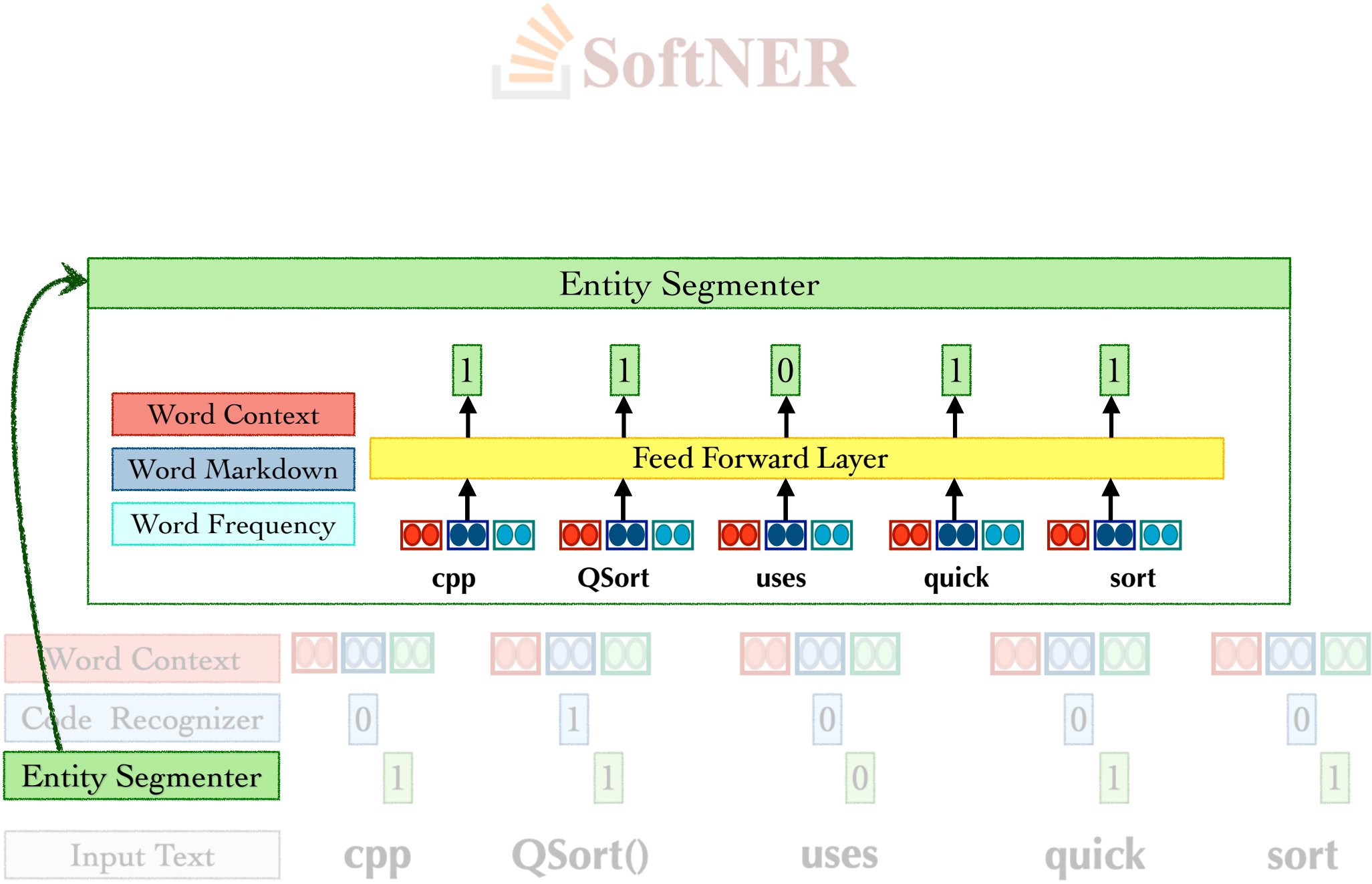

Code Recognizer
0/1 Binary Output for the input word
Feed Forward Layer - Sigmoid
Feed Forward Layer
FastText
Gaussian vectorization
$PP(G_{char})$
$PP(G_{word})$
$PP(SC_{word})$
$PP(SC_{char})$
GigaWord Corpus
StackOverflow Code Snippet Corpus
QSort()
1

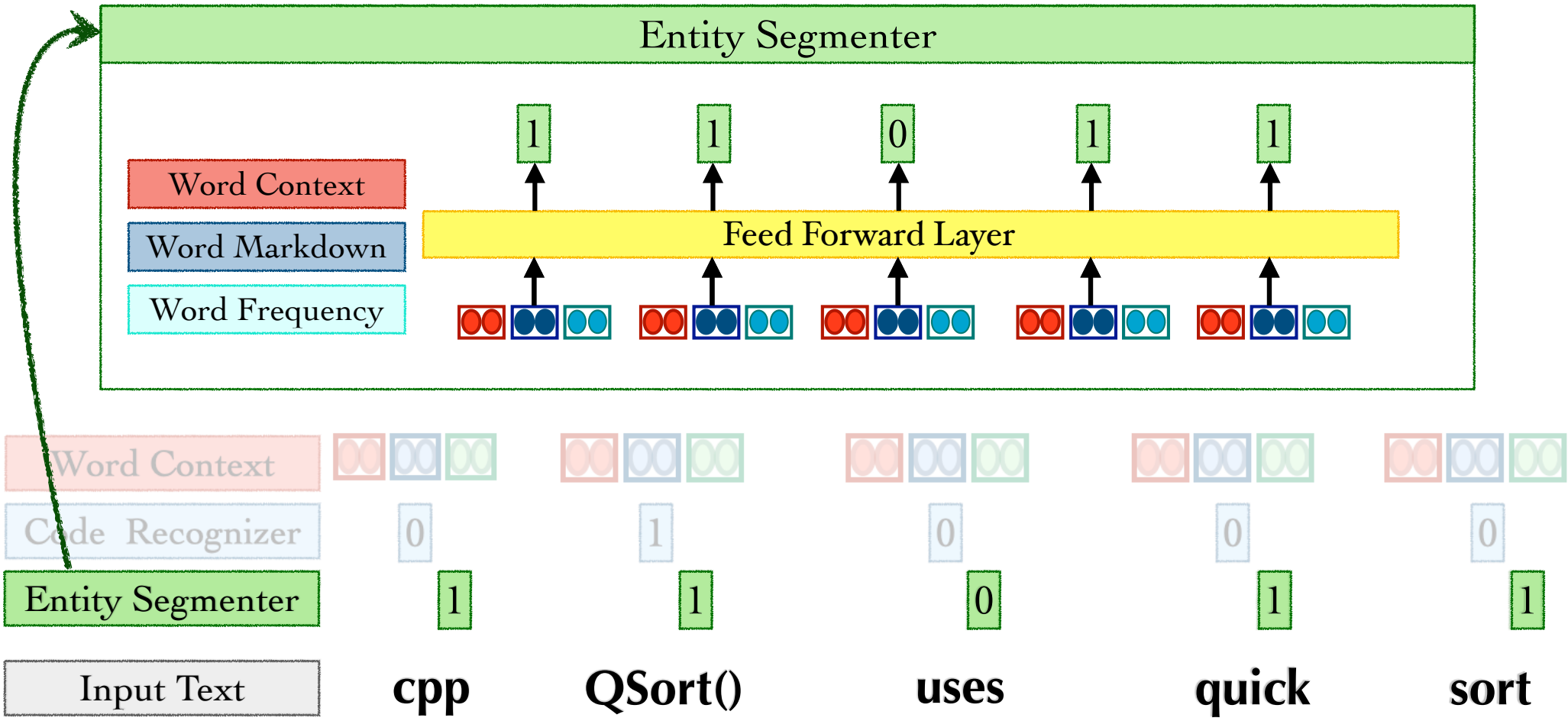

SoftNER
Code Recognizer
0/1 Binary Output for the input word
Feed Forward Layer - Sigmoid
Feed Forward Layer
FastText
Gaussian vectorization
$PP(G_{char})$
$PP(G_{word})$
$PP(SC_{word})$
$PP(SC_{char})$
GigaWord Corpus
StackOverflow Code Snippet Corpus
Word Context
Code Recognizer
0
1
0
0
0
Input Text
cpp
QSort()
uses
quick
sort

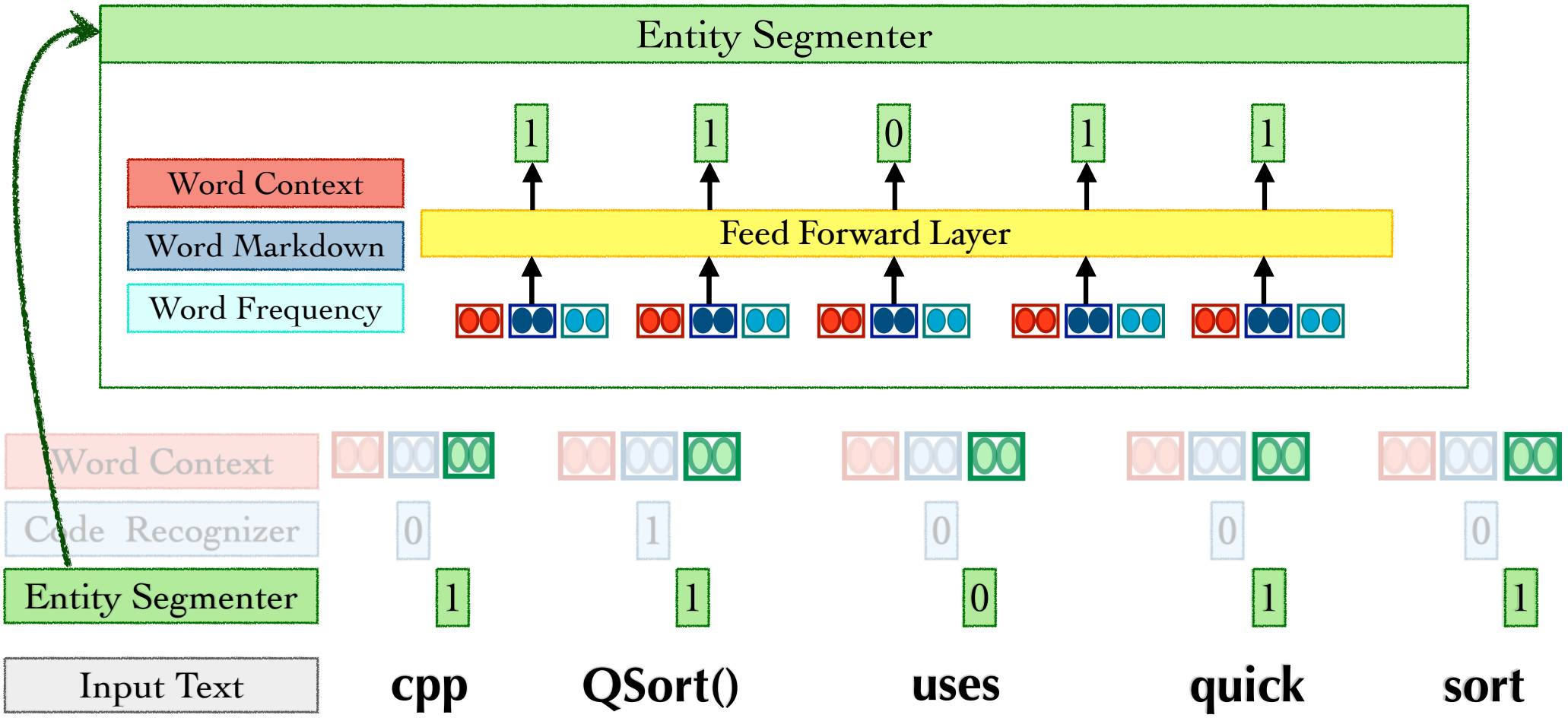

SoftNER
Code Recognizer
0/1 Binary Output for the input word
Feed Forward Layer - Sigmoid
Feed Forward Layer
FastText
Gaussian vectorization
$PP(G_{char})$
$PP(G_{word})$
$PP(SC_{word})$
$PP(SC_{char})$
GigaWord Corpus
StackOverflow Code Snippet Corpus
Word Context
Code Recognizer
0
1
0
0
0
Input Text
cpp
QSort()
uses
quick
sort

SoftNER

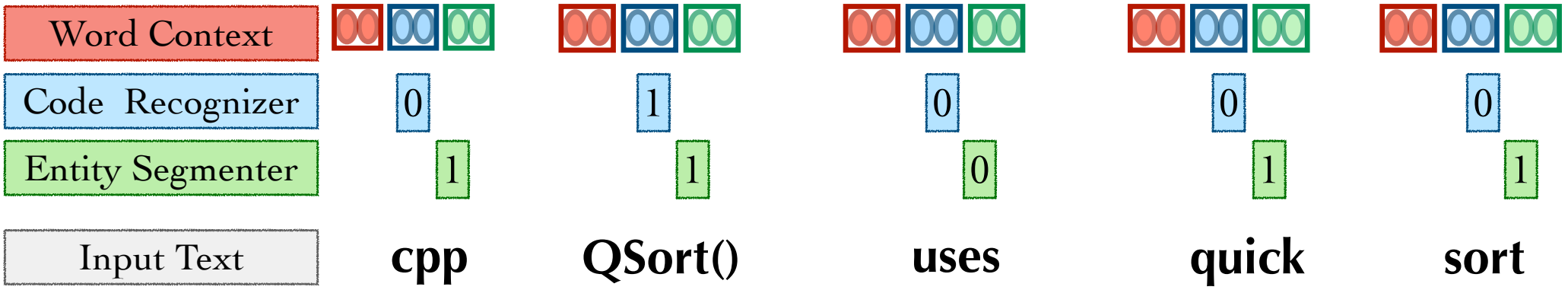

Word Context
Code Recognizer
Entity Segmenter
Input Text
0
1
cpp
1
1
QSort()
0
0
uses
0
1
quick
0
1
sort


SoftNER
Entity Segmenter
Word Context
Word Markdown
Word Frequency
1
1
0
1
1
Feed Forward Layer
cpp
QSort
uses
quick
sort
Word Context
Code Recognizer
Entity Segmenter
Input Text
0
1
0
0
1
1
1
0
1
cpp
QSort()
uses
quick
sort

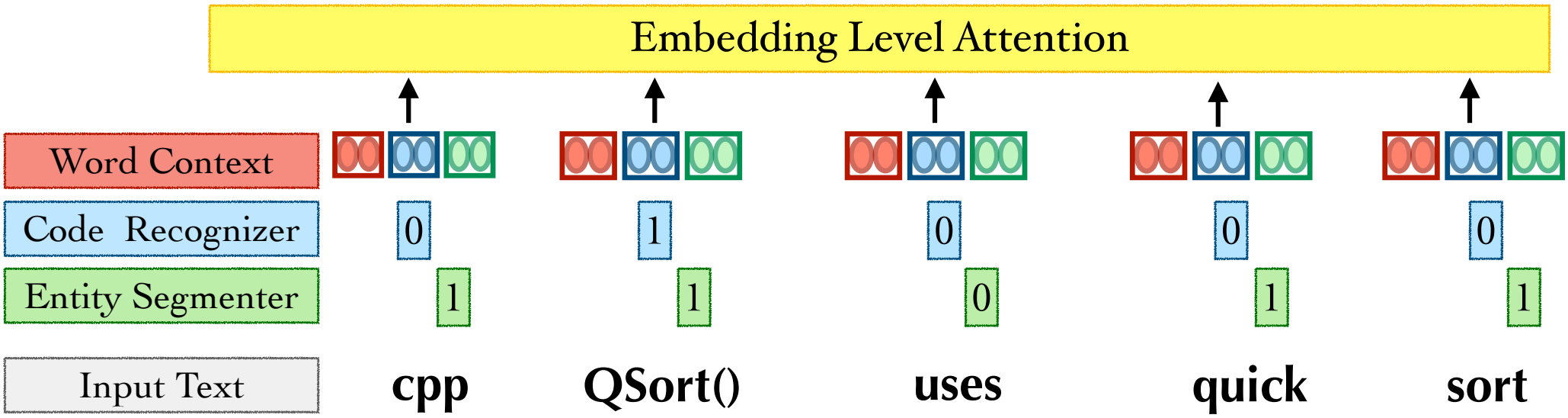

Entity Segmenter
Word Context
Word Markdown
Word Frequency
1
1
0
1
1
Feed Forward Layer
Word Context
Code Recognizer
Entity Segmenter
Input Text
0
1
1
0
1
0
1
0
0
1
cpp
QSort()
uses
quick
sort


Entity Segmenter
Word Context
Word Markdown
Word Frequency
1
1
0
1
1
Feed Forward Layer
Word Context
Code Recognizer
Entity Segmenter
Input Text
0
1
1
0
1
0
1
0
0
1
1
cpp
QSort()
uses
quick
sort

SoftNER


Word Context
Code Recognizer
Entity Segmenter
Input Text
0
1
cpp
1
1
QSort()
0
0
uses
0
1
quick
0
1
sort

SoftNER


Embedding Level Attention
Word Context
Code Recognizer
Entity Segmenter
Input Text
0
1
0
1
0
0
1
0
1
1
cpp
QSort()
uses
quick
sort

Embedding Level Attention

input embeddings

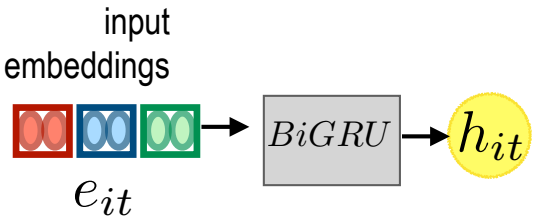
e_{it}

Entity Segmenter 1 1 0 1 1

Input Text cpp QSort() uses quick sort

SoftNER

Embedding Level Attention



Entity Segmenter

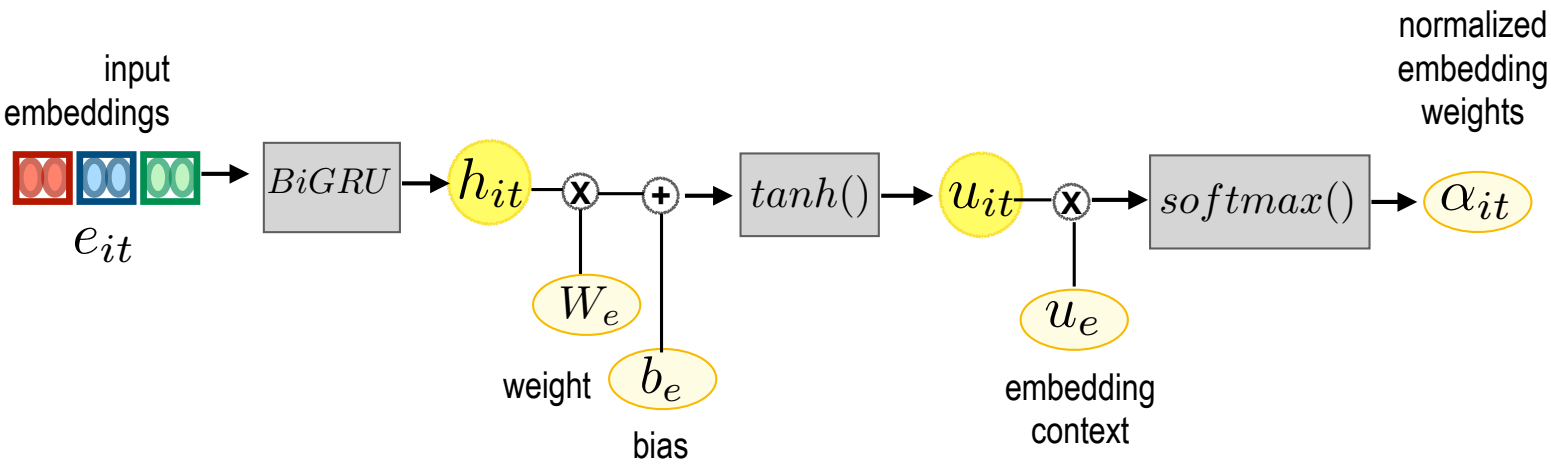
1 1 0 1 1

Input Text

cpp QSort() uses quick sort

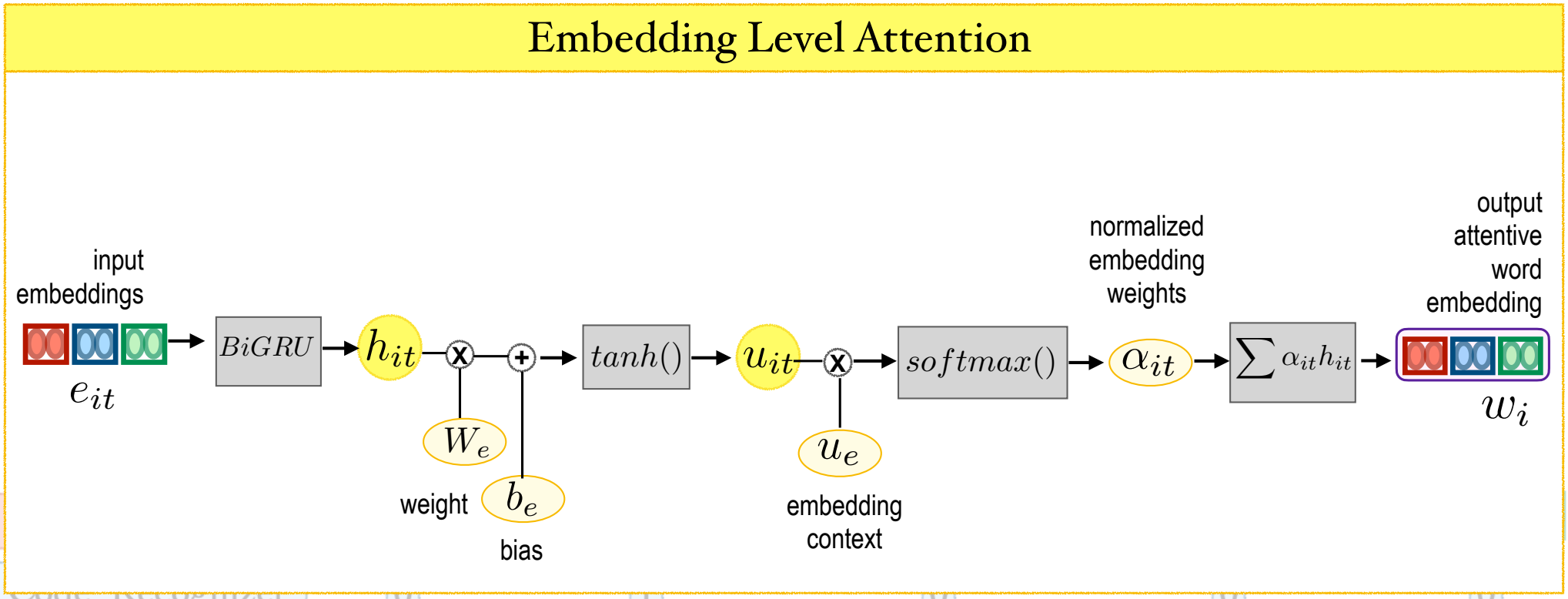
SoftNER

Embedding Level Attention

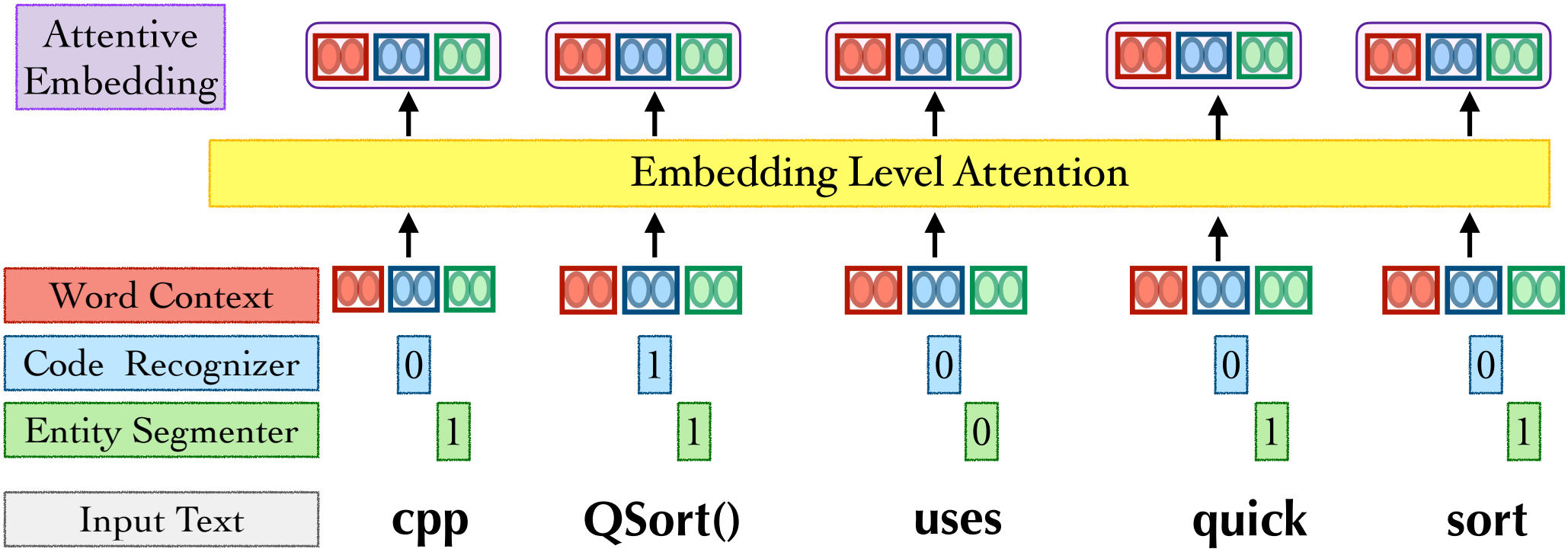


Entity Segmenter 1 1 0 1 1

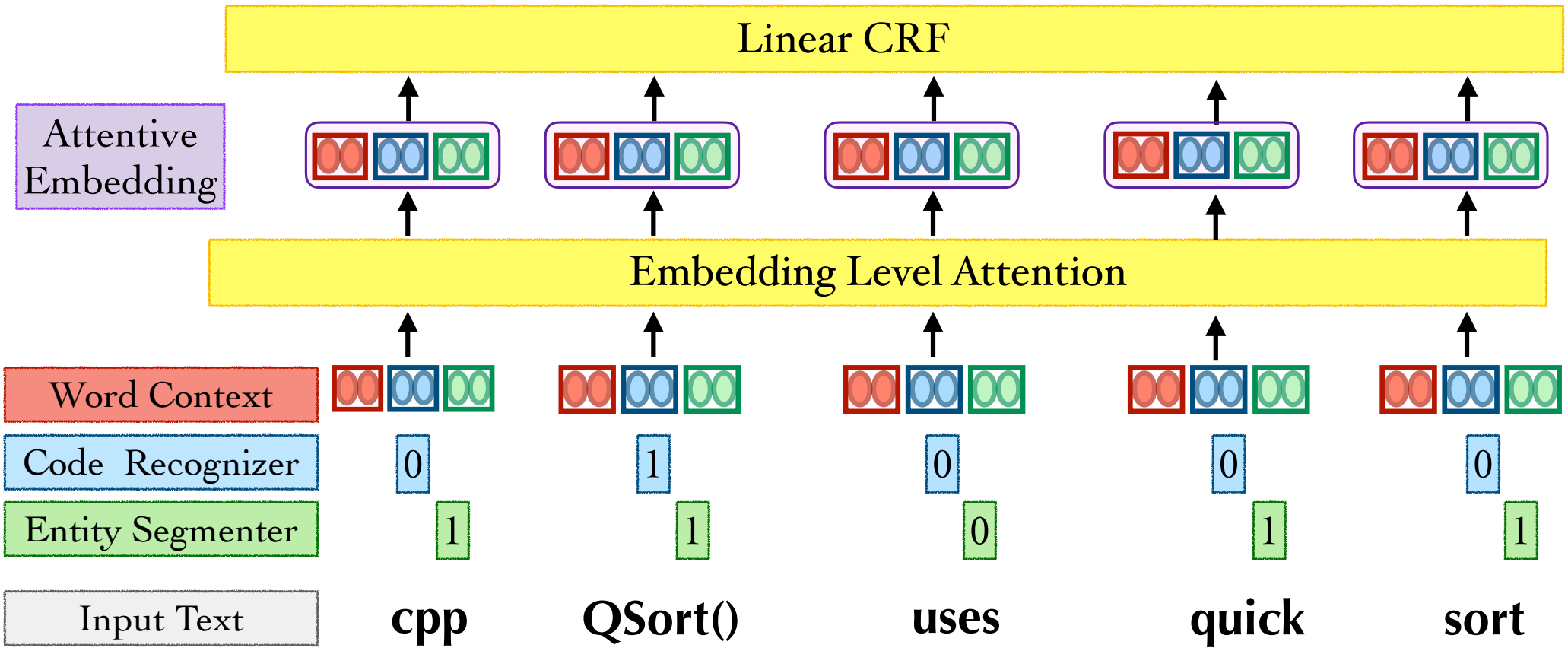
Input Text cpp QSort() uses quick sort

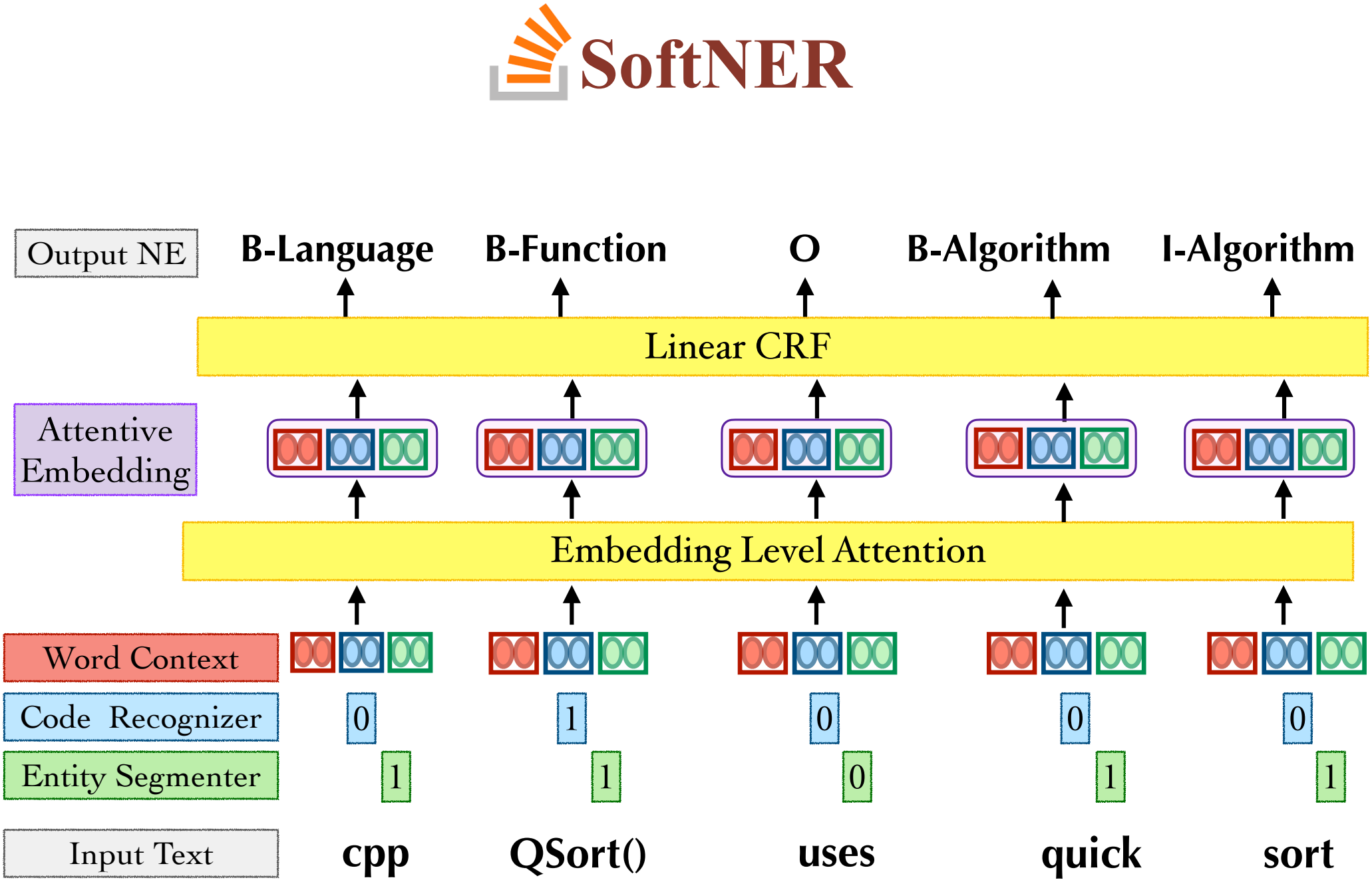


SoftNER


Attentive Embedding
Embedding Level Attention
Word Context
Code Recognizer
Entity Segmenter
Input Text
0
1
0
0
1
1
0
1
0
1
0
1
cpp
QSort()
uses
quick
sort

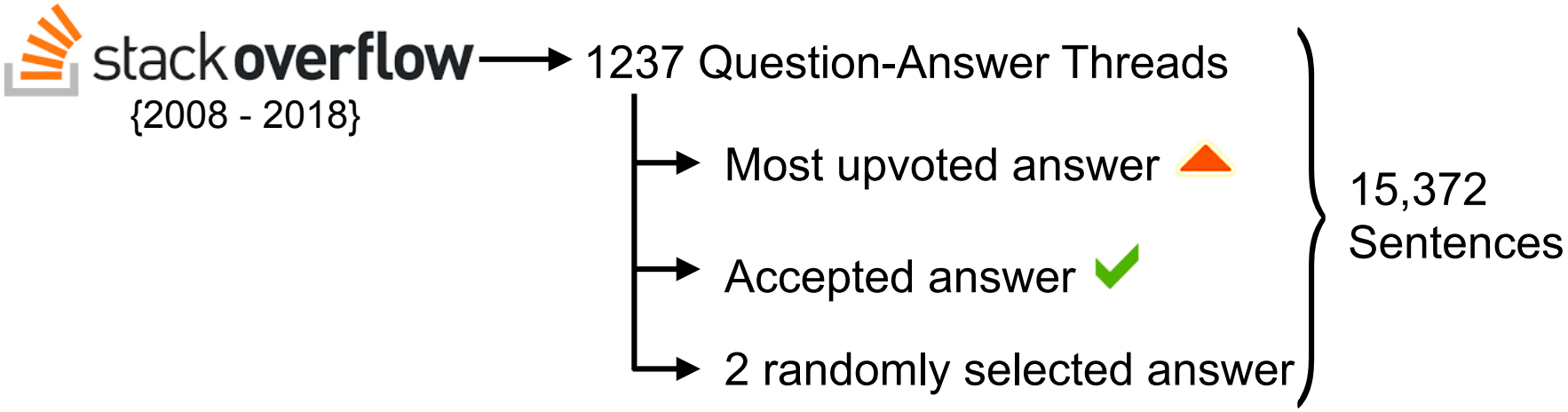
SoftNER


Linear CRF
Attentive Embedding
Embedding Level Attention
Word Context
Code Recognizer
Entity Segmenter
Input Text
0
1
0
0
0
1
1
0
1
1
cpp
QSort()
uses
quick
sort


SoftNER
Output NE
B-Language
B-Function
O
B-Algorithm
I-Algorithm
Linear CRF
Attentive Embedding
Embedding Level Attention
Word Context
Code Recognizer
Entity Segmenter
Input Text
0
1
0
0
0
1
1
0
1
1
cpp
QSort()
uses
quick
sort

SoftNER

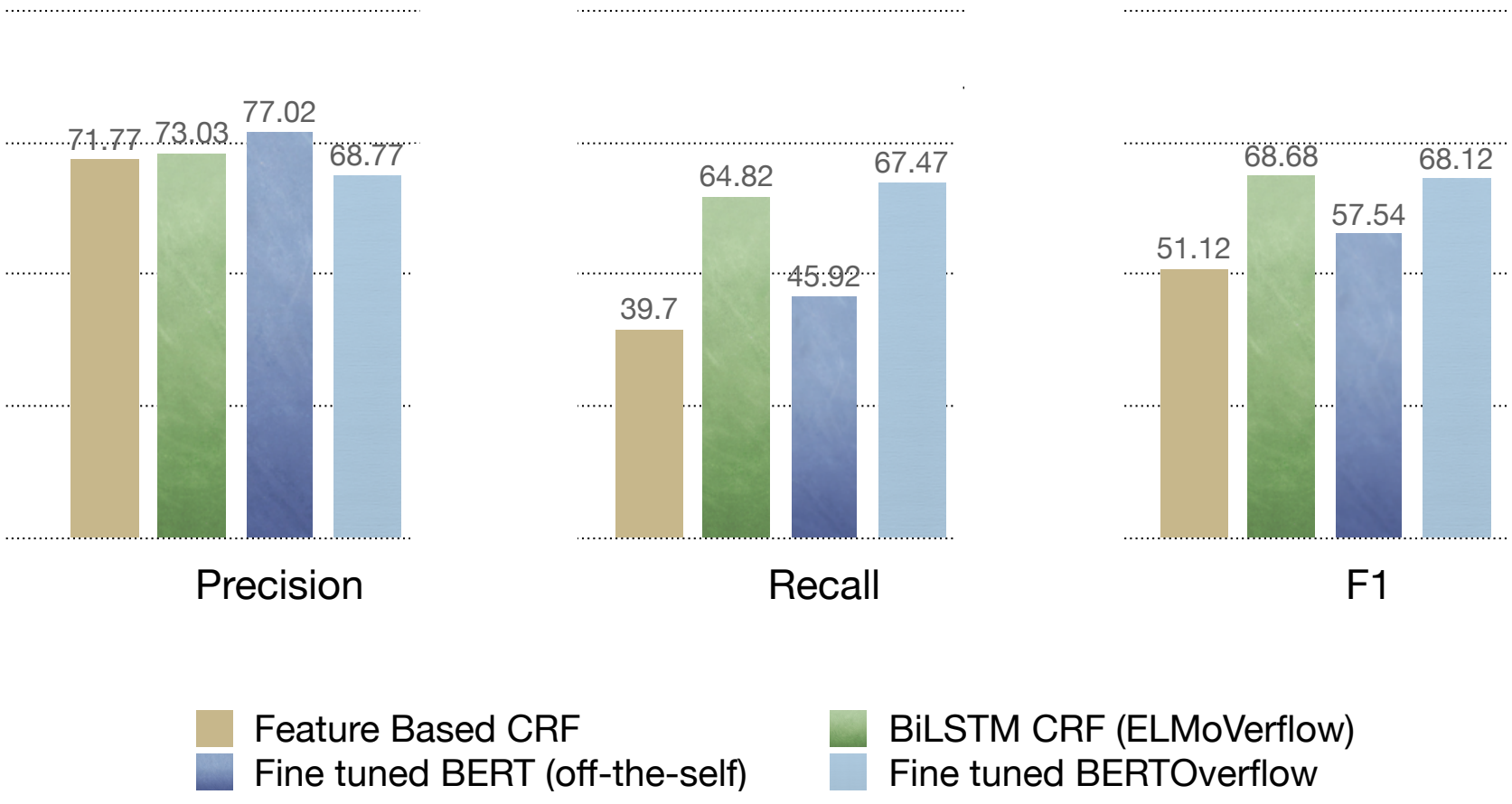
Annotated Corpus



Single annotated	**Double annotated**	
Train Set	Dev Set	Test Set
741 Question-Answers 9,315 Sentences	247 Question-Answers 2,942 Sentences	249 Question-Answers 3,115 Sentences

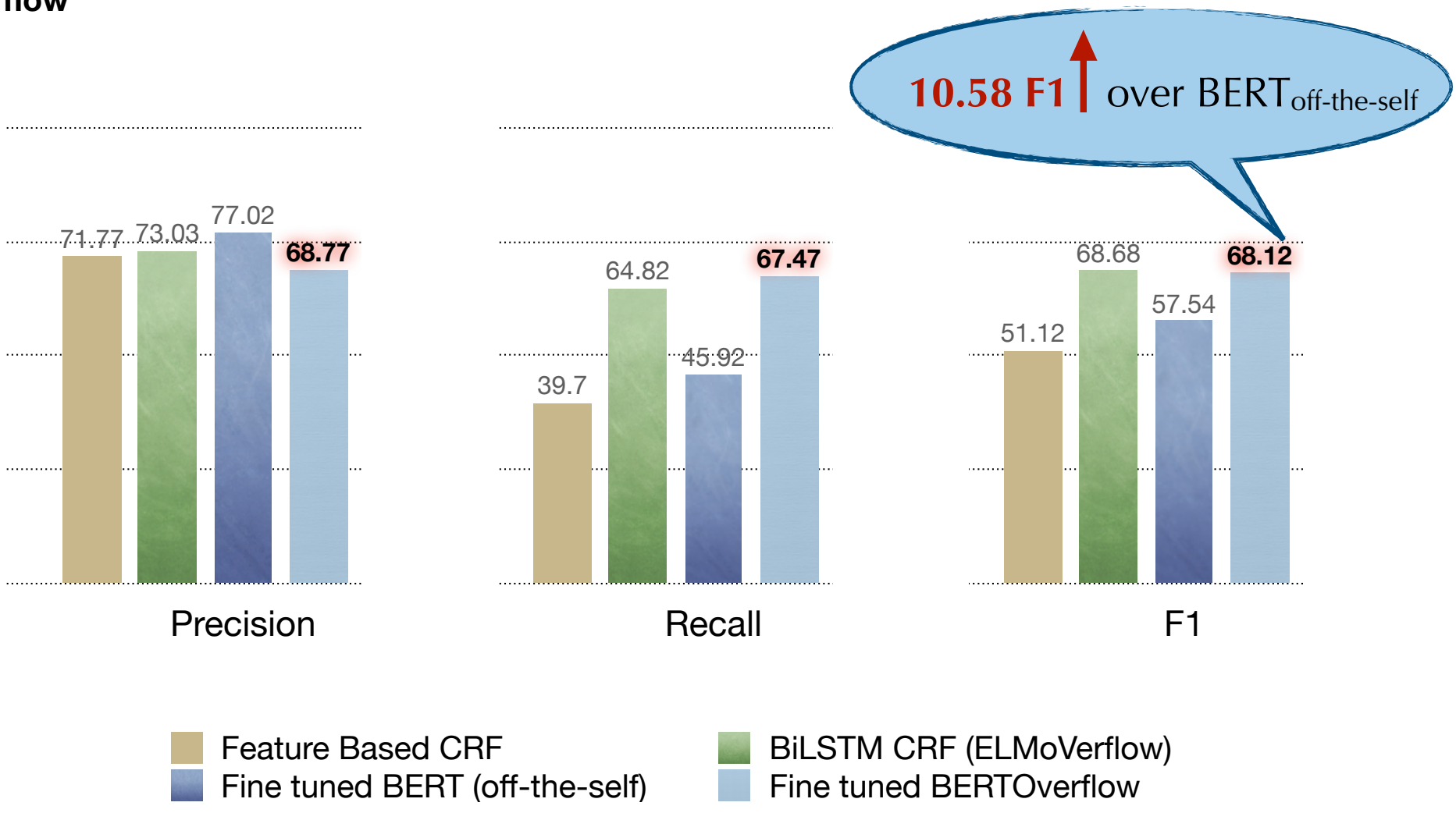
StackOverflow

Performance



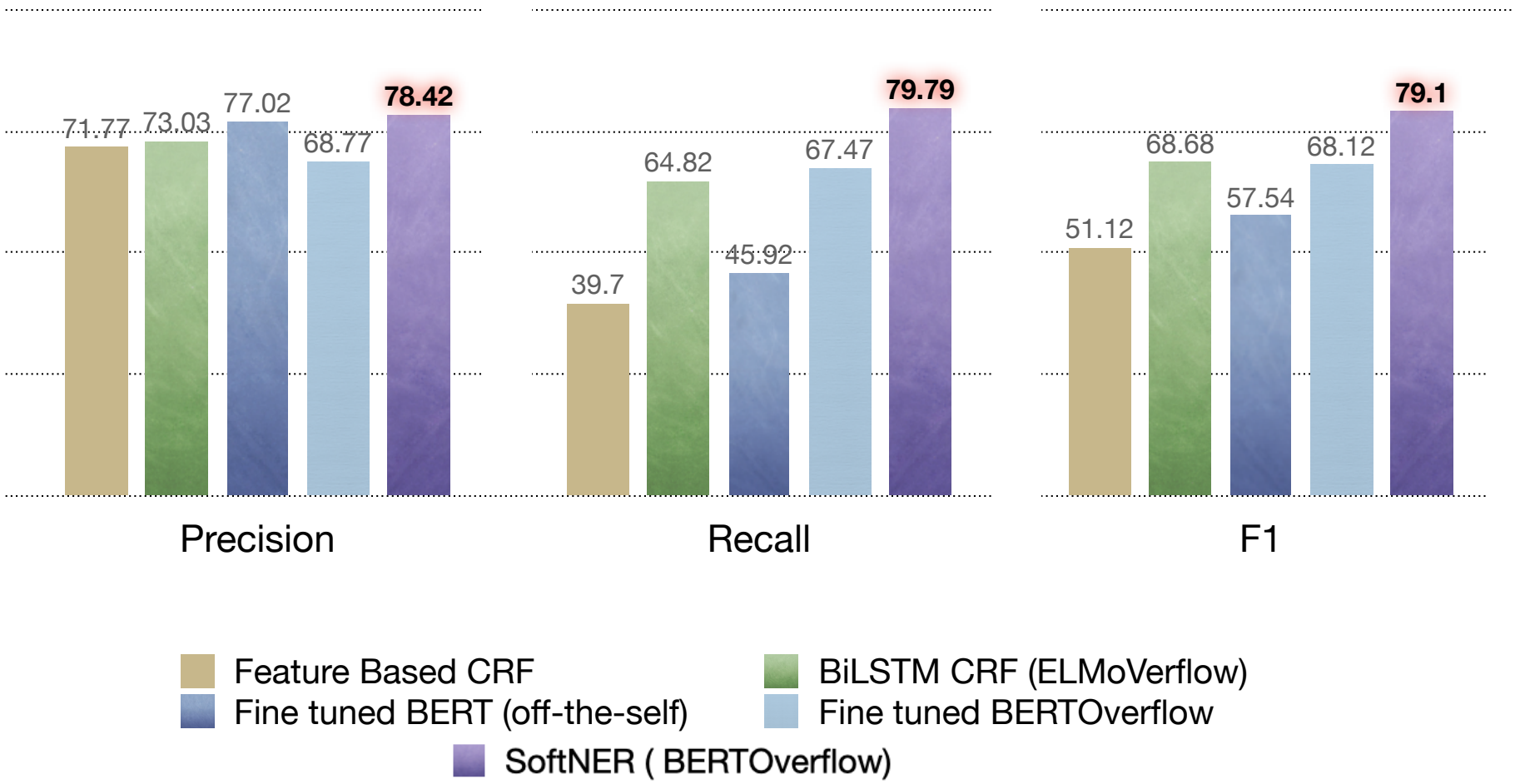
StackOverflow

Performance



StackOverflow

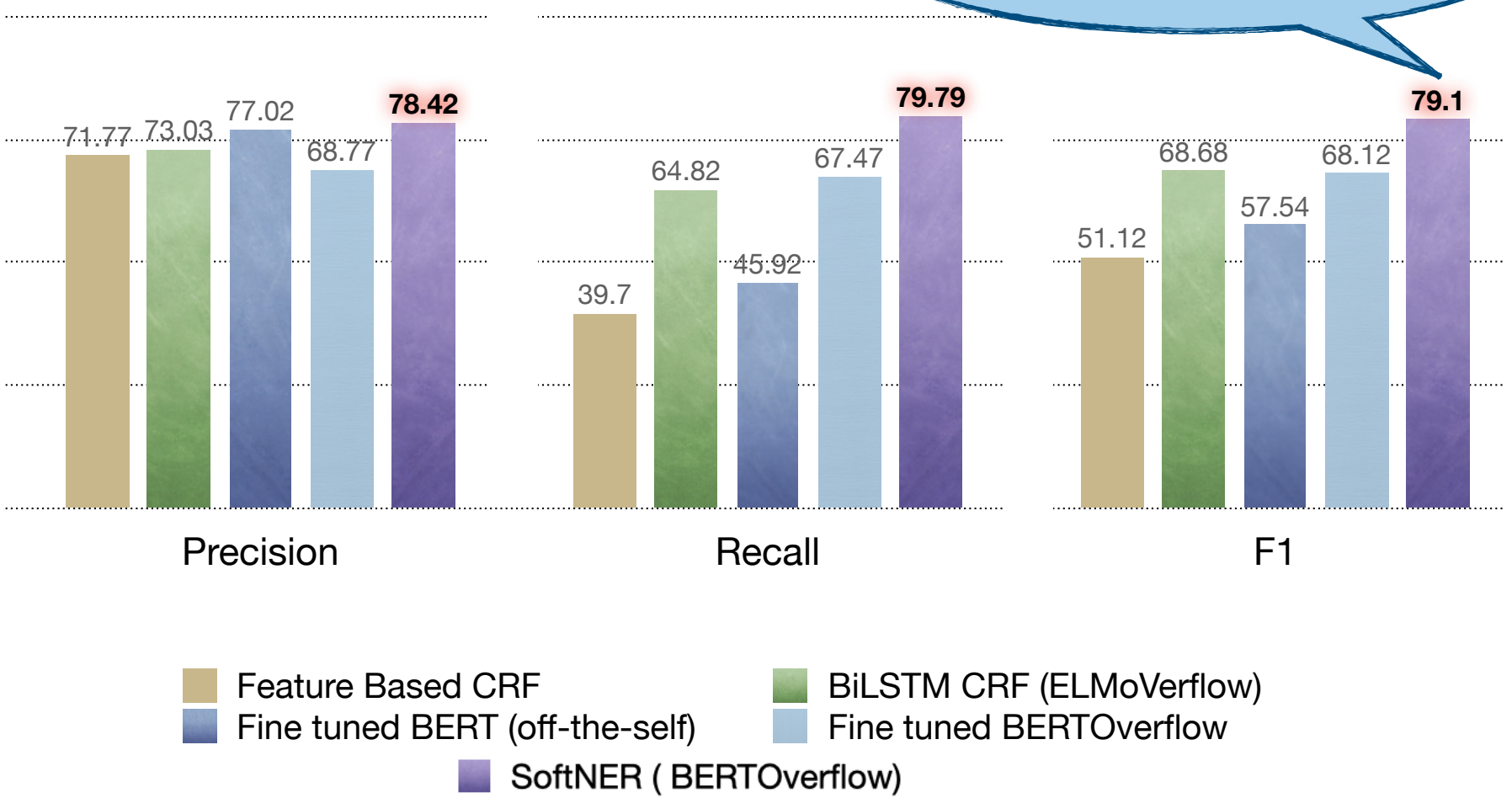
Performance



StackOverflow

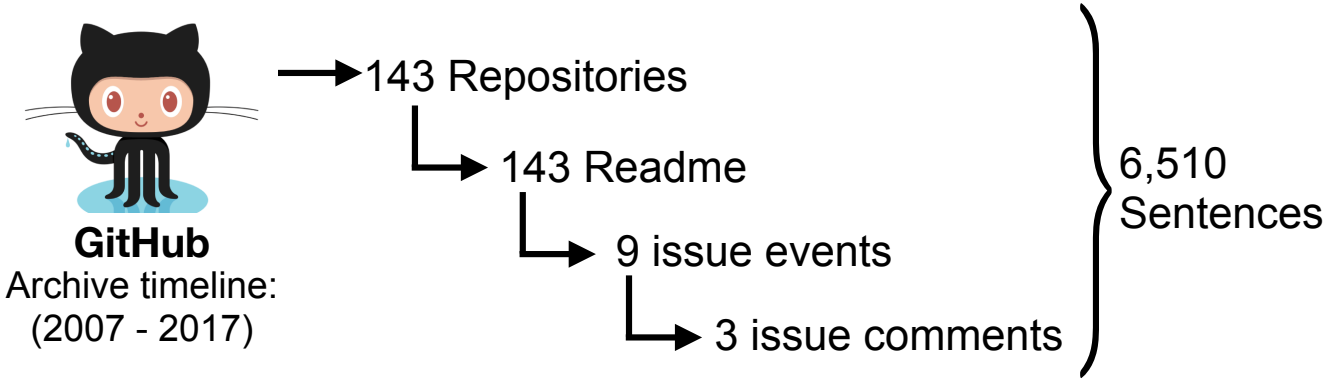
Performance

10.98 F1 over fine-tuned BERTOverflow

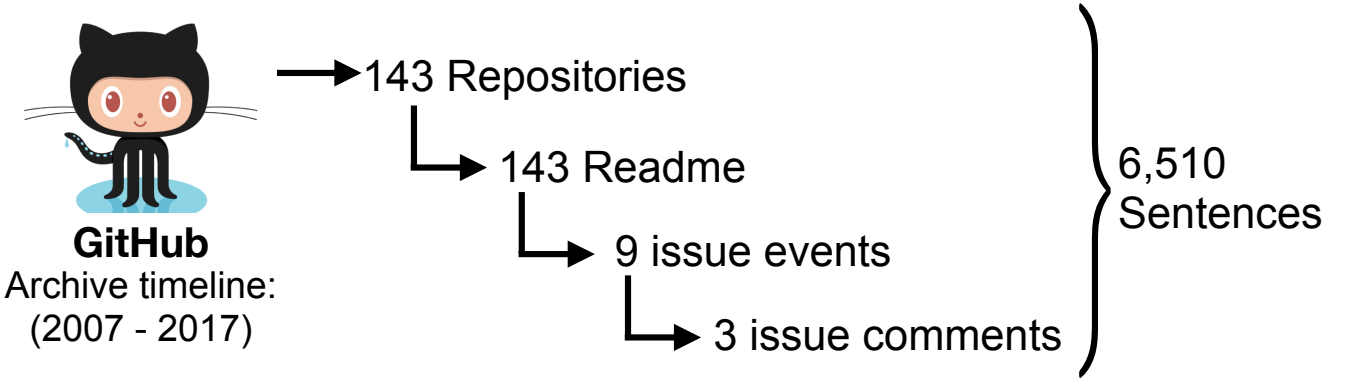


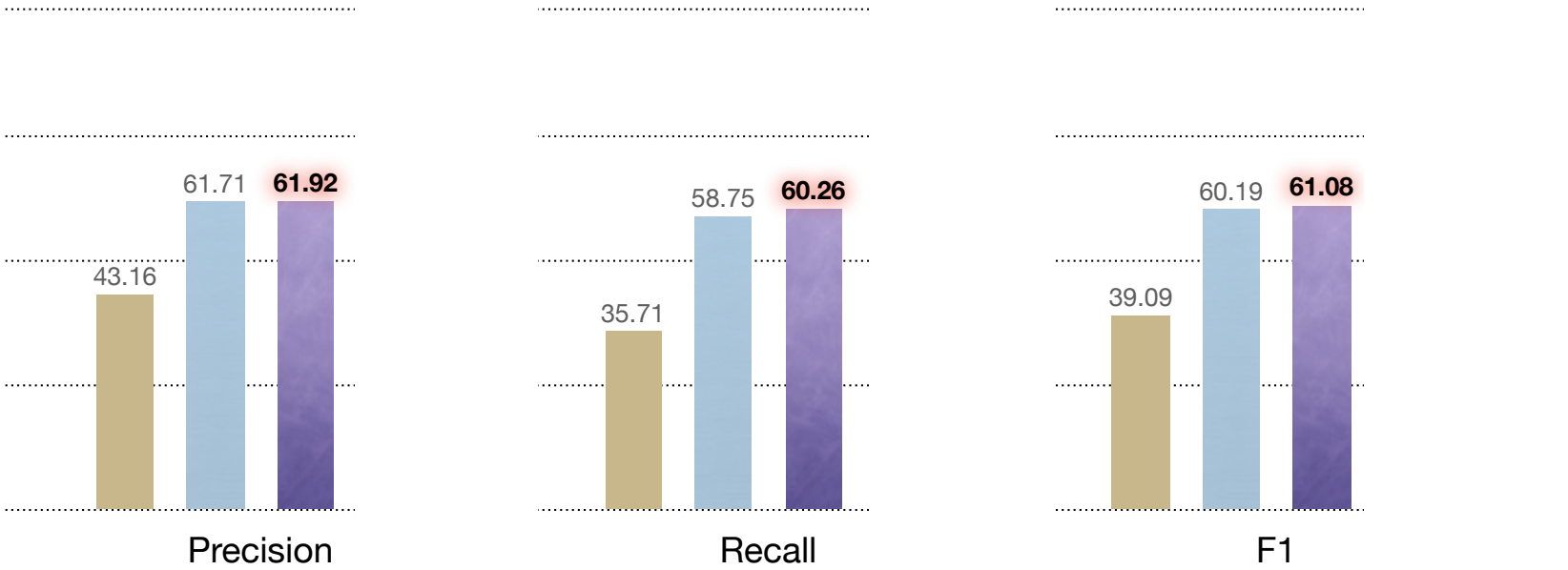
Does SoftNER work on **other social software domain** text?

Performance on GitHub Test Set



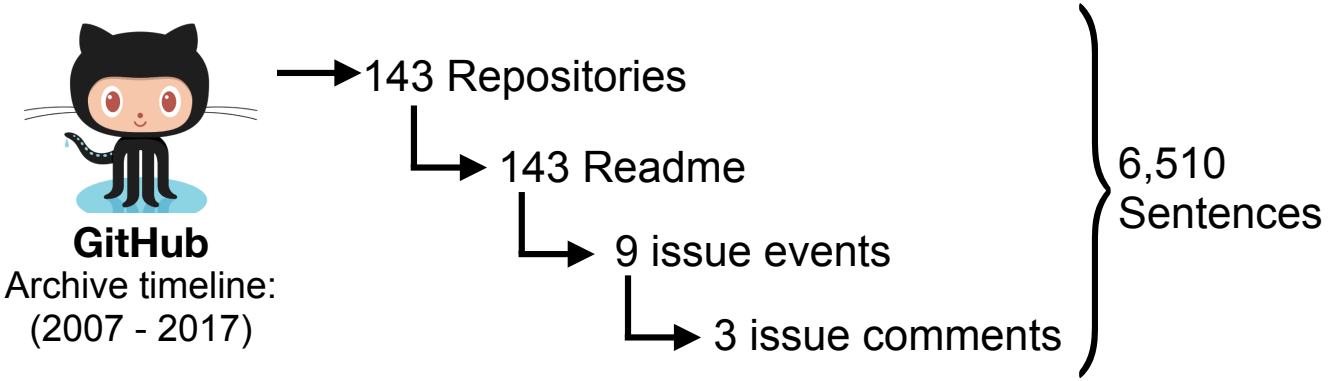
Performance on GitHub Test Set

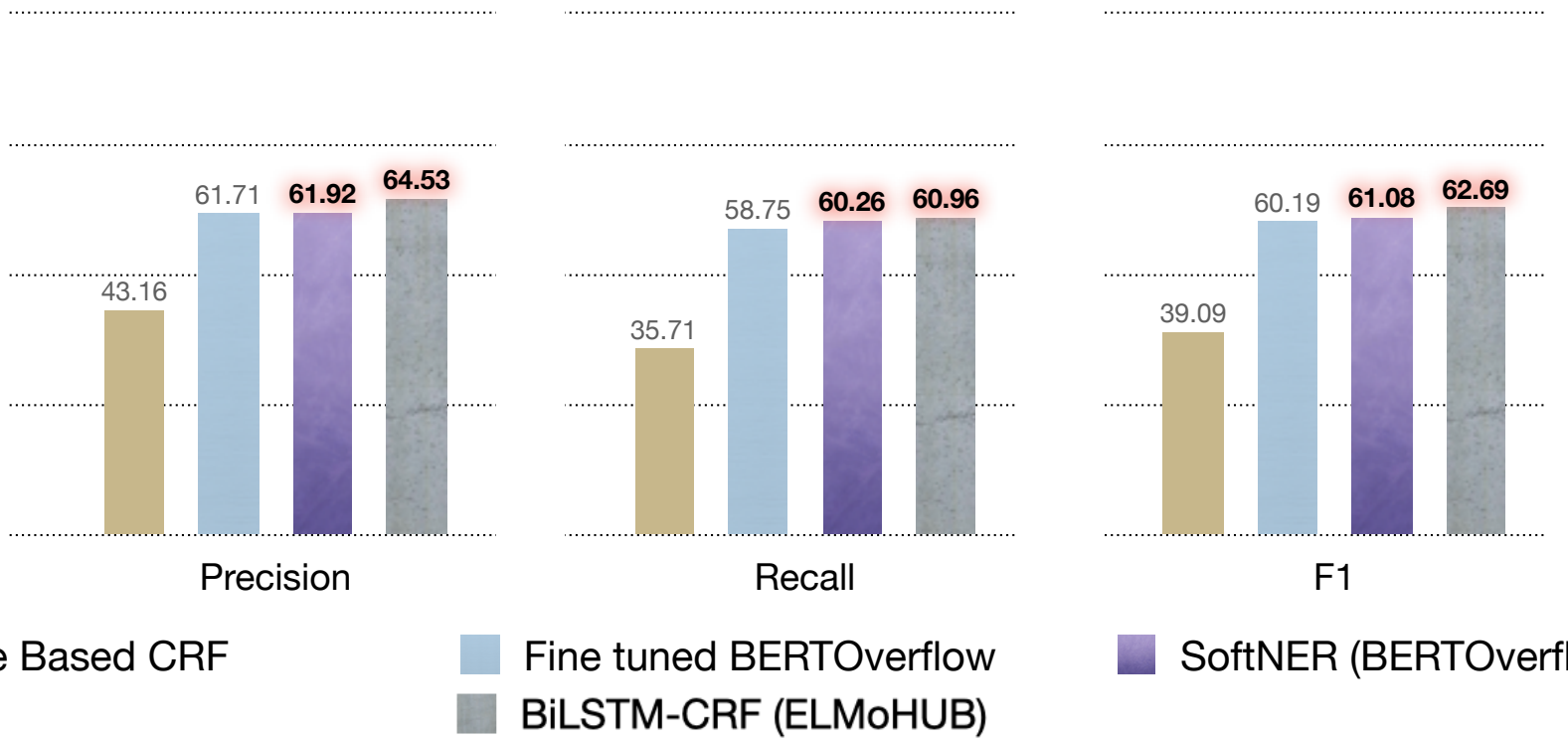




Trained on 9,315 manually annotated **StackOverflow** Sentences

Performance on GitHub Test Set





Trained on 9,315 manually annotated **StackOverflow** Sentences

Takeaway

Task: Fine-Grained Software Entity Extraction

- Covers **20 types** of fine-grained named entities
- Works on StackOverflow and Github Data

StackOveflow

code and data at: https://github.com/jeniyat/StackOverflowNER

GitHub

Takeaway

Task: Fine-Grained Software Entity Extraction

- Covers **20 types** of fine-grained named entities
- Works on StackOverflow and Github Data

New Data

- **15k** manually annotated sentences with fine grained software entities
- Software domain word vectors trained on **152M** sentences.
  - ❖ BERTOverflow, ELMoVerflow, GLoVerflow

StackOveflow

code and data at: https://github.com/jeniyat/StackOverflowNER

GitHub

Takeaway

Task: Fine-Grained Software Entity Extraction

- Covers **20 types** of fine-grained named entities
- Works on StackOverflow and Github Data

New Data

- **15k** manually annotated sentences with fine grained software entities
- Software domain word vectors trained on **152M** sentences.
  - BERTOverflow, ELMoVerflow, GLoVerflow

New Model

- Attentive NER tagger to combine the domain knowledge with contextual knowledge
  - **10.98 F1 increase** over fine tuned BERT
- Standalone code token recognizer

StackOveflow

code and data at: https://github.com/jeniyat/StackOverflowNER

GitHub

Takeaway

7 July
5:00 PM UTC
8:00 PM UTC

Task: Fine-Grained Software Entity Extraction

- Covers **20 types** of fine-grained named entities
- Works on StackOverflow and Github Data

New Data

- **15k** manually annotated sentences with fine grained software entities
- Software domain word vectors trained on **152M** sentences.
  - BERTOverflow, ELMoVerflow, GLoVerflow

New Model

- Attentive NER tagger to combine the domain knowledge with contextual knowledge
  - **10.98 F1 increase** over fine tuned BERT
- Standalone code token recognizer

StackOveflow

code and data at: https://github.com/jeniyat/StackOverflowNER

GitHub