

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

Language Application

Using their JavaScript SDK, 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.
Using their JavaScript **SDK**, the following works:

Language **Application**

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.

Using their JavaScript **SDK**, 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**



StackOverflow Entity Recognition



BRAT

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

Library

Language Application

Using their JavaScript SDK, 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 $F_1 \uparrow$



StackOverflow Entity Recognition



BRAT

I am trying to do a search query on the SoundCloud API.
Using their **JavaScript SDK**, the following works:

Language Application
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 $F_1 \uparrow$

BERTOverflow

Pre-trained on 152M Sentences





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]



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]



Code Retrieval

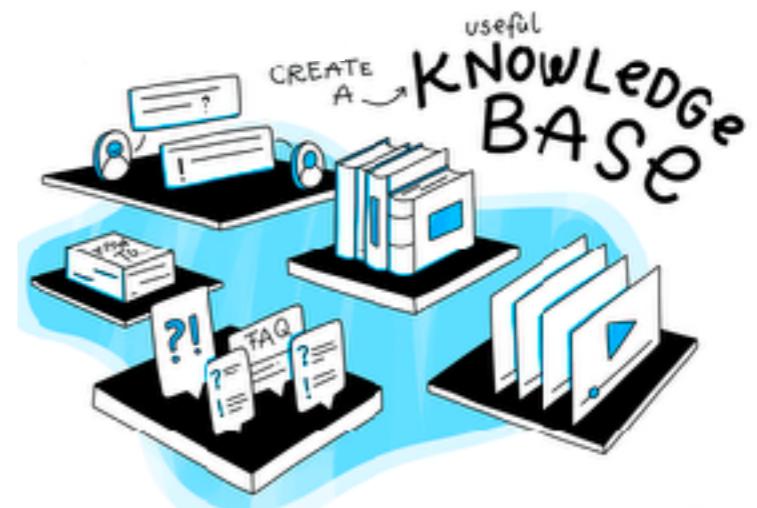
NL Query:

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
```

KB Creation



[Yao et al., 2019]

[Iyer et al. 2016]

[Giorgi and Bader 2018]

[Movshovitz-Attias and Cohen 2015]



Code Retrieval

NL Query:

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

Code Snippet

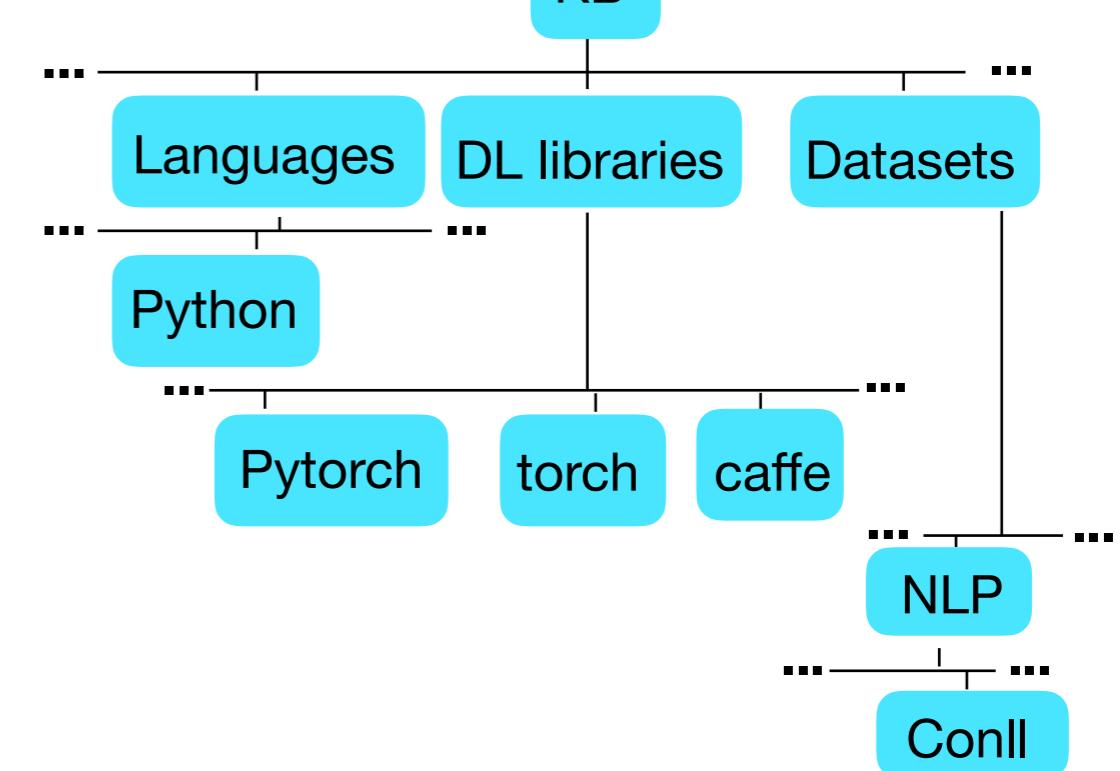
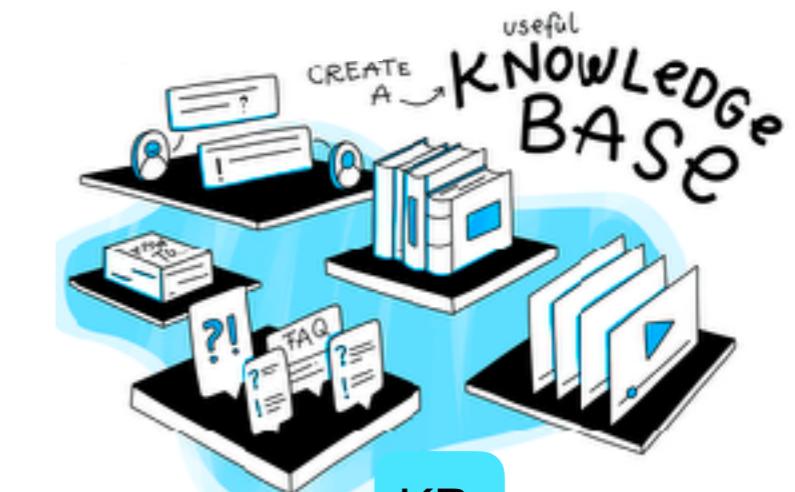
```
SELECT url FROM links WHERE date =  
  ( SELECT date FROM links WHERE  
    date < "2014/02/25" ORDER BY date  
    DESC LIMIT 1)  
ORDER BY date DESC, clicks DESC LIMIT 200
```

[Yao et al., 2019]

[Iyer et al. 2016]

[Giorgi and Bader 2018]

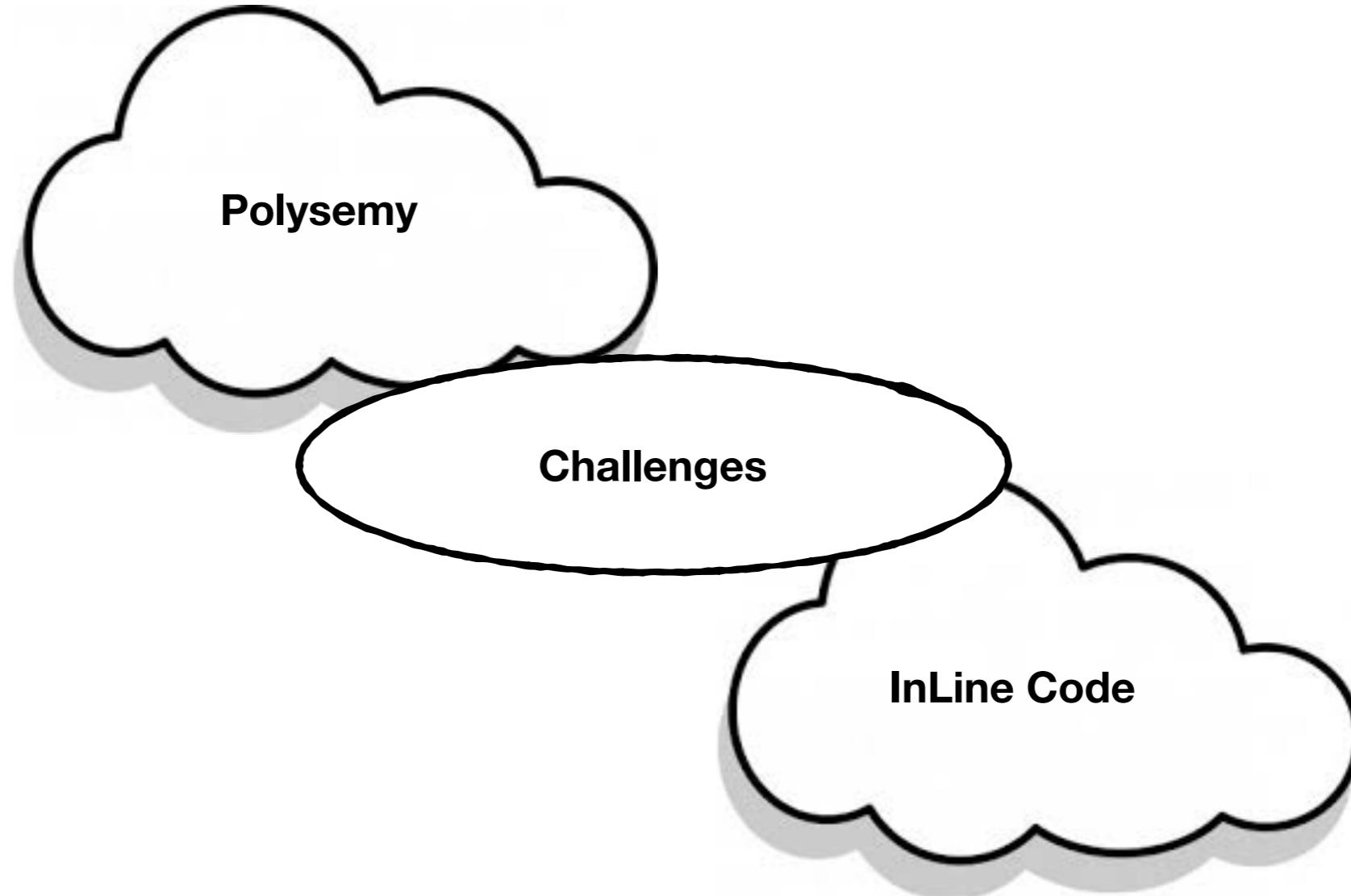
KB Creation



[Movshovitz-Attias and Cohen 2015]

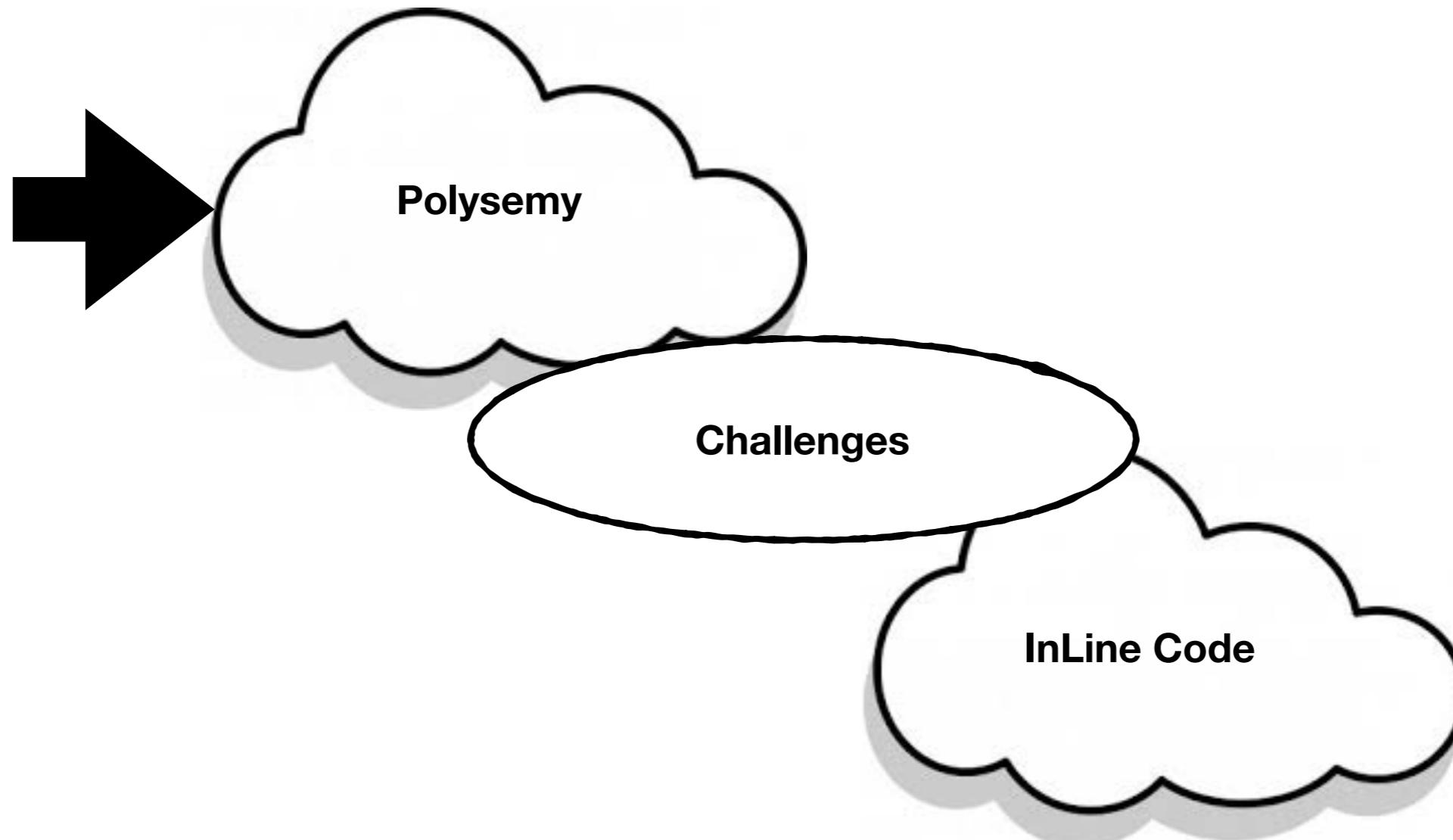


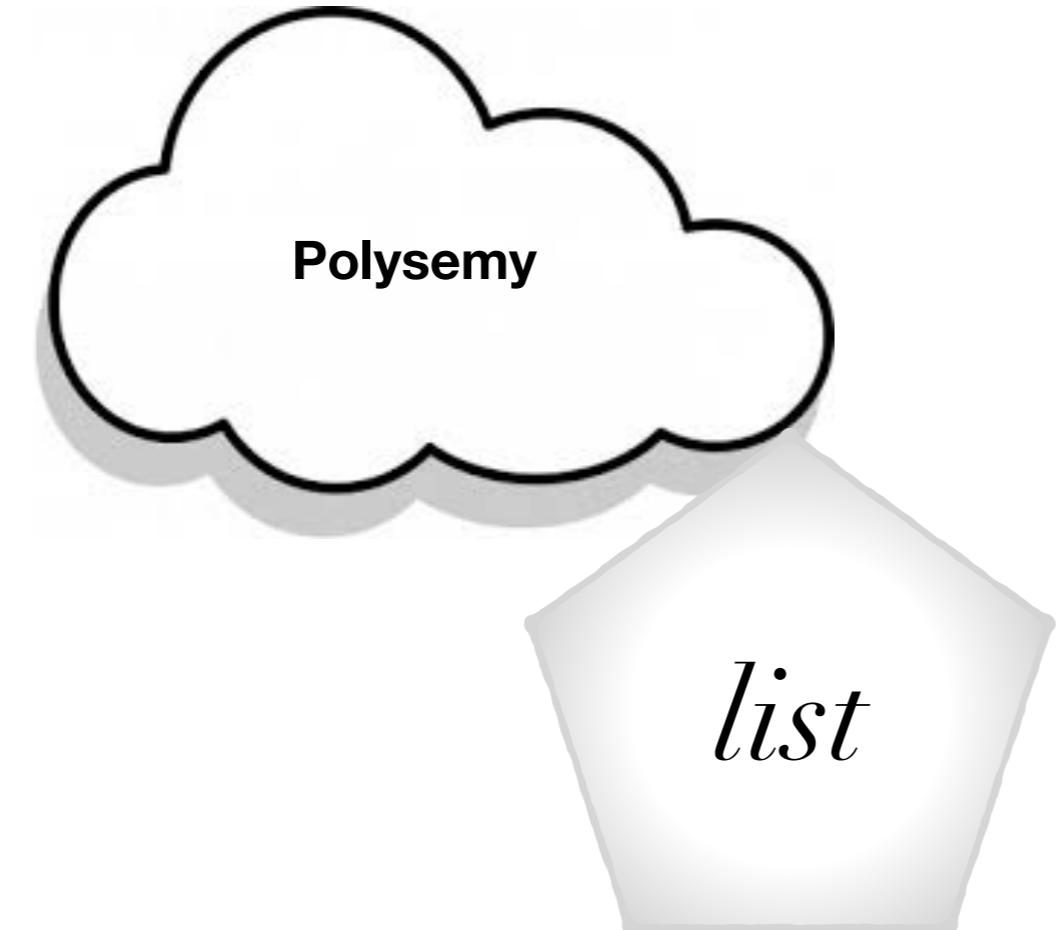
StackOverflow Entity Recognition





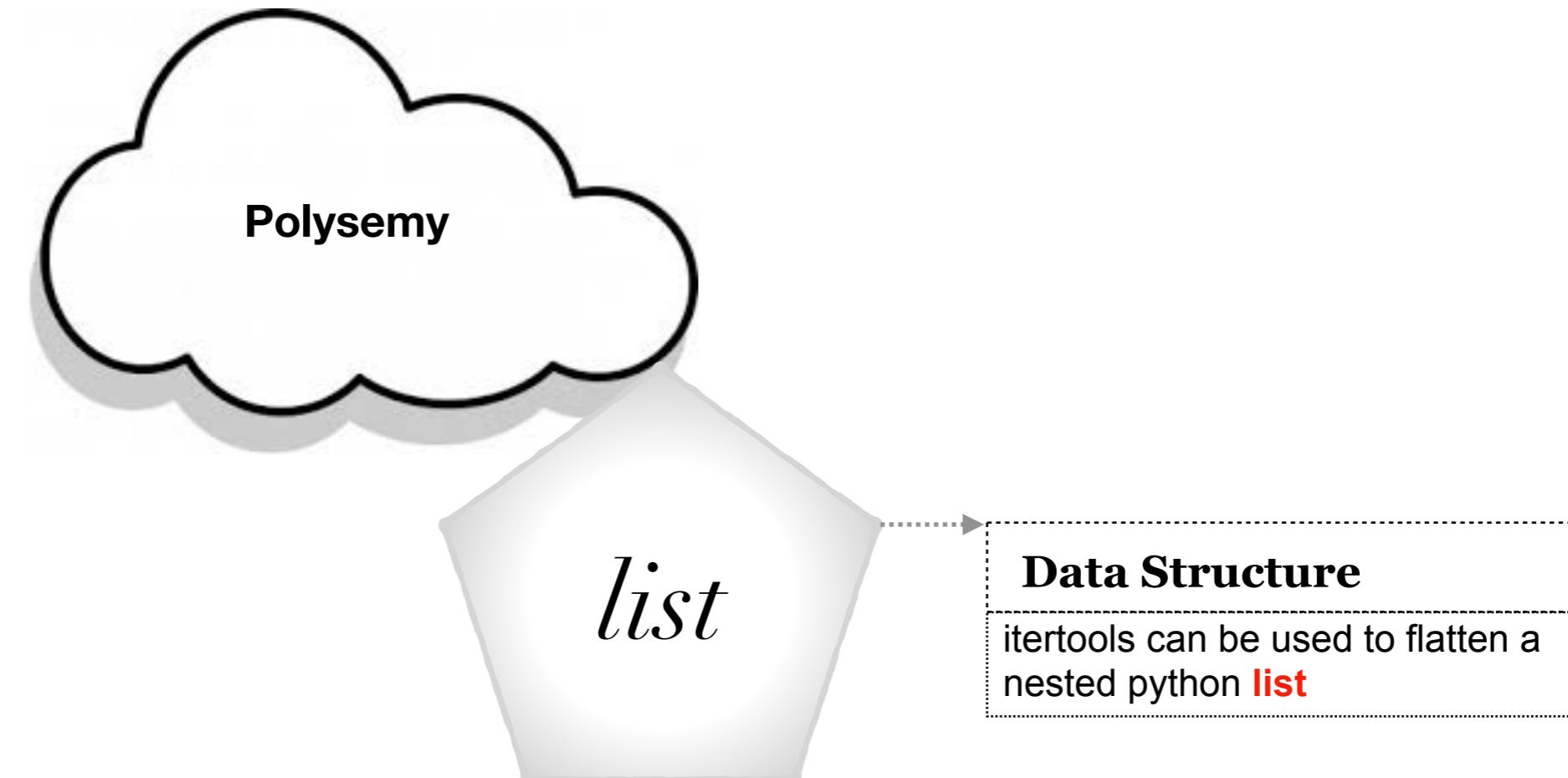
StackOverflow Entity Recognition





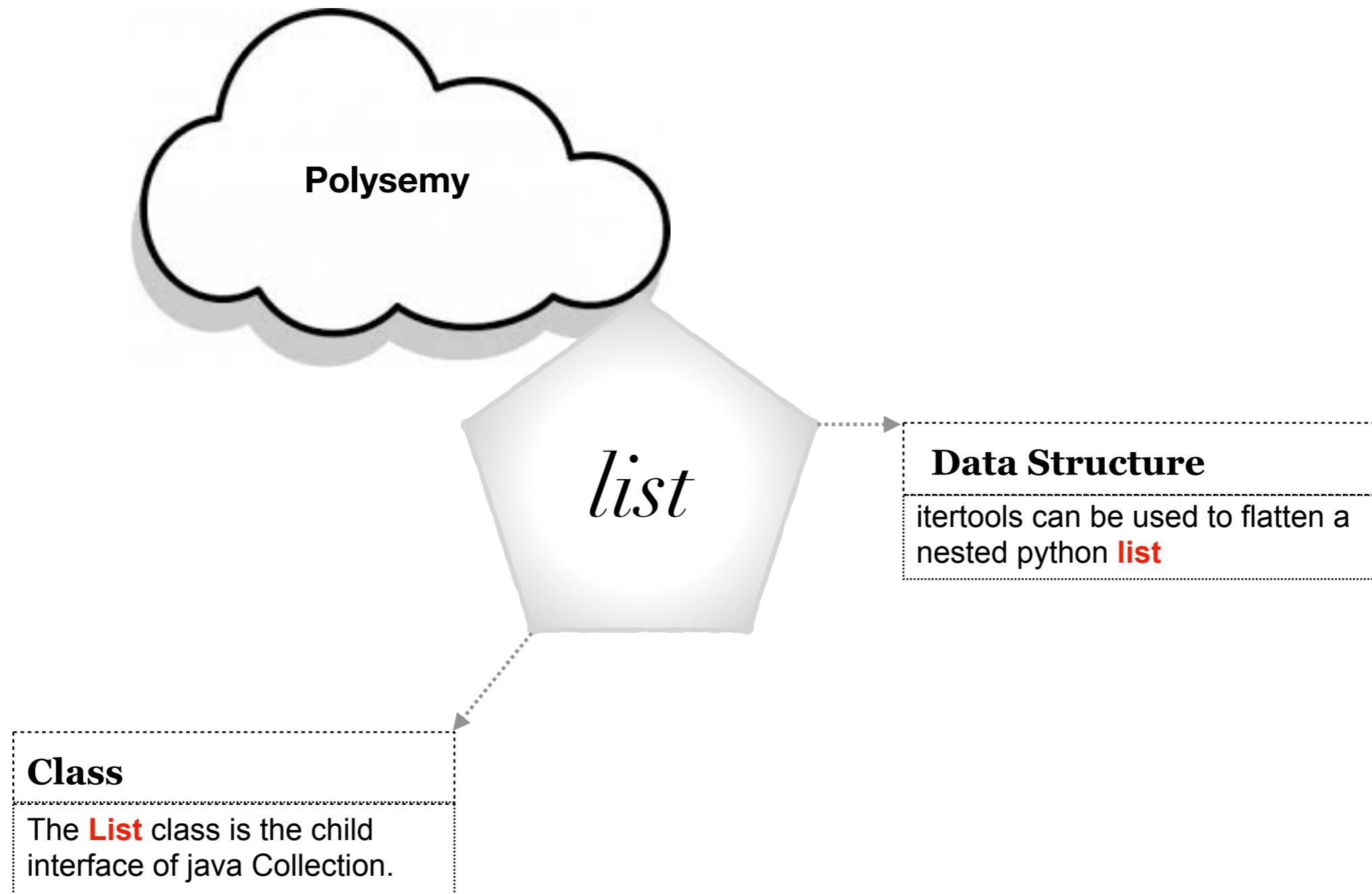


StackOverflow Entity Recognition



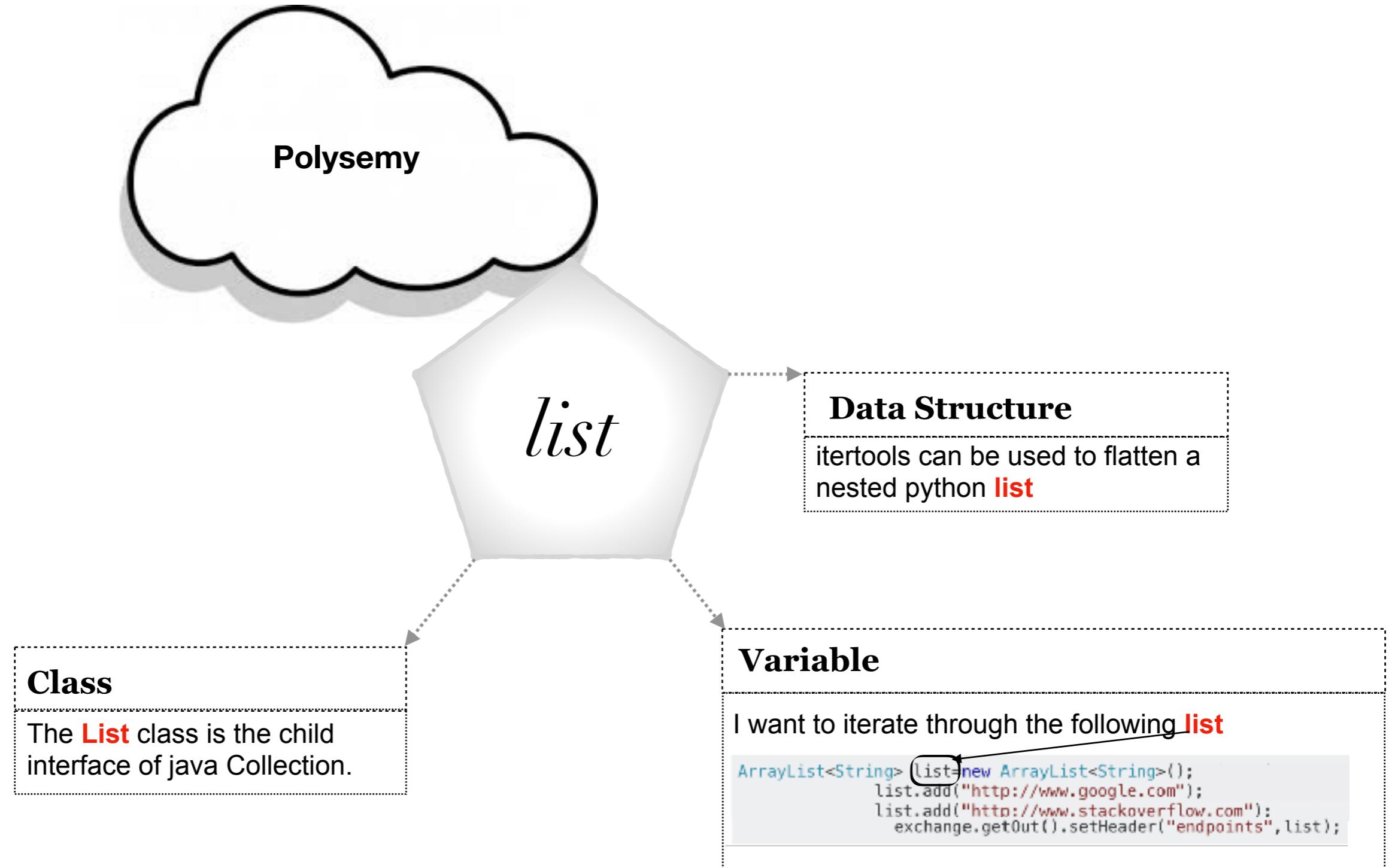


StackOverflow Entity Recognition



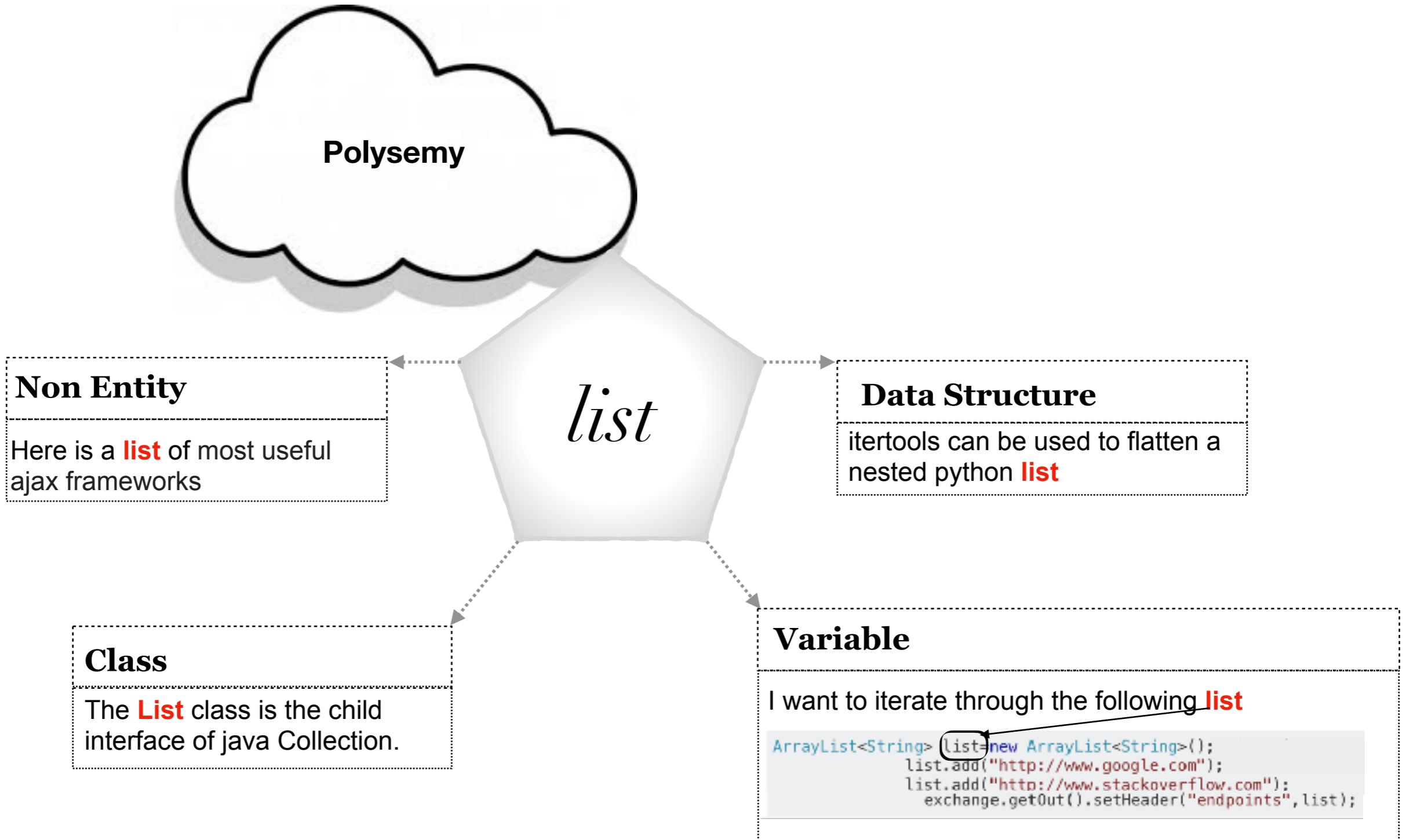


StackOverflow Entity Recognition



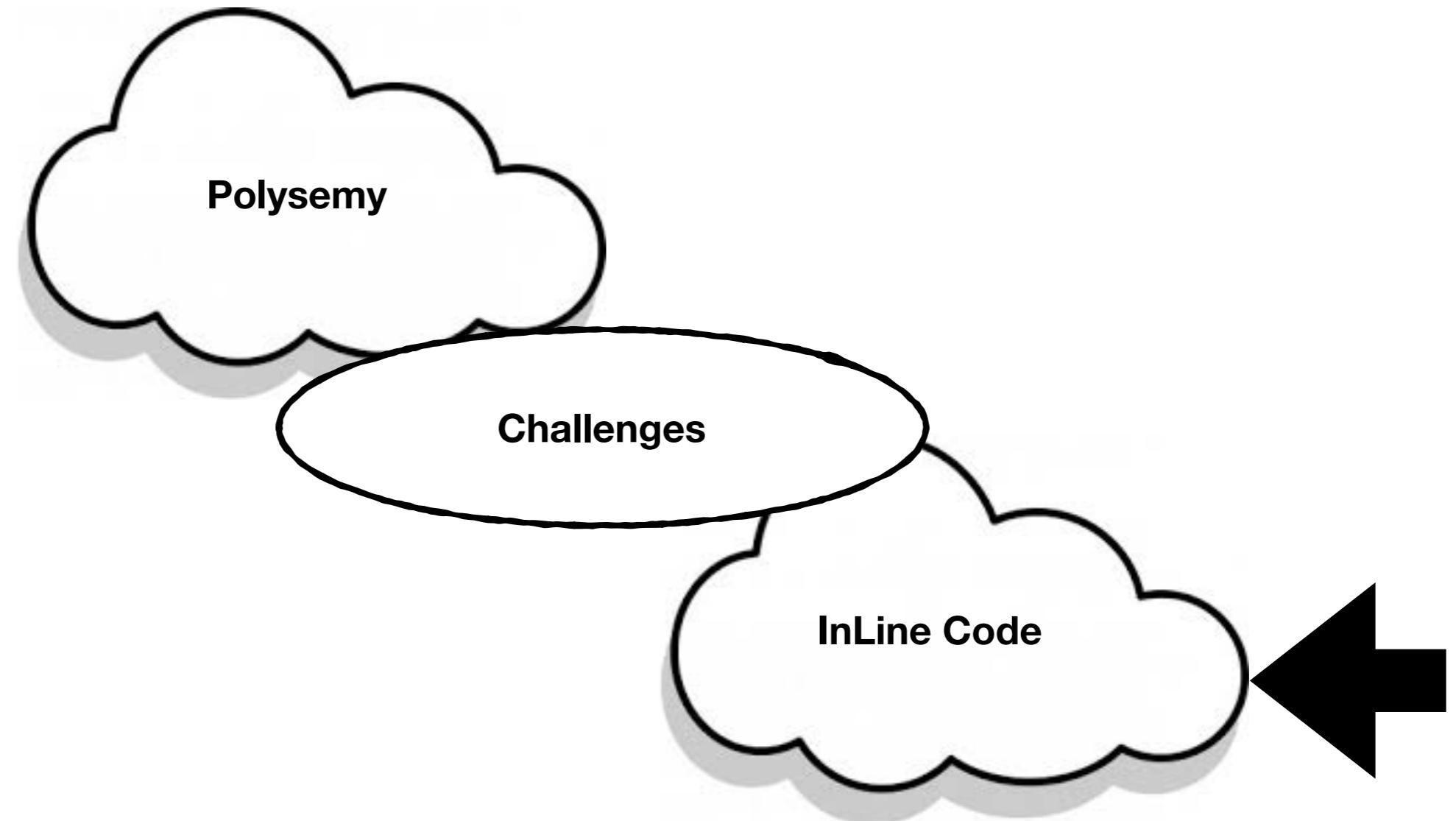


StackOverflow Entity Recognition



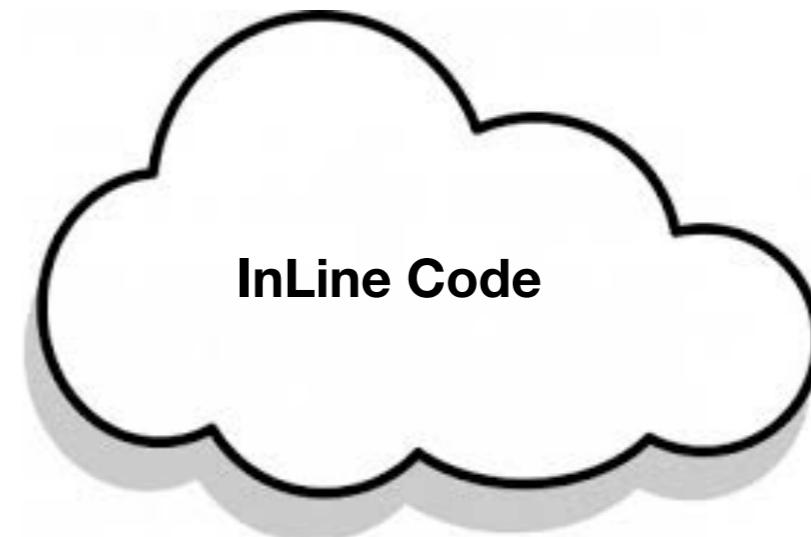


StackOverflow Entity Recognition





StackOverflow Entity Recognition



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



StackOverflow Entity Recognition



InLine Code

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



StackOverflow Entity Recognition

980 Unique Function Names in 15k sentences



InLine Code

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($key)` function.
If it return false, then, convert key to integer using typecasting, `(int)$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

StackOverflow Entity Recognition

980 Unique Function Names in 15k sentences



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

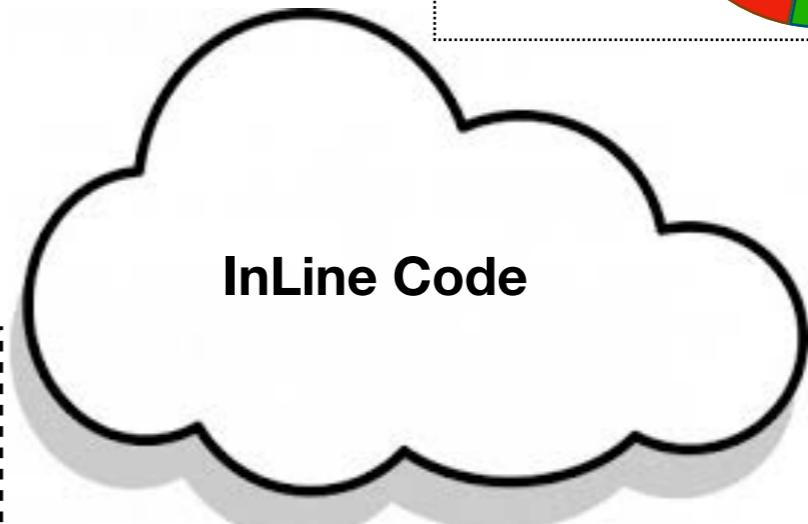
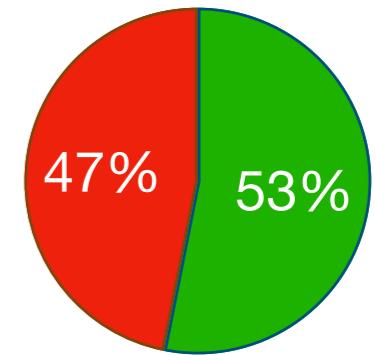


300 1 5

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

User Identified Code Entity

- Code Entity inside <code> tag
- Code Entity outside <code> tag



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

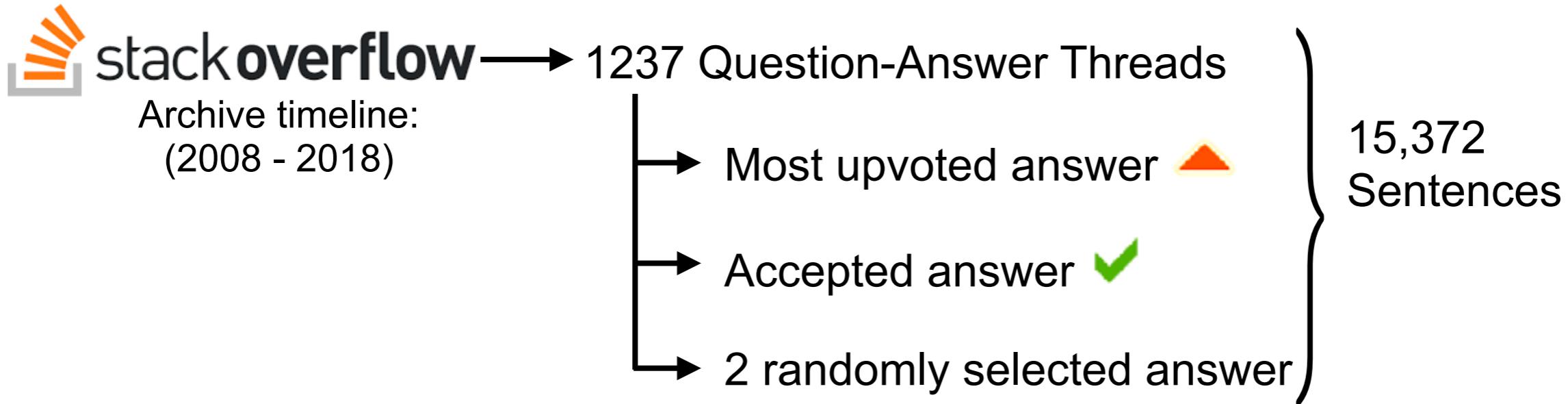
Contribution #2: Propose a **new model** for Software Entities

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



stack**overflow**

Archive timeline:
(2008 - 2018)

→ 1237 Question-Answer Threads

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

15,372
Sentences

20 Entity
Types

Class	Value	Function
HTML_XML_Tag		In_Line_code
Variable	Data_Type	Library

Code Entity Types

Algorithm	Application	Data_Structure
	File_Type	Version
Website	File_Name	Language
	Operating_System	User_Name

Natural Language Entity

Annotated StackOverflow Corpus



stack**overflow**

Archive timeline:
(2008 - 2018)

→ 1237 Question-Answer Threads

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

15,372
Sentences

20 Entity
Types



Class	Value	Function
HTML_XML_Tag	In_Line_code	
Variable	Data_Type	Library

Code Entity Types

Algorithm	Application	Data_Structure
	File_Type	Version
Website	File_Name	Language
	Operating_System	User_Name

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

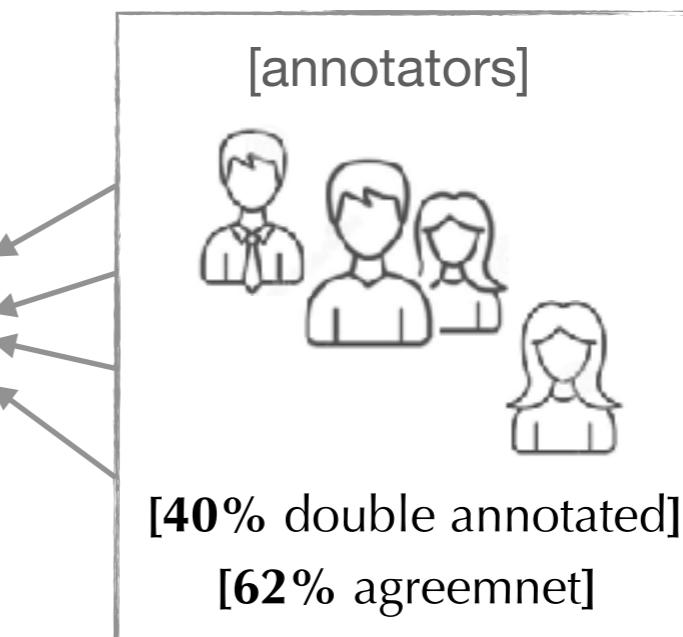


1 Answer to Question ID: 27926052
2 Question URL: <https://stackoverflow.com/questions/27926052/>

4 You can now add customised selector(please refer 'TQUITView+TQKeyboardToolbar.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 'TQUITView+TQKeyboardToolbar.h', for 'how to use?'
please refer 'TextFieldViewController.m'.



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



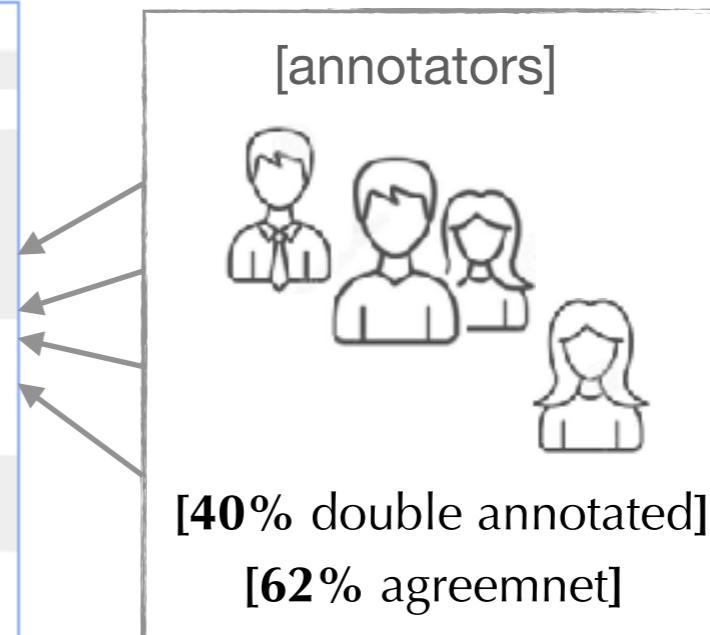
1 Answer to Question ID: 27926052
2 Question URL: <https://stackoverflow.com/questions/27926052/>

4 You can now add customised selector(please refer 'TQUITView+TQKeyboardToolbar.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 'TQUITView+TQKeyboardToolbar.h', for 'how to use?'

7 please refer 'TextFieldViewController.m'.



- 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

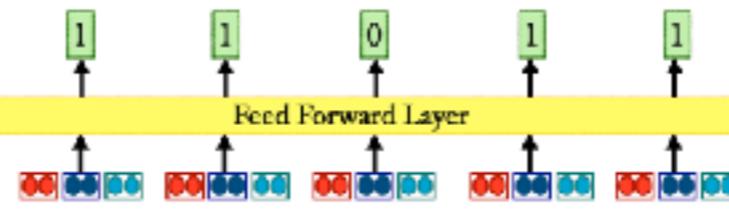
- Attentive NER tagger
 - combined with in-domain contextual representation



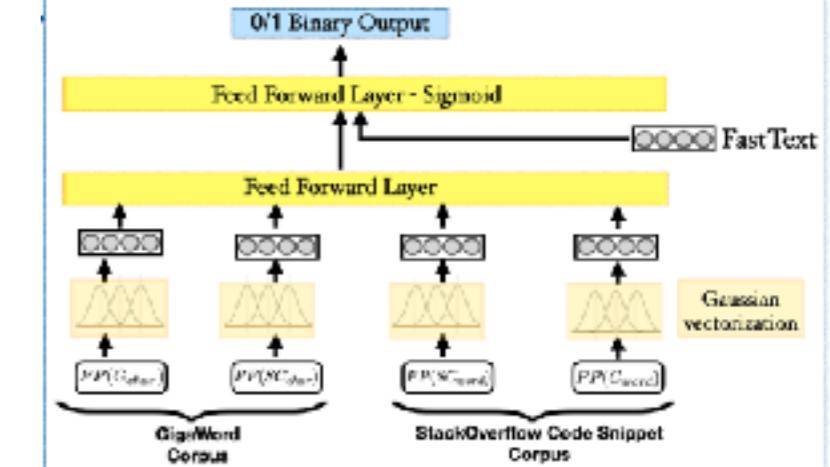
Contextual Word Representation



Entity Segmenter



Code Recognizer



Polysemy

Challenges

InLine Code



Input Text

cpp

QSort()

uses

quick

sort



Word Context



Input Text

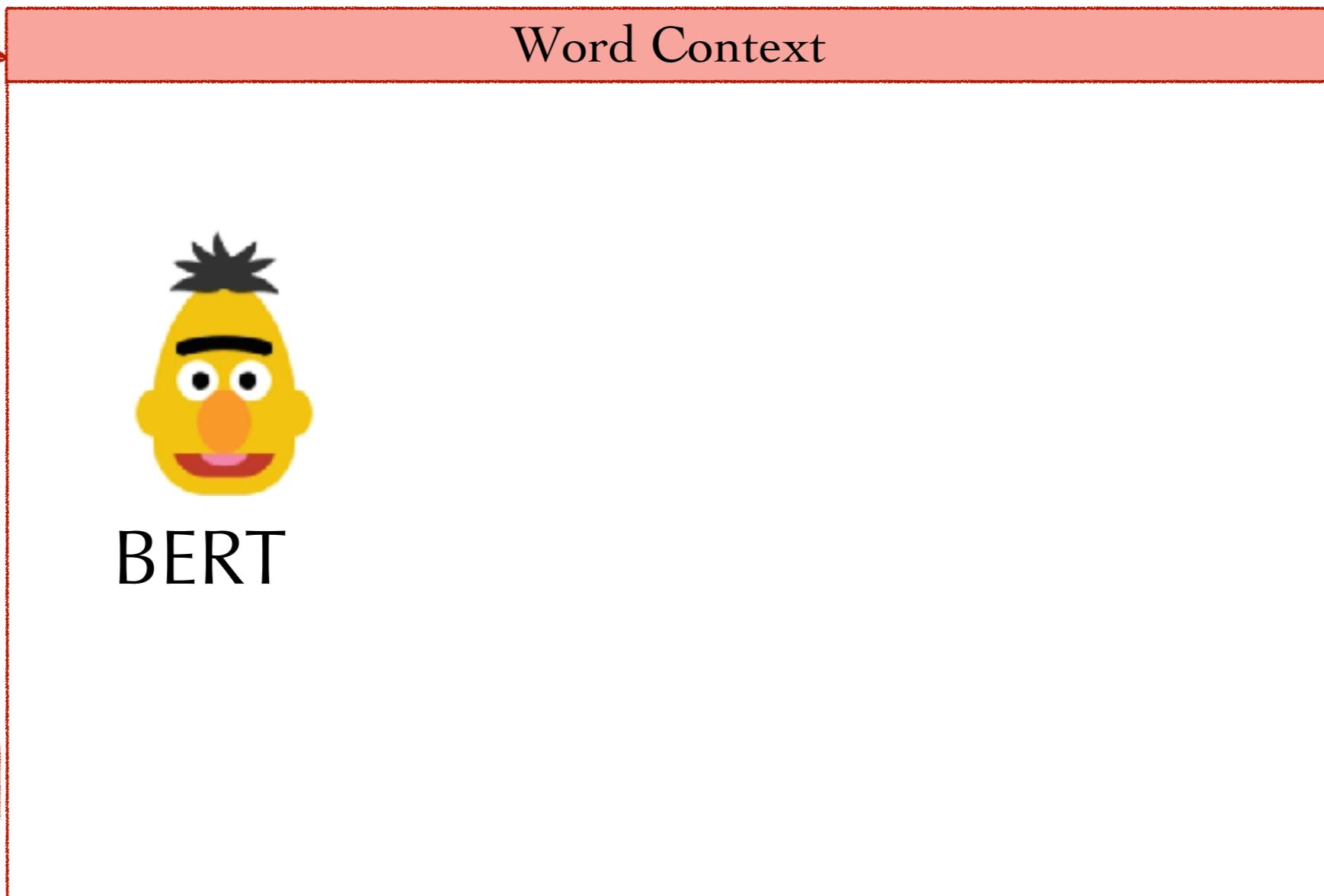
cpp

QSort()

uses

quick

sort



Word Context

Input Text

cpp

QSort()

uses

quick

sort

Word Context



BERTOverflow

Word Context

Input Text

cpp

QSort()

uses

quick

sort



Word Context



BERTOverflow

152M sentences
StackOverflow 10 year archive

Word Context

Input Text

cpp

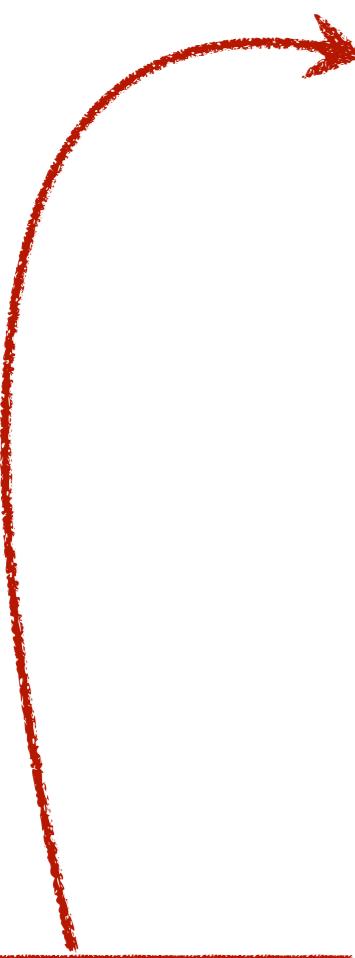
QSort()

uses

quick

sort

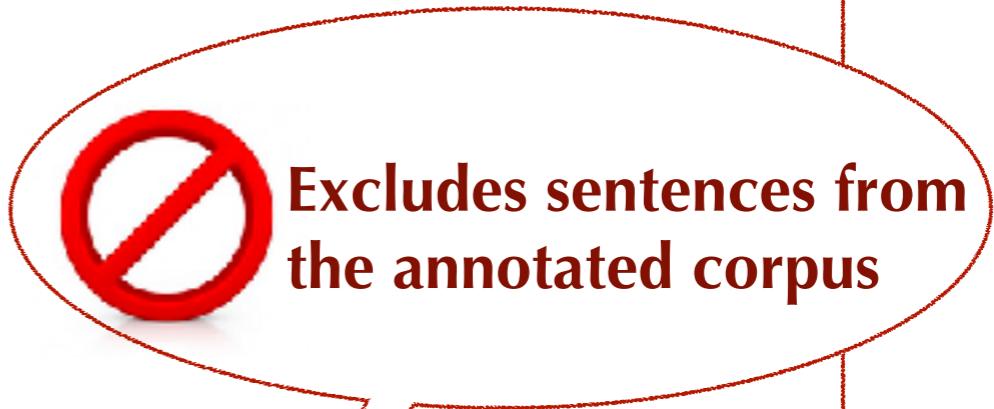




Word Context



BERTOverflow



Word Context

152M sentences

StackOverflow 10 year archive



Input Text

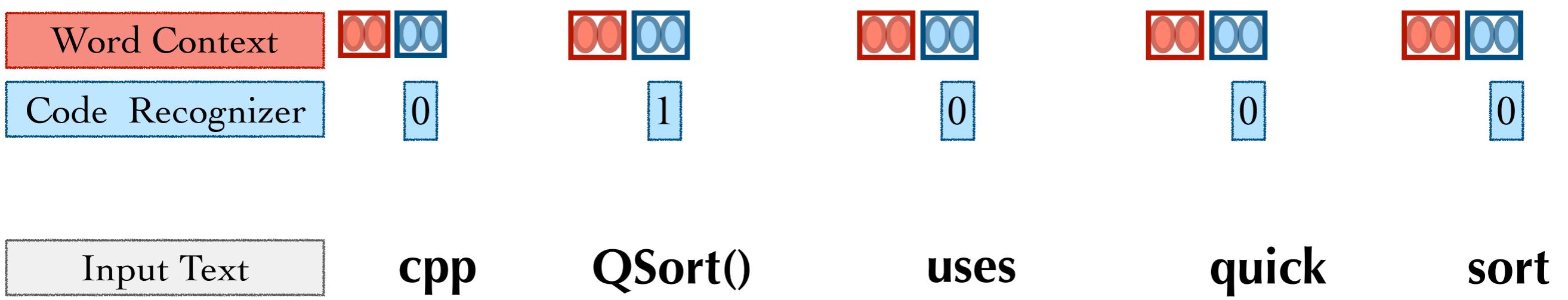
cpp

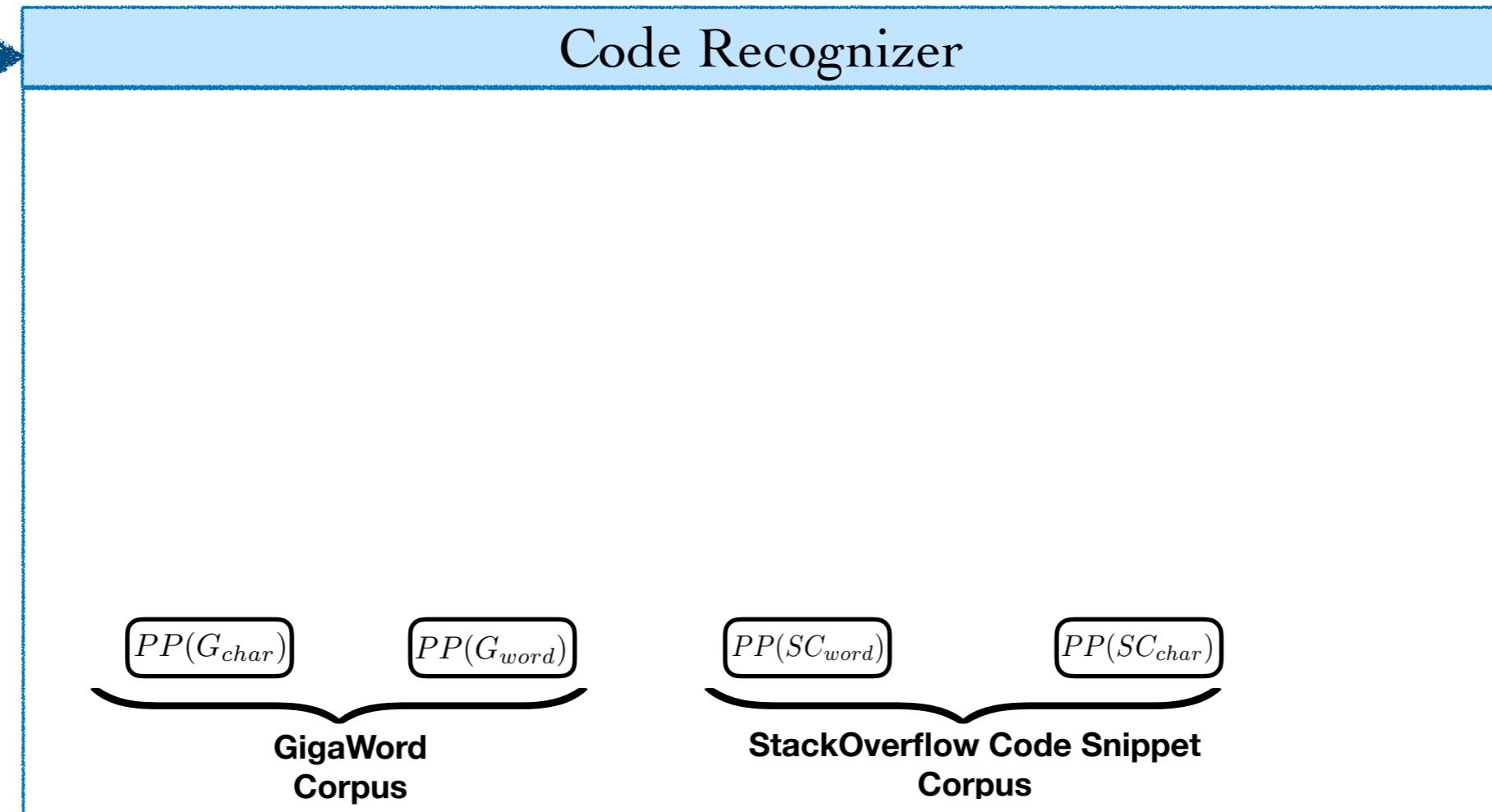
QSort()

uses

quick

sort





Input Text

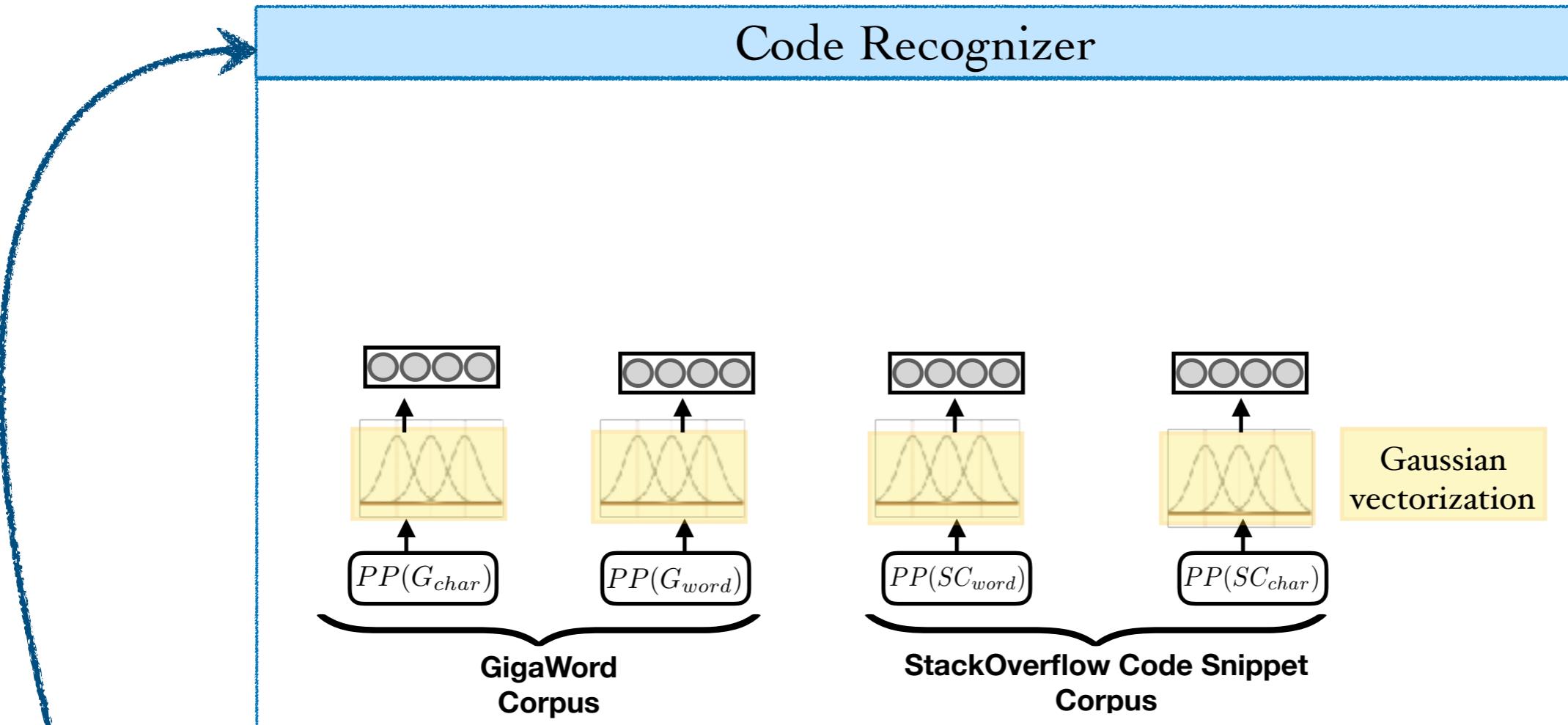
cpp

QSort()

uses

quick

sort



Input Text

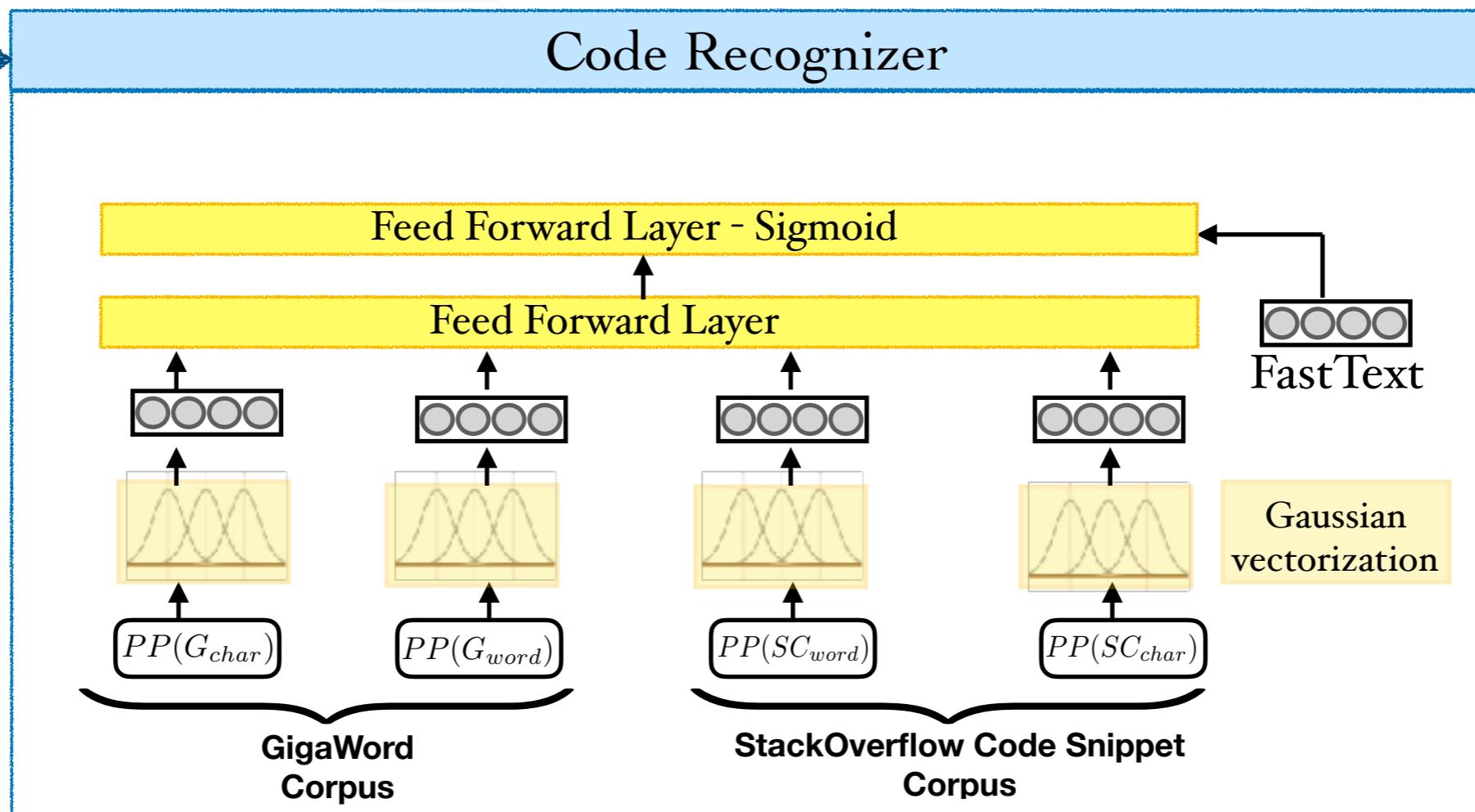
cpp

QSort()

uses

quick

sort



Word Context

Code Recognizer

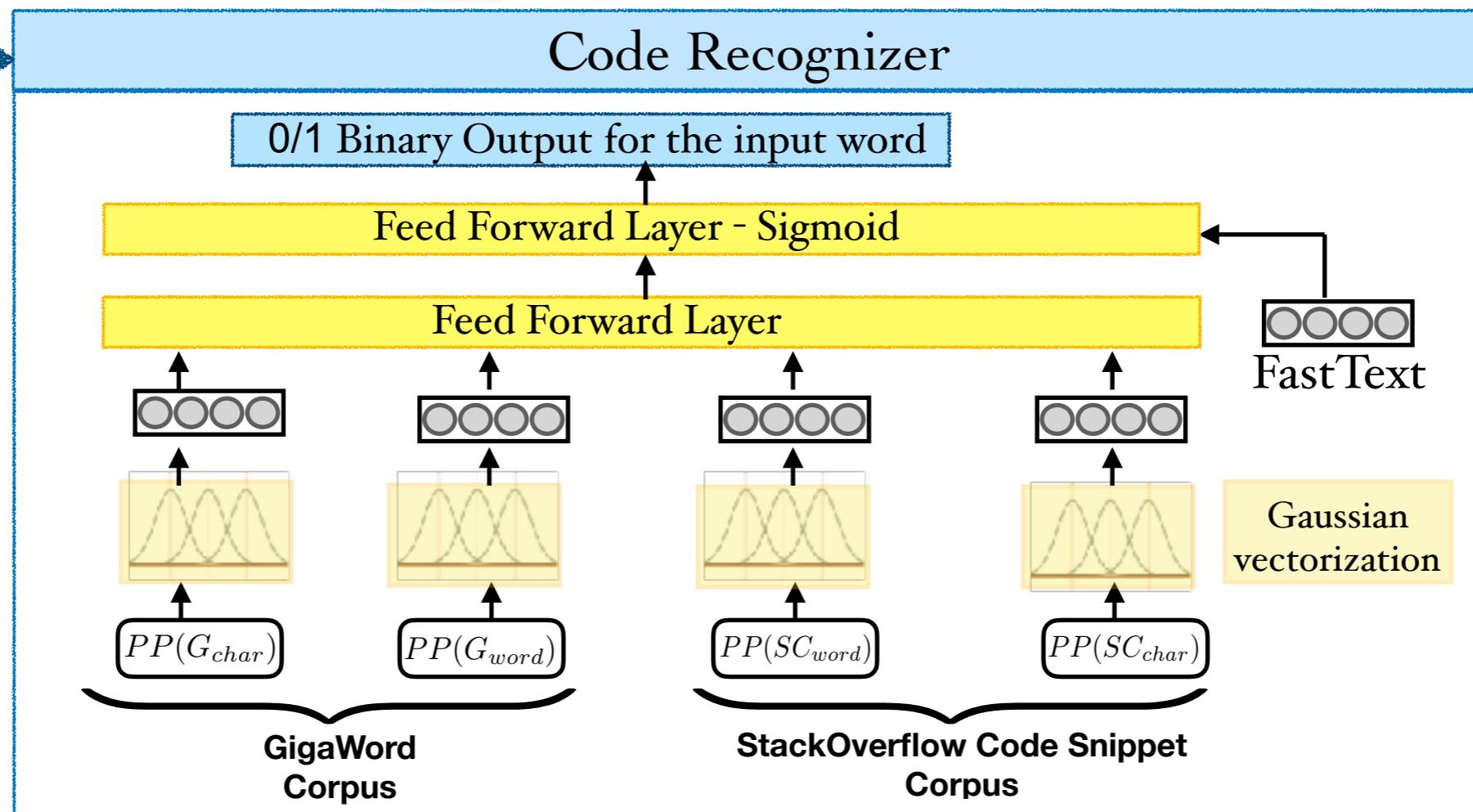
cpp

QSort()

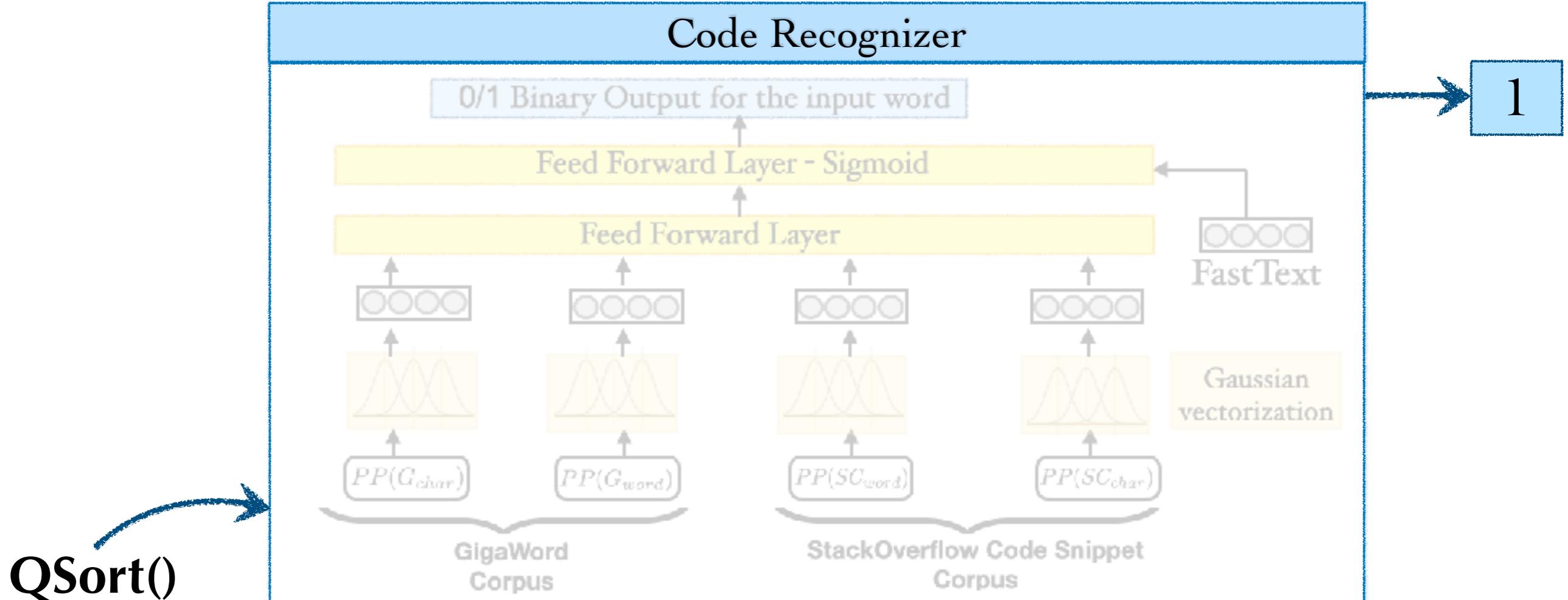
uses

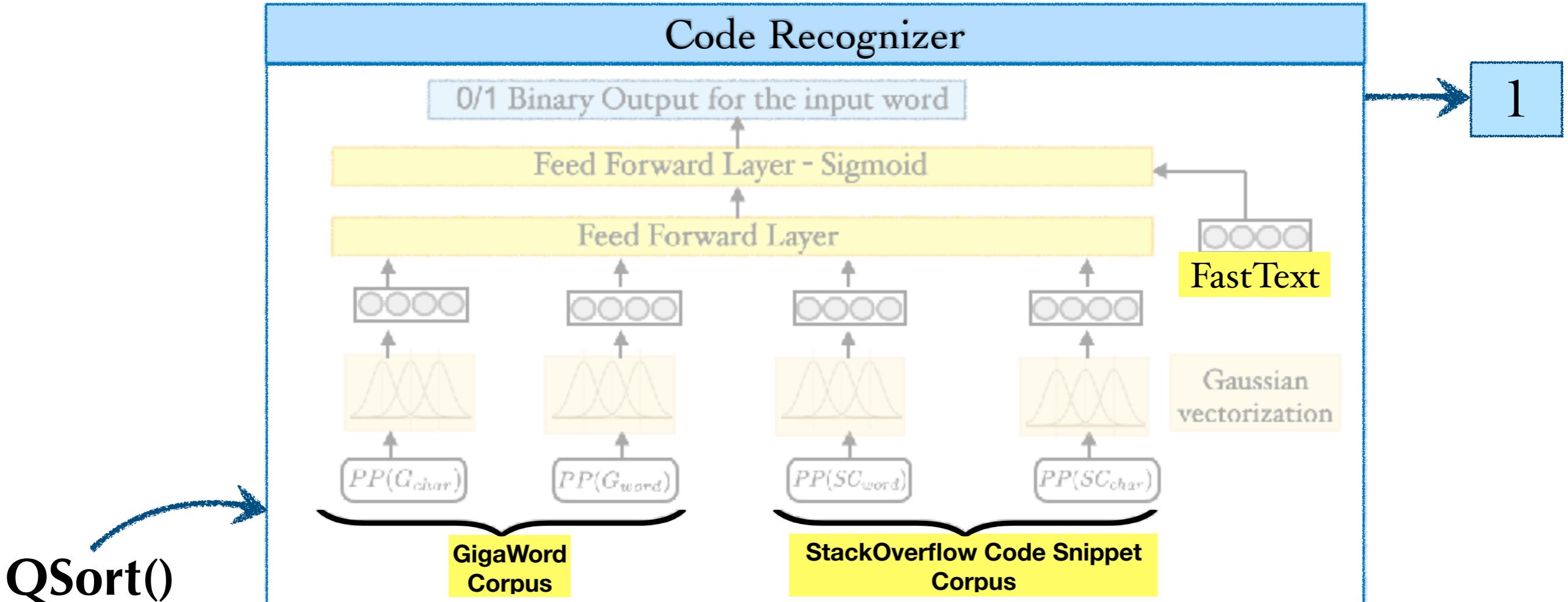
quick

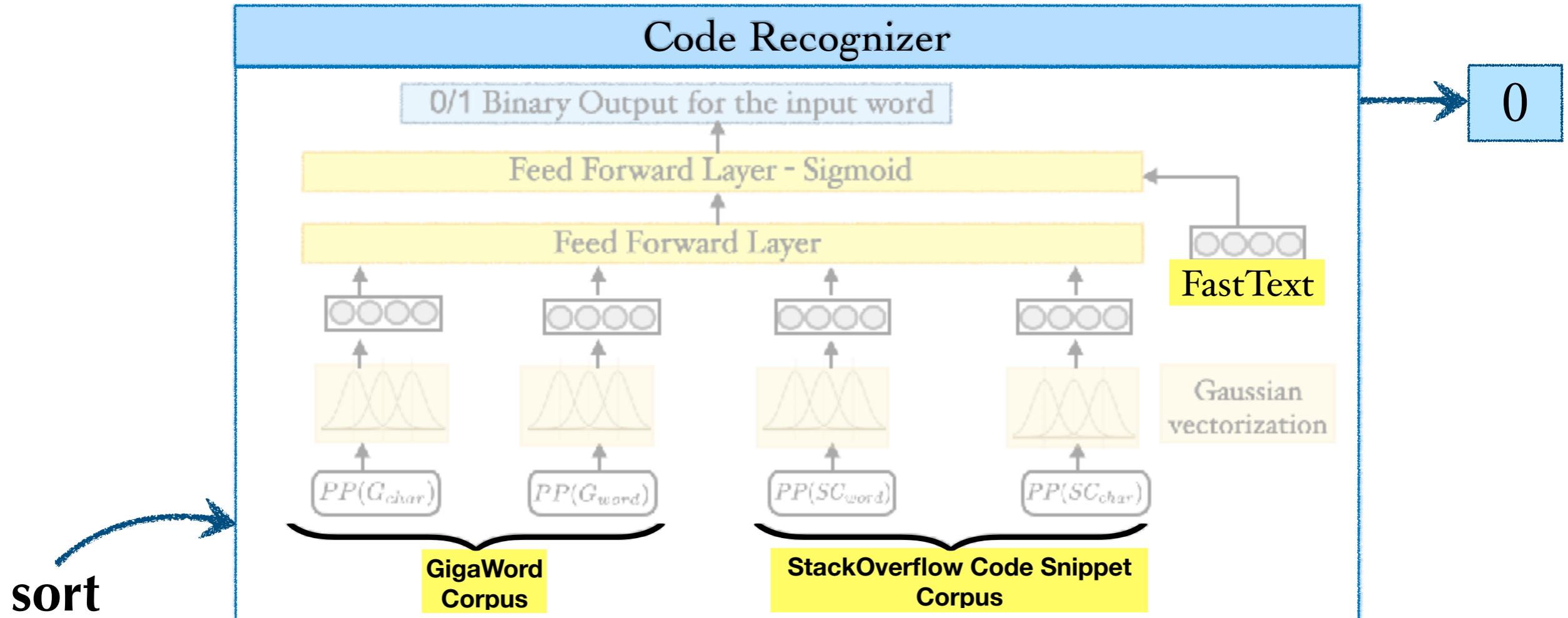
sort

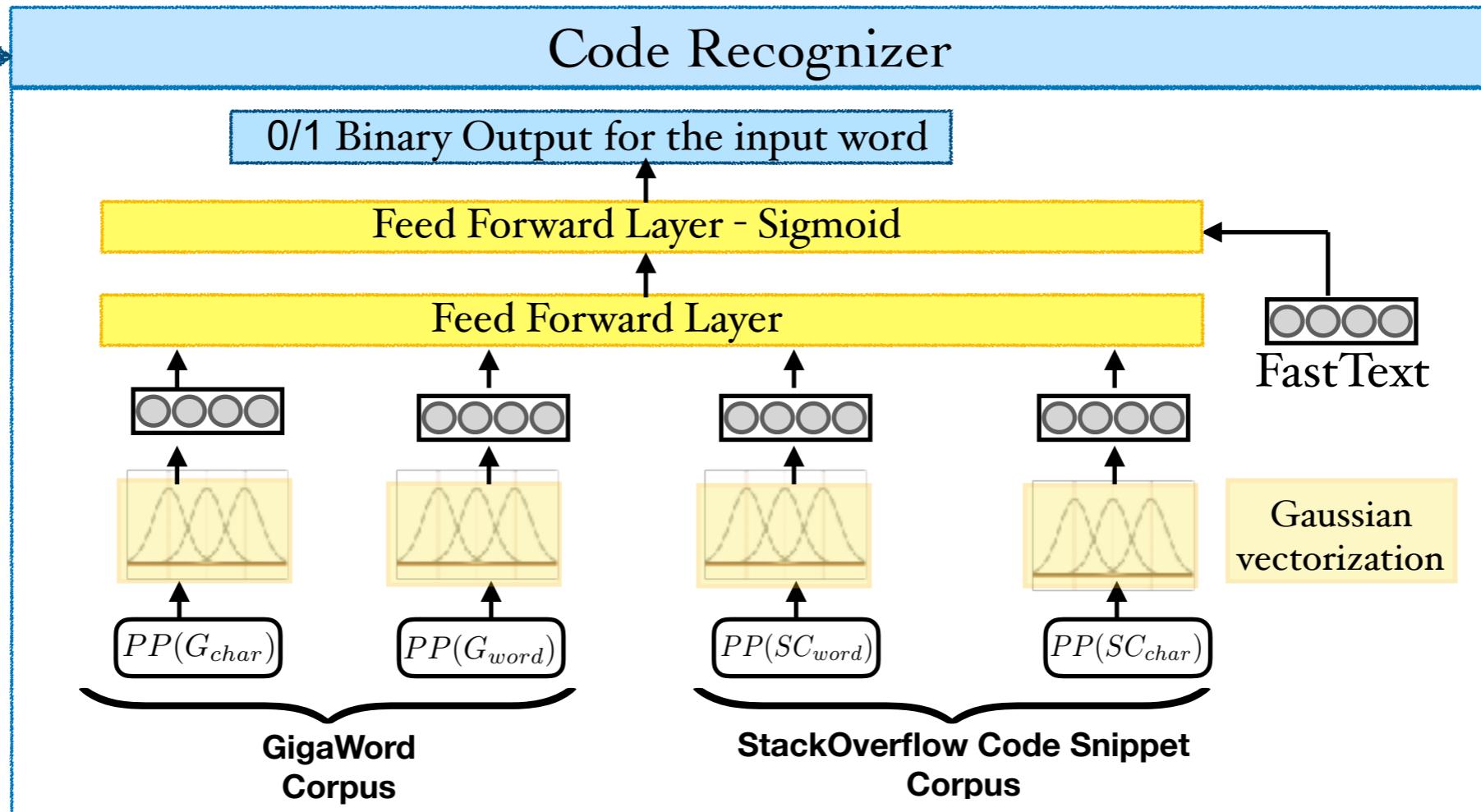


Input Text cpp QSort() uses quick sort









Input Text

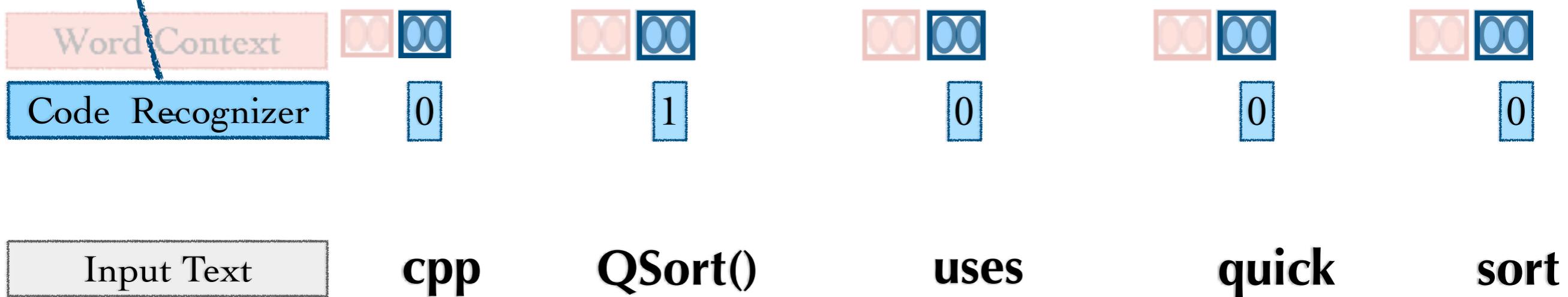
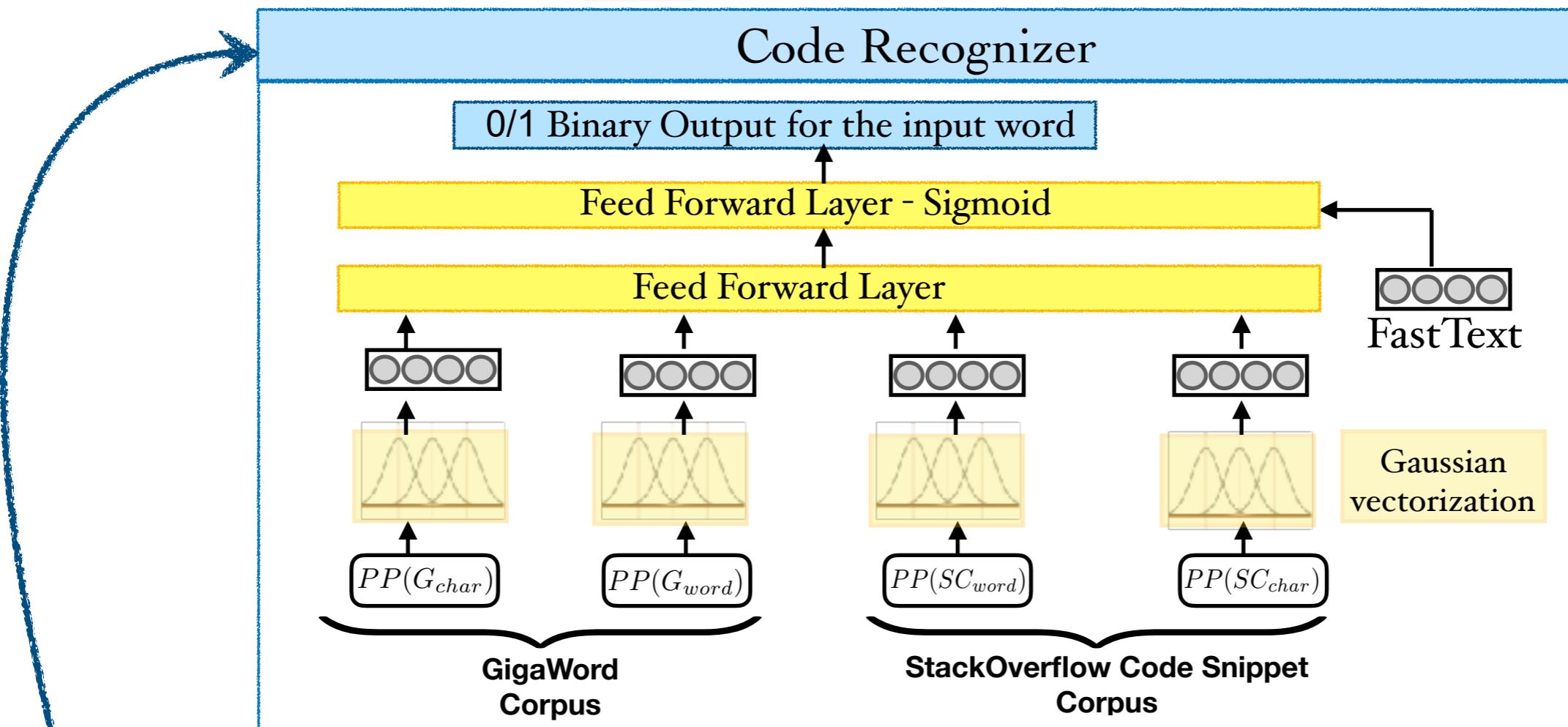
cpp

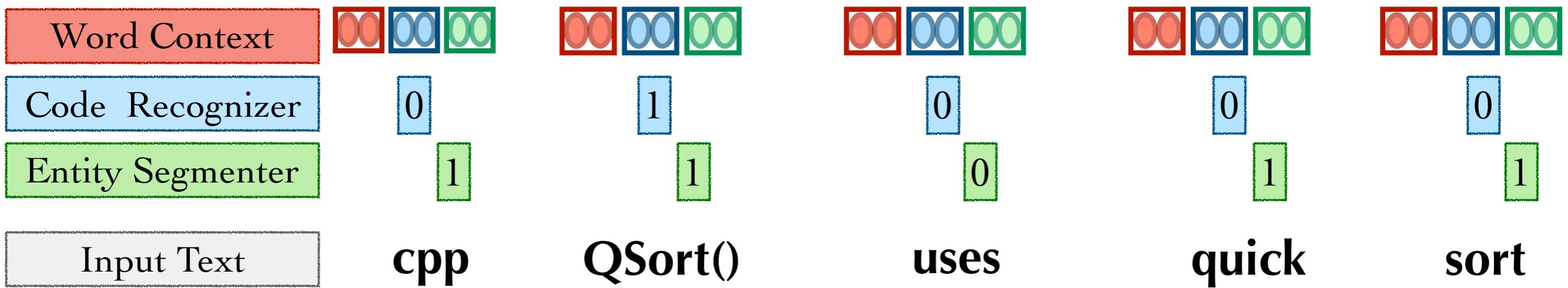
QSort()

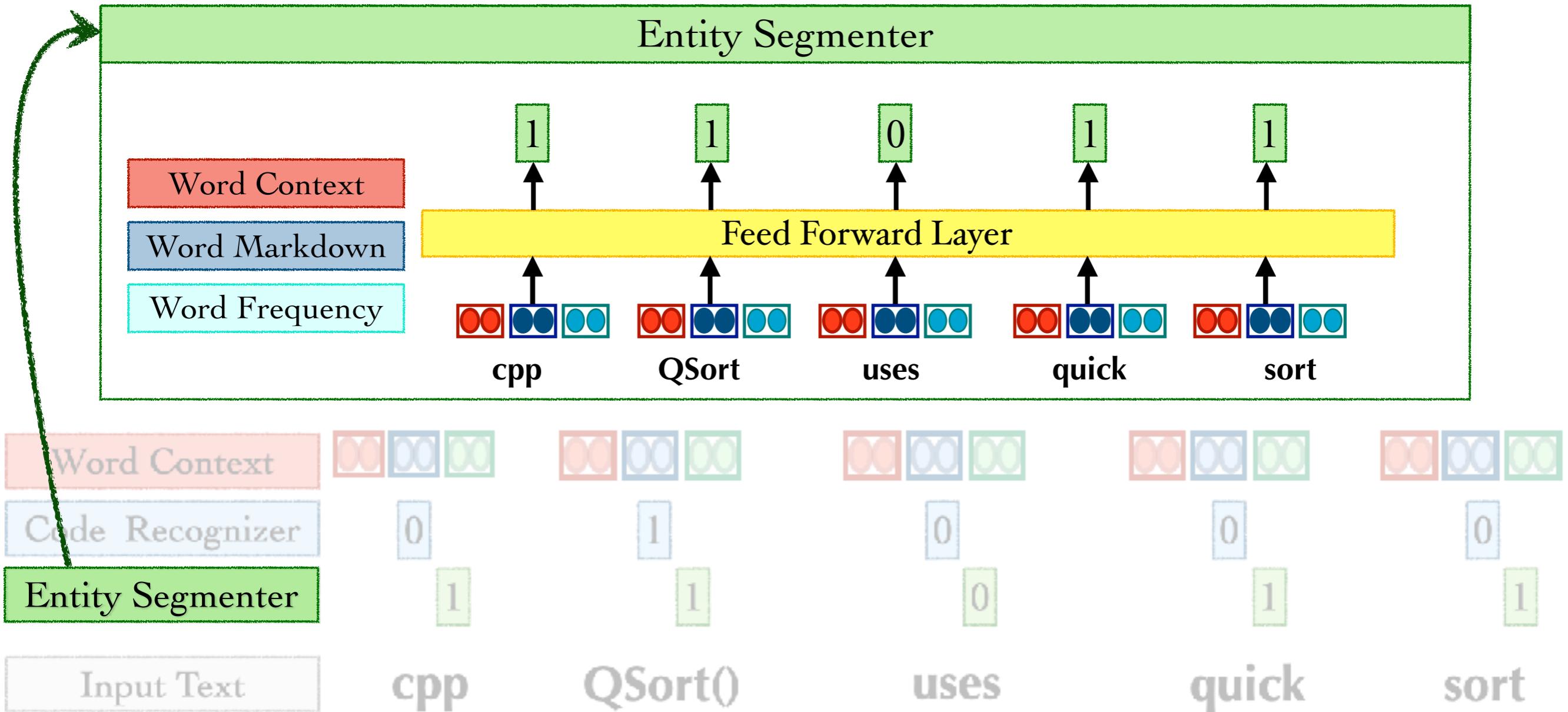
uses

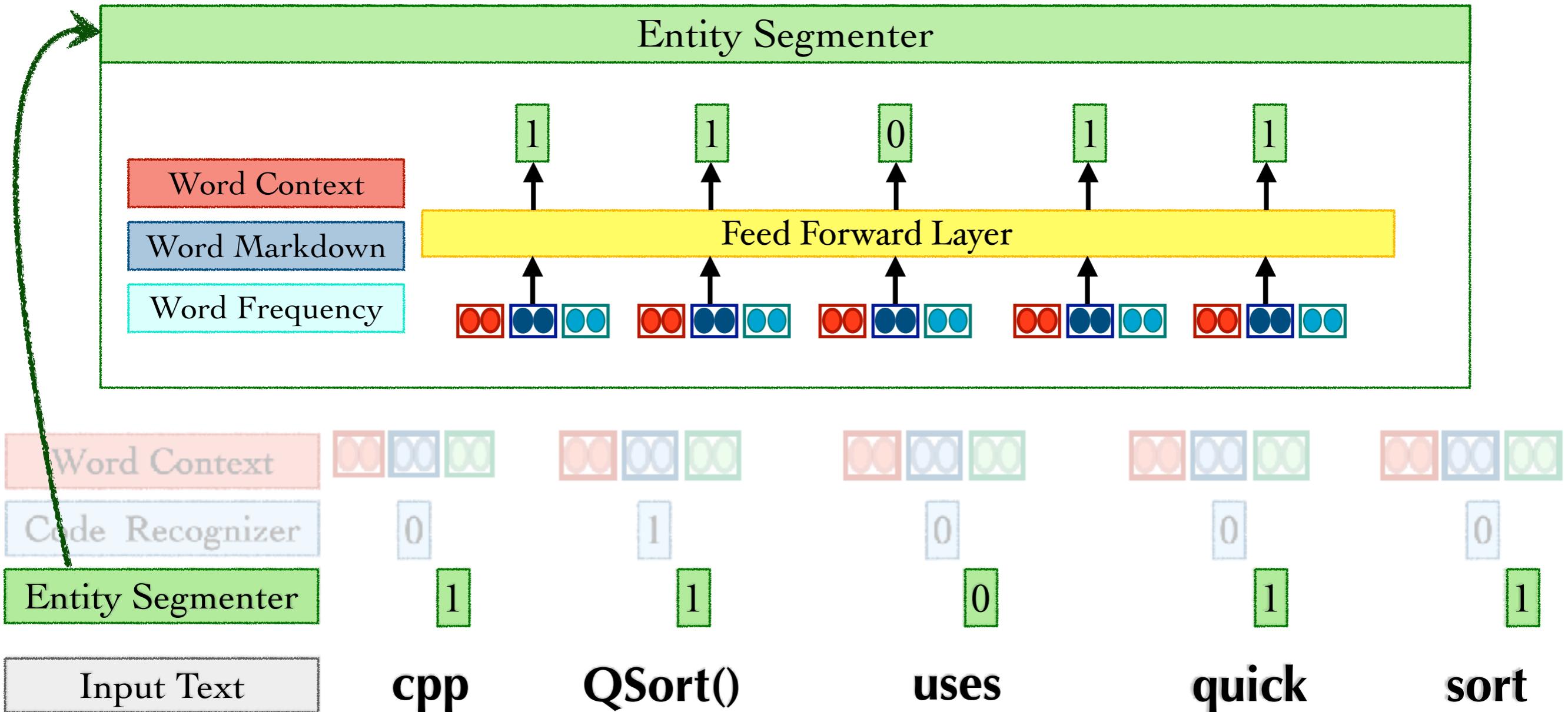
quick

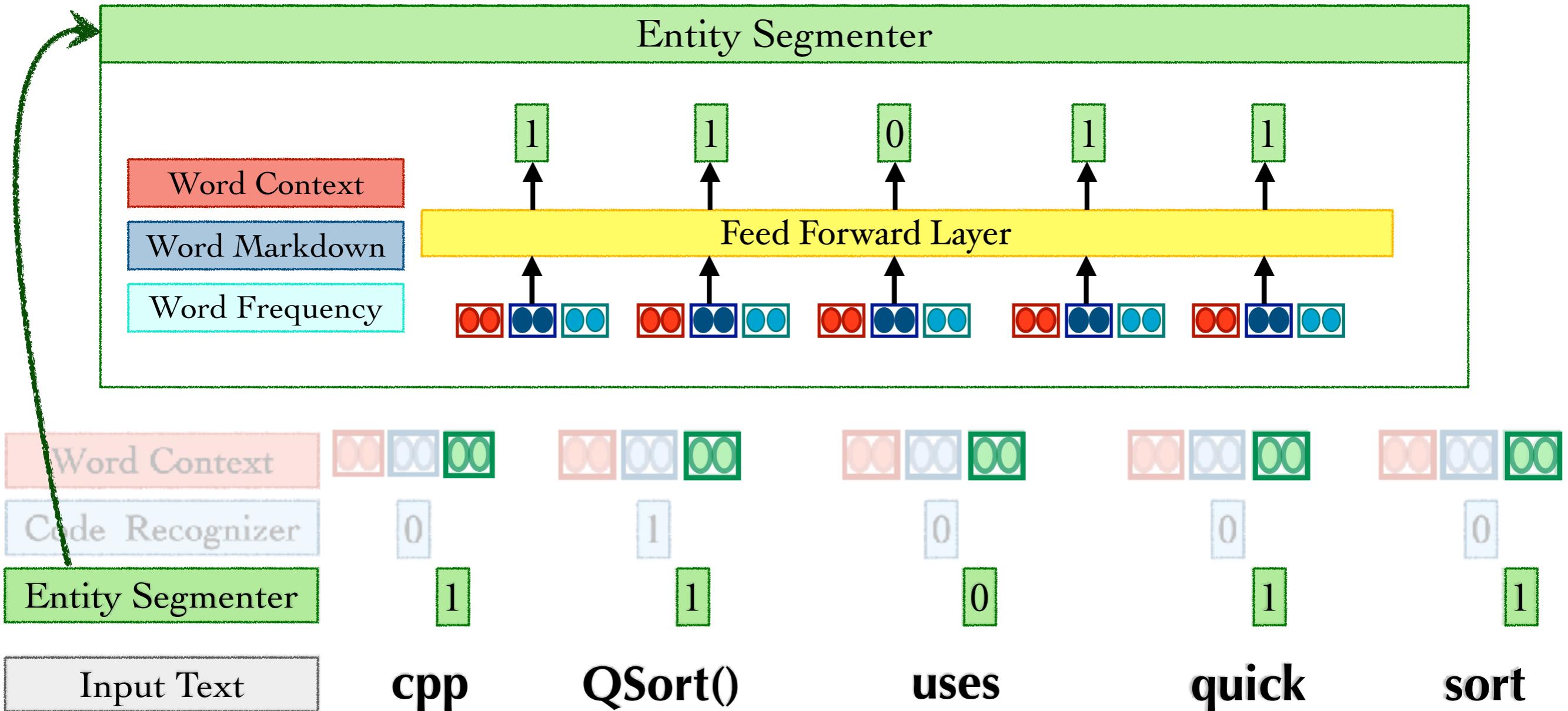
sort

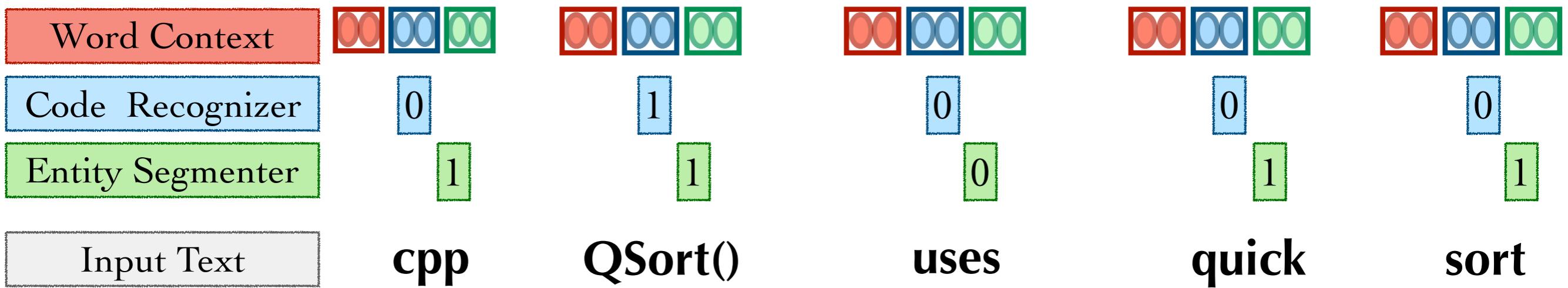


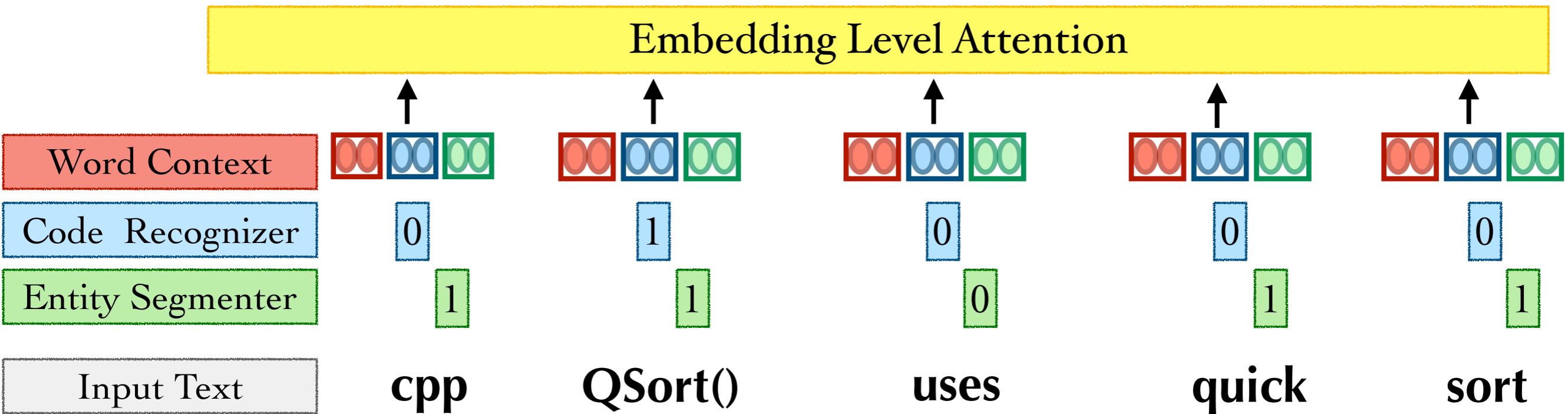












Embedding Level Attention

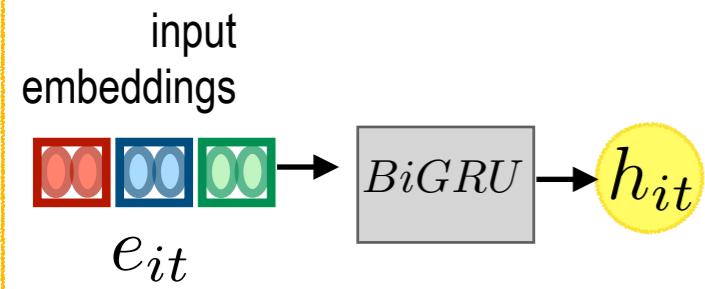
input
embeddings



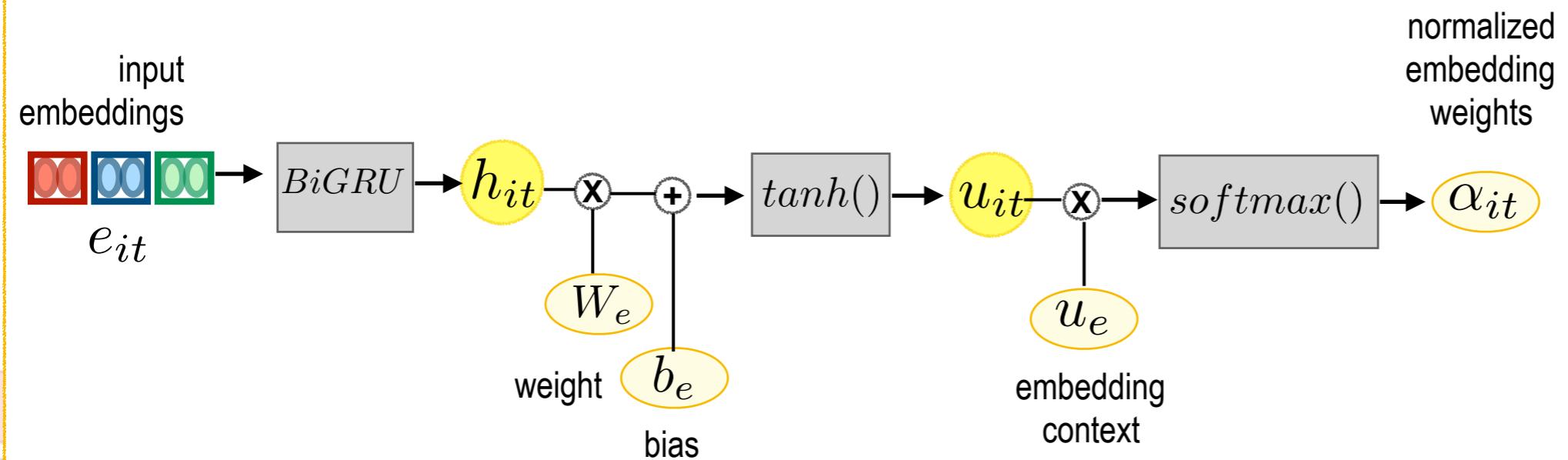
e_{it}



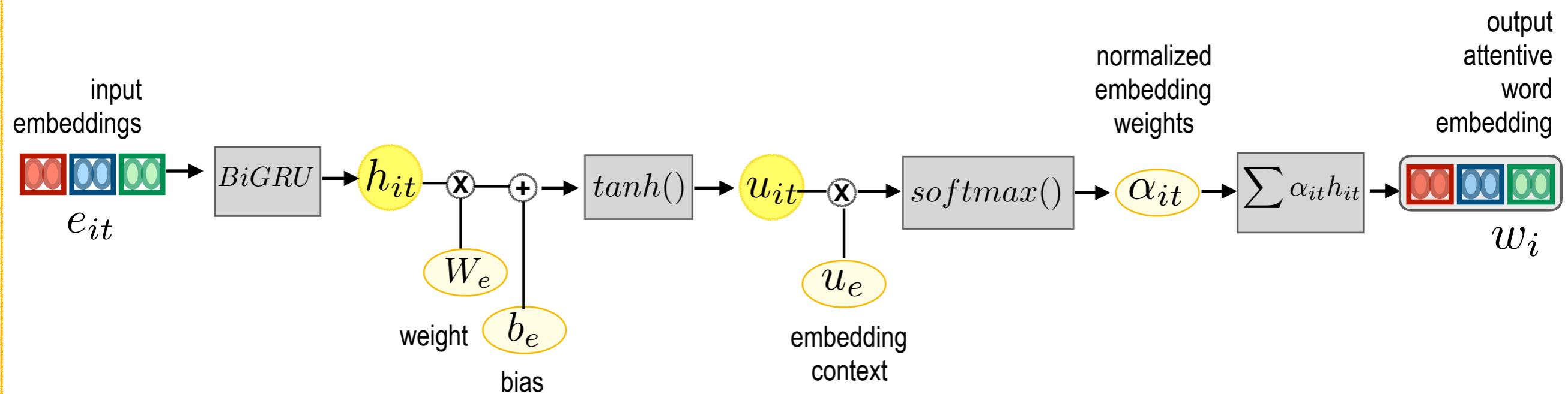
Embedding Level Attention



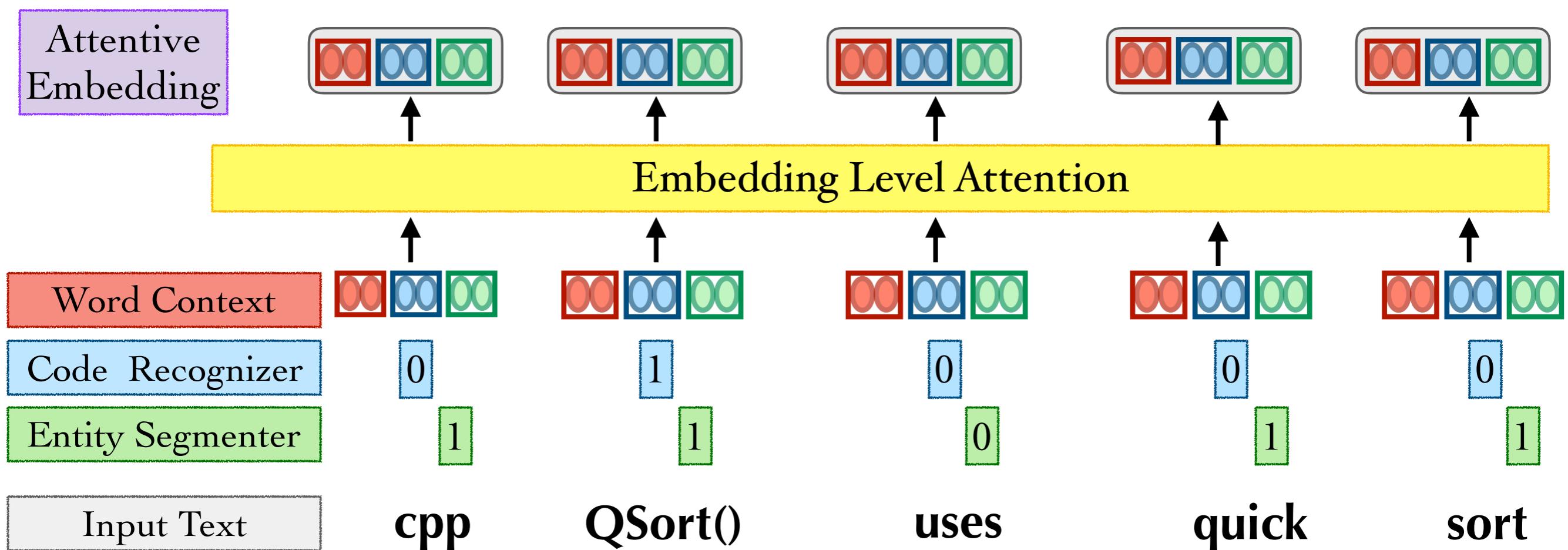
Embedding Level Attention



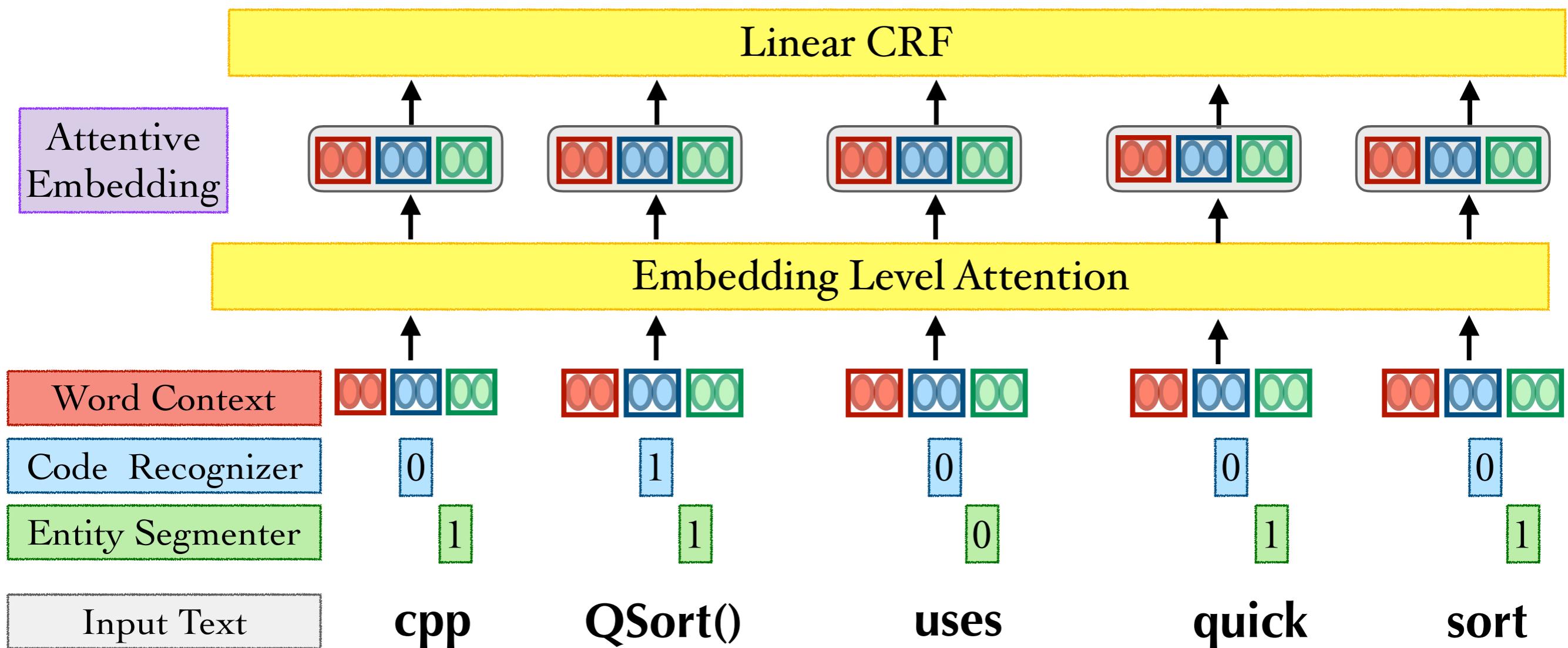
Embedding Level Attention



SoftNER

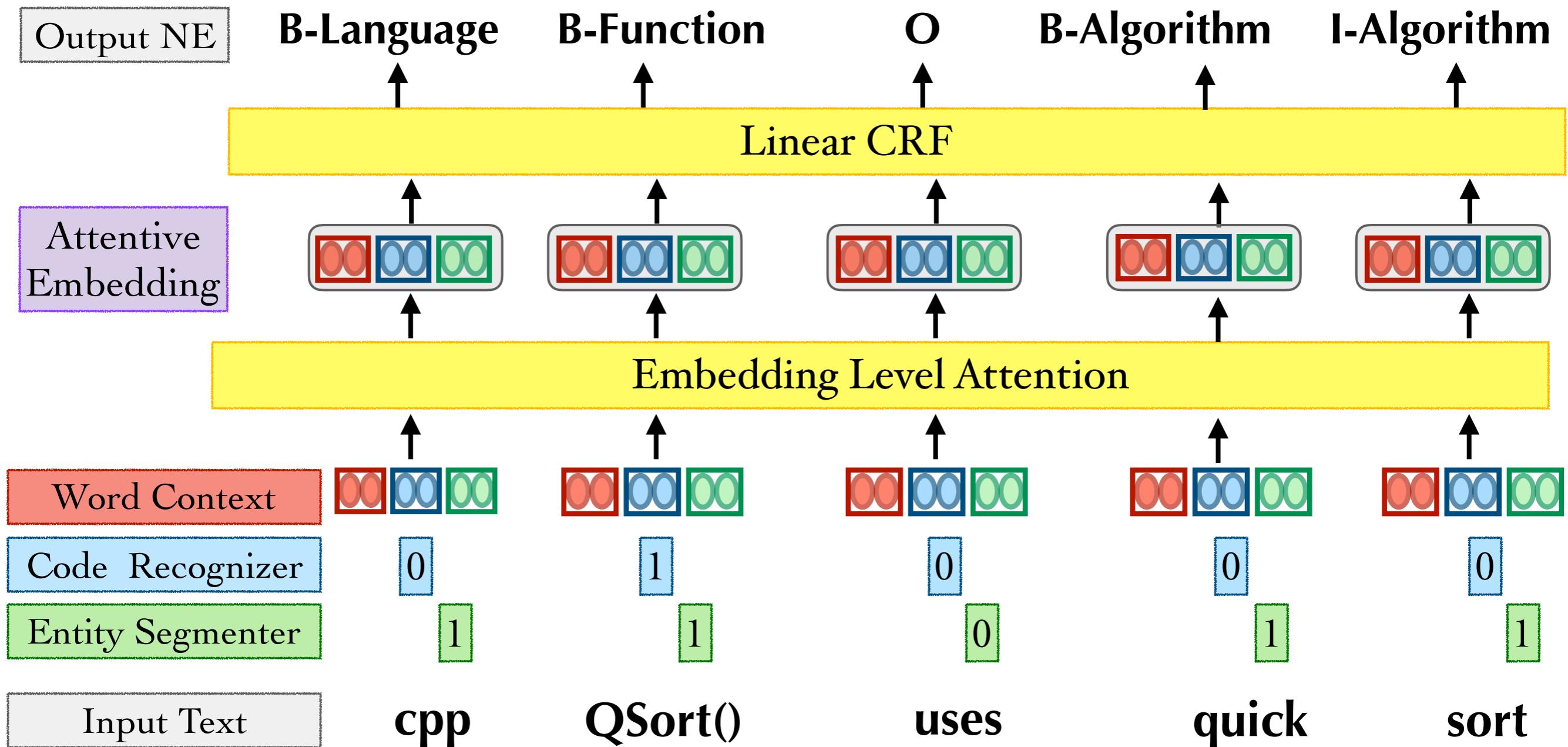


SoftNER



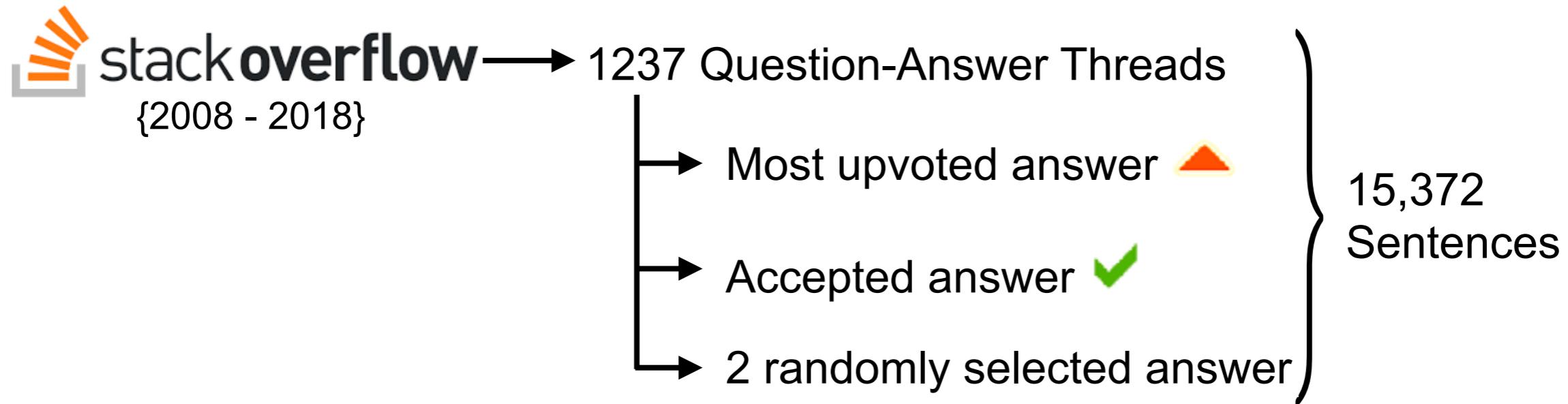


SoftNER





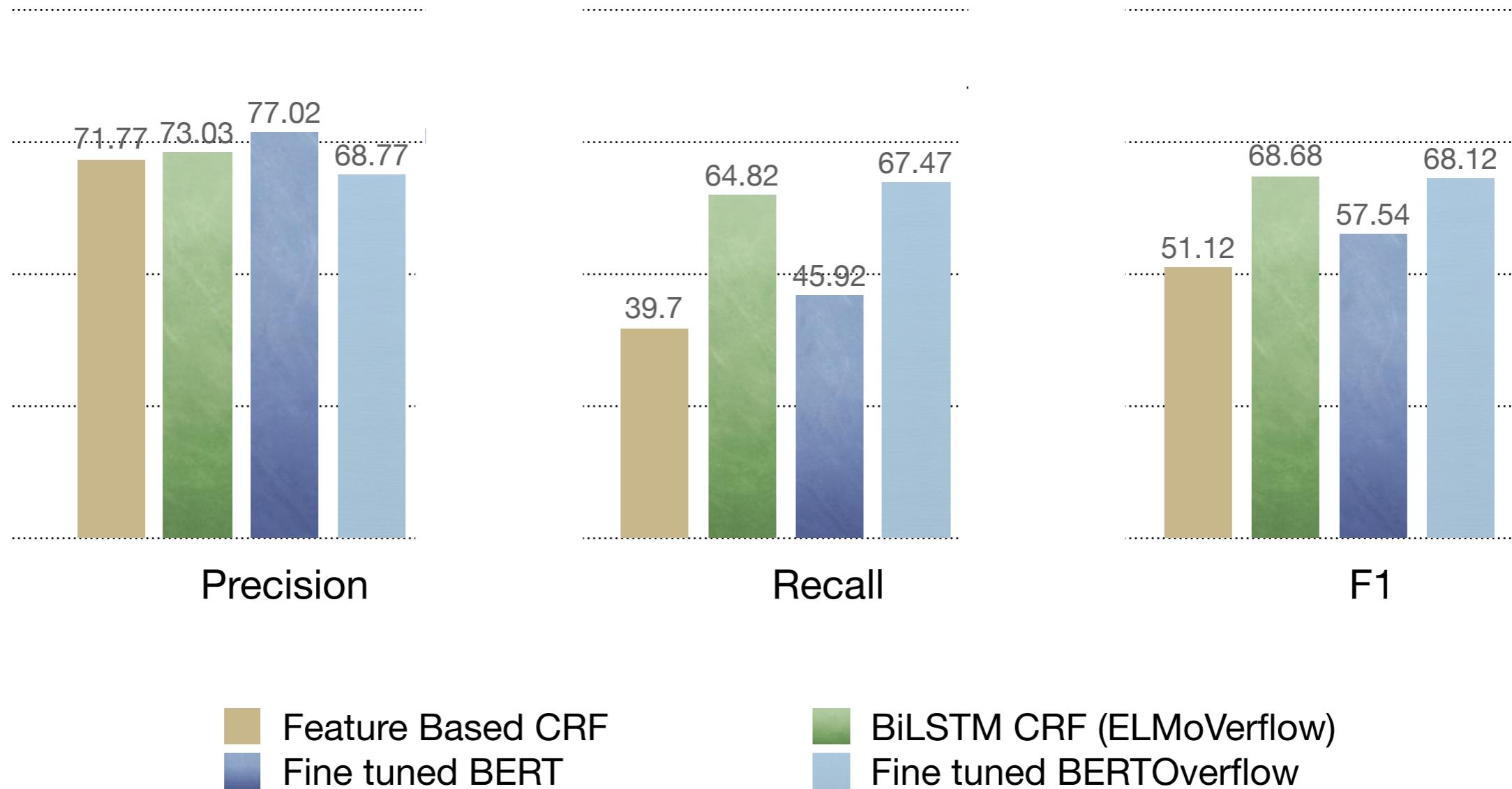
Annotated Corpus



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



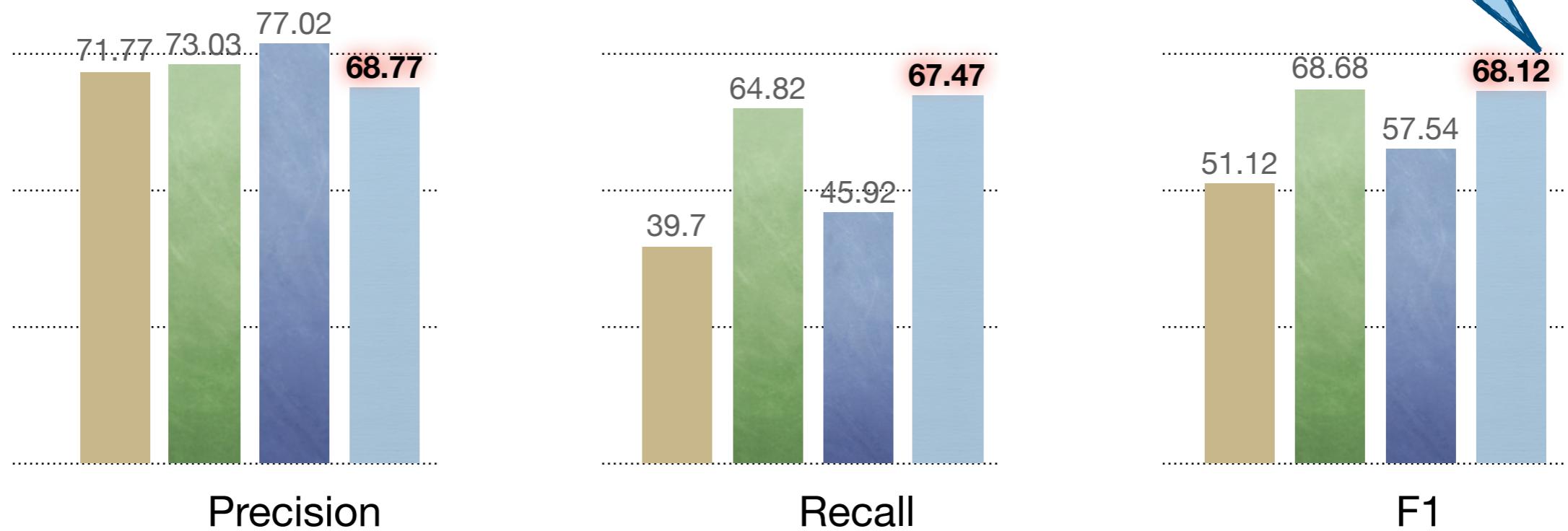
Performance





StackOverflow

Performance



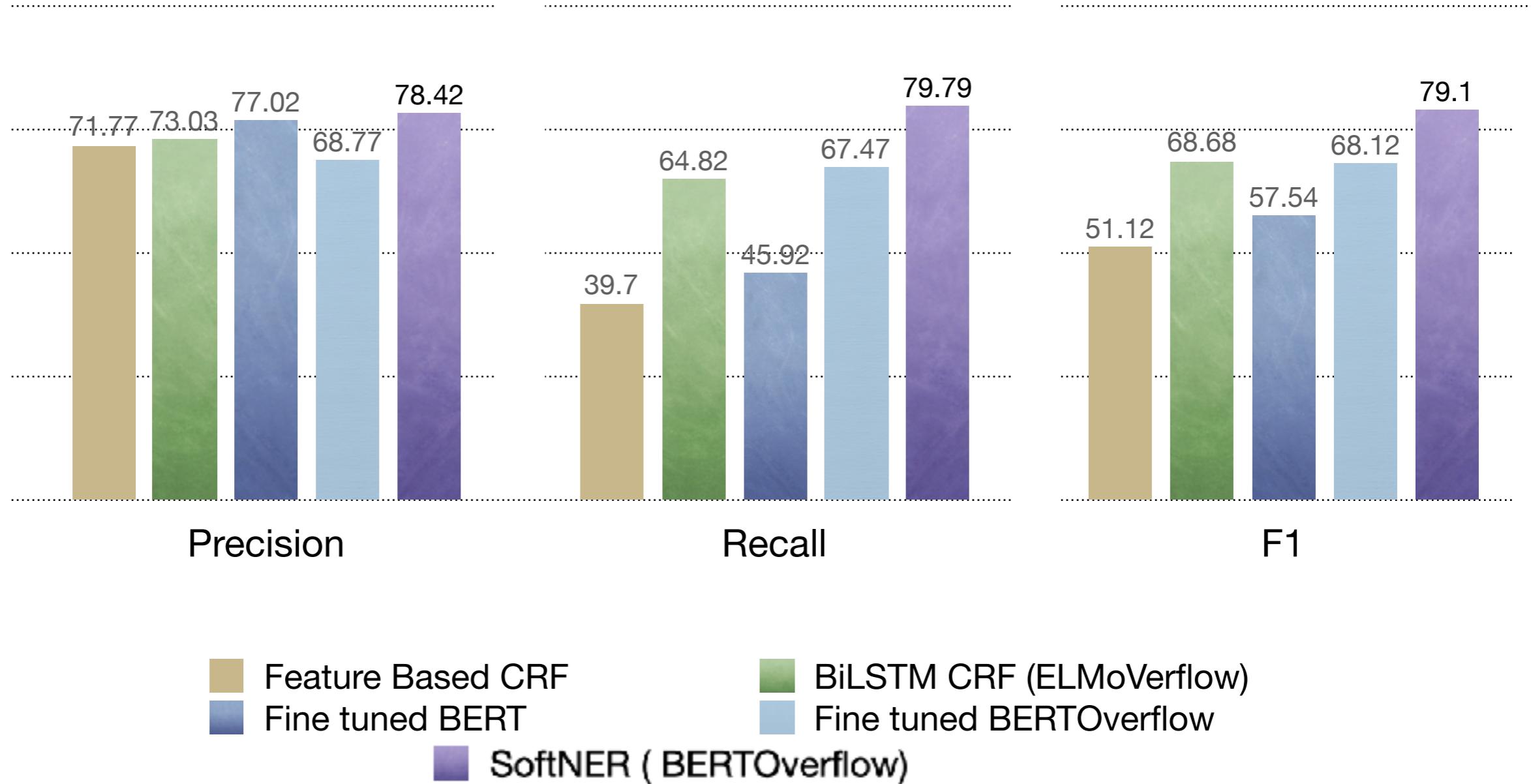
Feature Based CRF
Fine tuned BERT

BiLSTM CRF (ELMoOverflow)
Fine tuned BERTOverflow

10.58 F1 ↑ over BERT_{off-the-self}

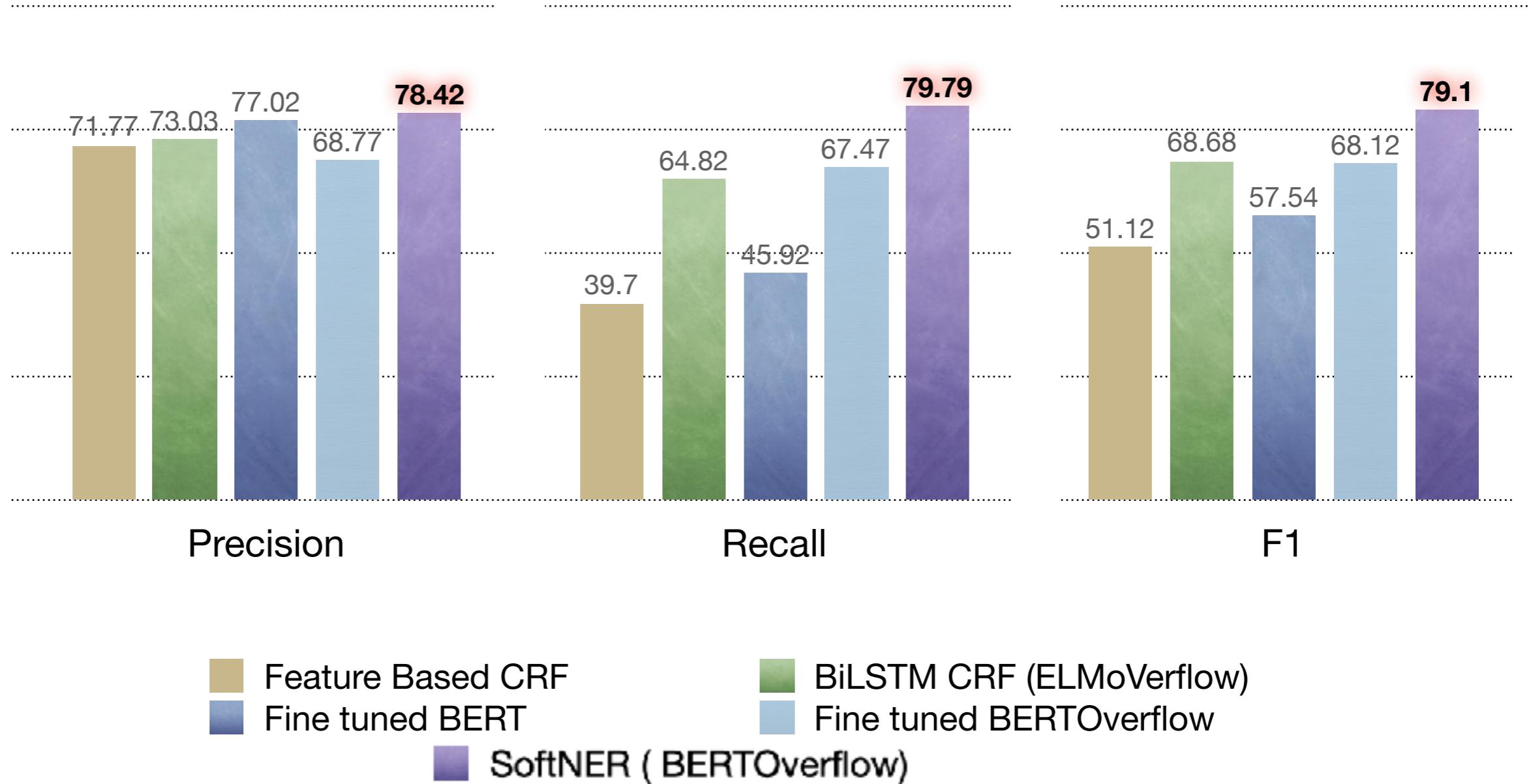


Performance





Performance

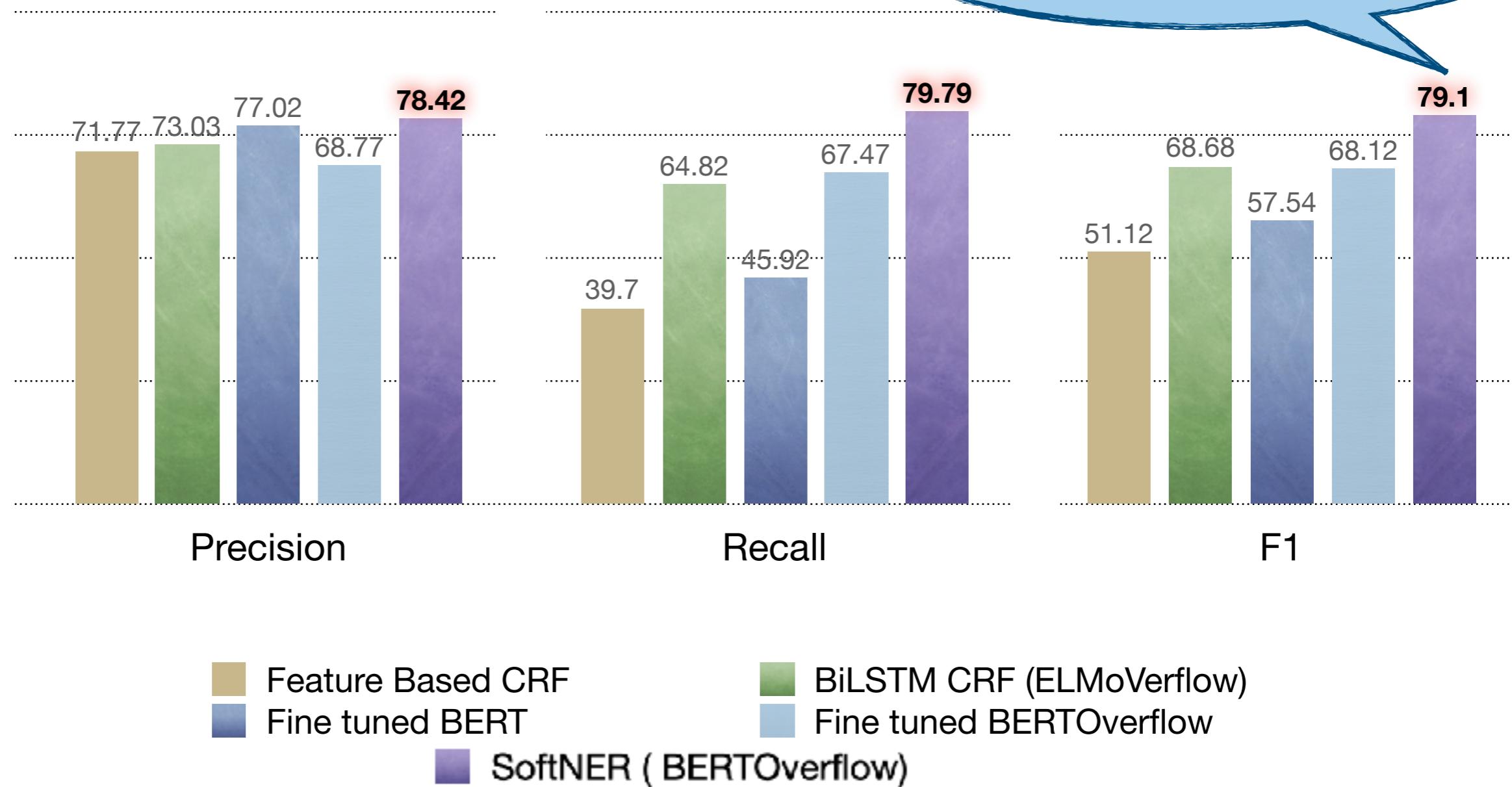




StackOverflow

Performance

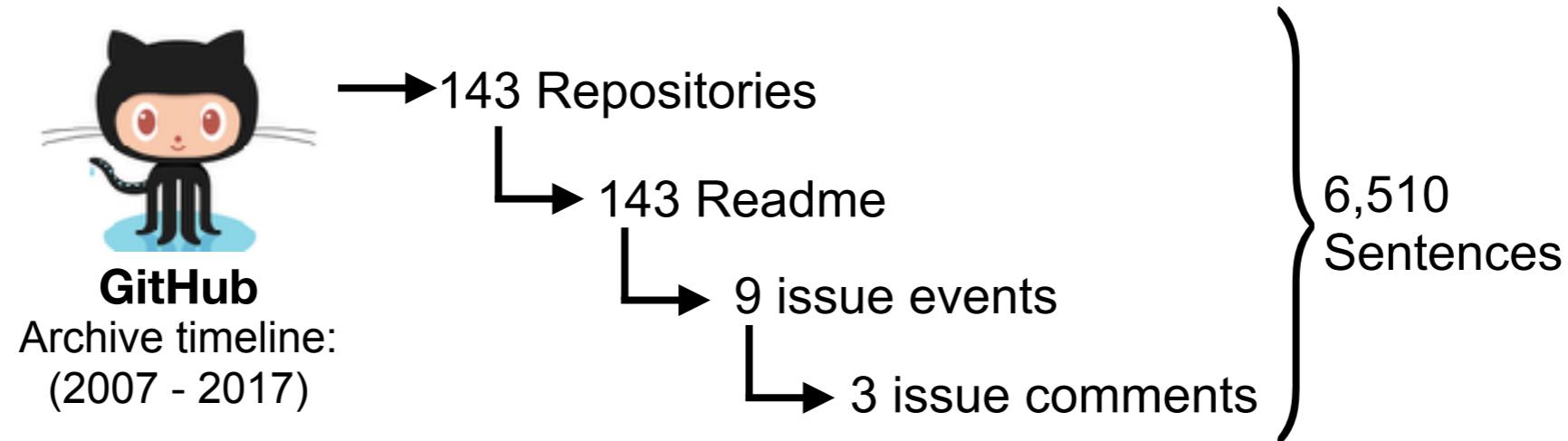
10.98 F1 ↑ over fine-tuned
BERTOverflow





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

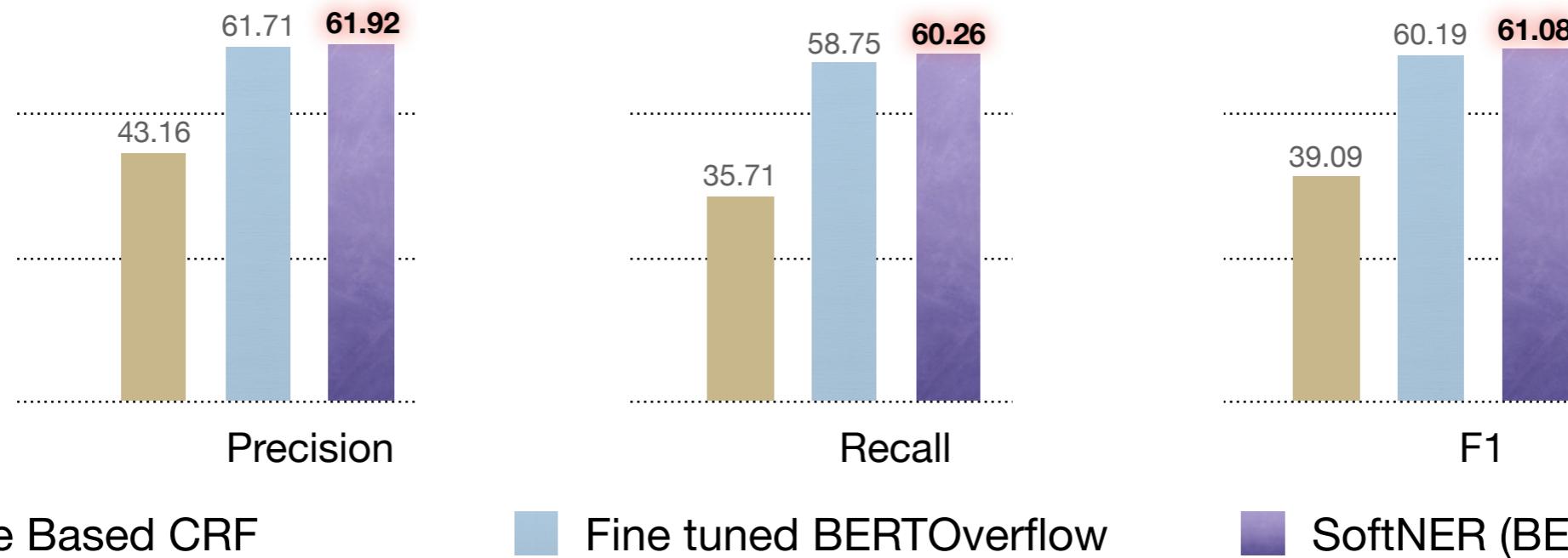
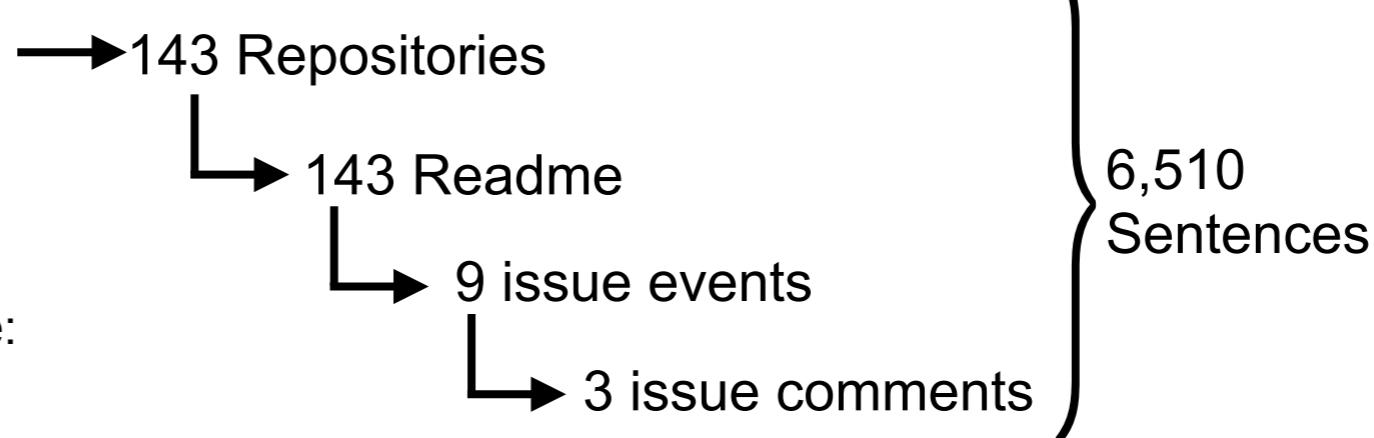
Performance on GitHub Test Set



Performance on GitHub Test Set



Archive timeline:
(2007 - 2017)



Trained on 9,315 manually annotated StackOverflow Sentences

Performance on GitHub Test Set



Archive timeline:
(2007 - 2017)

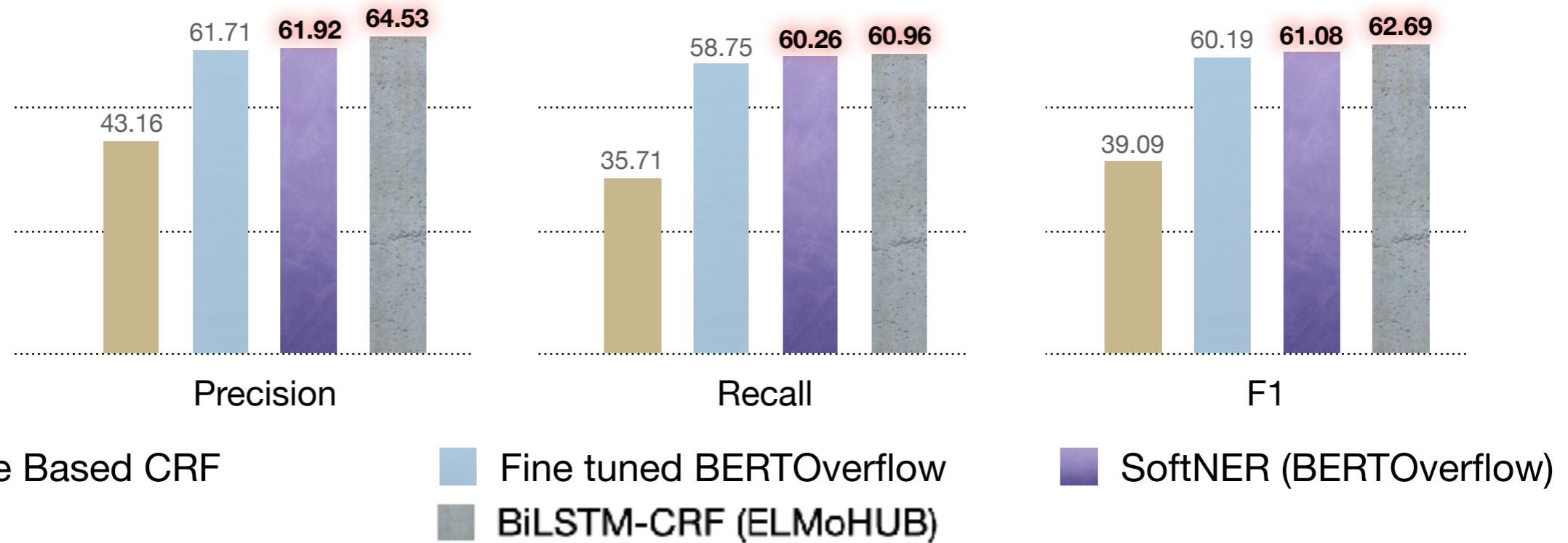
→ 143 Repositories

↳ 143 Readme

↳ 9 issue events

↳ 3 issue comments

} 6,510
Sentences



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
-



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



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, ELMoOverflow, GLoOverflow



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

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, ELMoOverflow, 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



StackOverflow

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



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, ELMoOverflow, GLoOverflow



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



StackOverflow

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

