Carnegie Mellon University
School of Computer Science

Learning to Mine Aligned Code and Natural Language Pairs from Stack Overflow

Pengcheng Yin*   Bowen Deng*   Edgar Chen   Bogdan Vasilescu   Graham Neubig

Carnegie Mellon University

Language Technologies Institute

NEULAB

isr institute for SOFTWARE RESEARCH

STRUDEL
SOCIO-TECHNICAL RESEARCH USING DATA EXCAVATION LAB

Background

- When Natural Language Processing (NLP) meets Software Engineering…

Code Summarization[1]

```
SELECT Max(marks) FROM records
WHERE marks <
    (SELECT Max(marks) FROM
records);
```

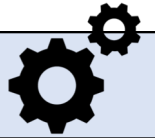*Get the second largest value of a column*

Code Retrieval[2]

load csv file to pandas

```
pandas.read_csv('example.csv')
```

Code Generation[3]

sort my_list in descending order

```
sort(my_list,reverse=True)
```

- These tasks, mostly powered by **data-driven models**, heavily rely on parallel training (and evaluation) corpora of source code and natural language in **high quality** and **large amount**

Language Technologies Institute

[1] Iyer et al., ACL 2016
[2] Zhang et al., FSE 2016
[3] Yin and Neubig., ACL 2017, ACL 2018

Collecting Intent/Snippet Pairs from SO

- Such data-driven models require parallel data of natural language **intents** and source code **snippets** and in high volume and high quality
  - **Intent** natural language description of what a programmer would like to do
  - **Snippet** a piece of source code that implements the intent

Intent get the maximum value of a column

Snippet SELECT MAX(marks) from records

Intent read a csv file into pandas

Snippet pandas.read_csv('example.csv')

Intent sort my_list in descending order

Snippet sort(my_list, reverse=True)

Can we collect such data from

stackoverflow ?

Heuristic approaches [Wong et al., 2013; Iyer et al., 2016]?
- Select **all** code blocks
- Select **all** code blocks in **accepted answers**

Language
Technologies
Institute

Are these Heuristic Approaches Good Enough?

Intent

Removing duplicates in lists

▲
566
▼

Pretty much I need to write a program to check if a list has any duplicates and if it does it removes them and returns a new list with the items that werent duplicated/removed. This is what I have but to be honest I do not know what to do.

```
def remove_duplicates():
    t = ['a', 'b', 'c', 'd']
```

▲
1000
▼

The common approach to get a unique collection of items is to use a `set`. Sets are *unordered* collections of *distinct* objects. To create a set from any iterable, you can simply pass it to the built-in `set()` function. If you later need a real list again, you can similarly pass the set to the `list()` function.

The following example should cover whatever you are trying to do:

✔

```
>>> t = [1, 2, 3, 1, 2, 5, 6, 7, 8]
>>> t
[1, 2, 3, 1, 2, 5, 6, 7, 8]
>>> list(set(t))
[1, 2, 3, 5, 6, 7, 8]
>>> s = [1, 2, 3]
>>> list(set(t) – set(s))
[8, 5, 6, 7]
```

Contextual Information
Actual Snippet
Auxiliary info.
Irrelevant Code

As you can see from the example result, the original order is not maintained. As mentioned above, sets themselves are unordered collections, so the order is lost. When converting a set back to a list, an arbitrary order is created.

However, acquiring such data from Stack Overflow posts may not be that straight-forward ....

1. Contextual Information package import statements, variable definition
2. Auxiliary information return values, example outputs
3. Irrelevant Code
4. Counter-examples

Heuristic approaches? ☹
- Select **all** code blocks
- Select **all** code blocks in **accepted answers**

Language
Technologies
Institute

Our Solution

- **CONALA,** a system to collect parallel data of source **co**de snippet and **na**tural **la**nguage intents from Stack Overflow
  - *data-driven:* learn patterns of "good" and "bad" intent/snippet pairs from data (using neural networks)
  - *language-agnostic:* applicable to different programming languages (e.g., Python and Java)
  - *Scalable:* capable of applying to full-scale Stack Overflow data (collected ~600K intent/snippet pairs for Python)

Project Website: *conala-corpus.github.io*
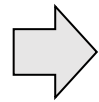
Language
Technologies
Institute

System Architecture

- Enumerate all possible candidate intent/snippet pairs in a Stack Overflow page
- Learn a classifier to rank each candidate intent/snippet pair

```
Question (i.e. intent)
removing duplicates in lists

An Answer Code Block
t = [1, 2, 3]
list(set(t))
s = [1, 2, 3]
list(set(t) − set(s))
```

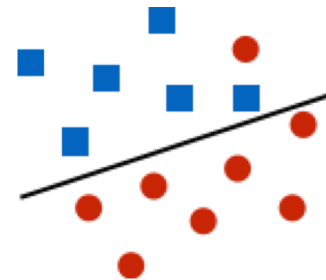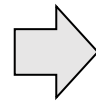Input SO question
("*how-to*" questions)

```
t = [1, 2, 3]
```

```
list(set(t))
```

```
list(set(t) − set(s))
```

```
t = [1, 2, 3]
list(set(t))
```

```
s = [1, 2, 3]
list(set(t) − set(s))
```

⋮

Candidate Snippets

Generated by enumerating contiguous sequence of code in answer code blocks

Feature-based Classifier

```
list(set(t))
```
p=0.6

```
list(set(t) − set(s))
```
p=0.2

```
t = [1, 2, 3]
list(set(t))
```
p=0.06

```
t = [1, 2, 3]
```
p=0.04

```
s = [1, 2, 3]
list(set(t) − set(s))
```
p=0.03

⋮

Ranked Snippets

Language
Technologies
Institute

Features

- **Purpose** measure the probability (plausibility) of each intent/snippet candidate

- **Two types of features** language independent, highly-indicative

$$P\left\{correct \mid \begin{matrix} \textbf{Intent} = \textit{removing duplicates in lists} \\ \textbf{Snippet} = \texttt{list(set(t))} \end{matrix} \right\}$$

Structural Features

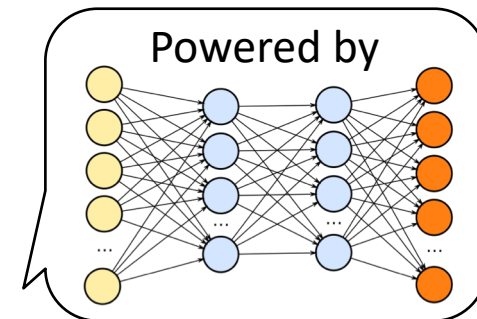Inspired by heuristic approaches, does **not** use intent information

Code Shape features: FULLBLOCK CONTAINSIMPORT STARTSWITHASSIGNMENT …

Answer quality features: ISACCEPTEDANSWER POSTRANK …
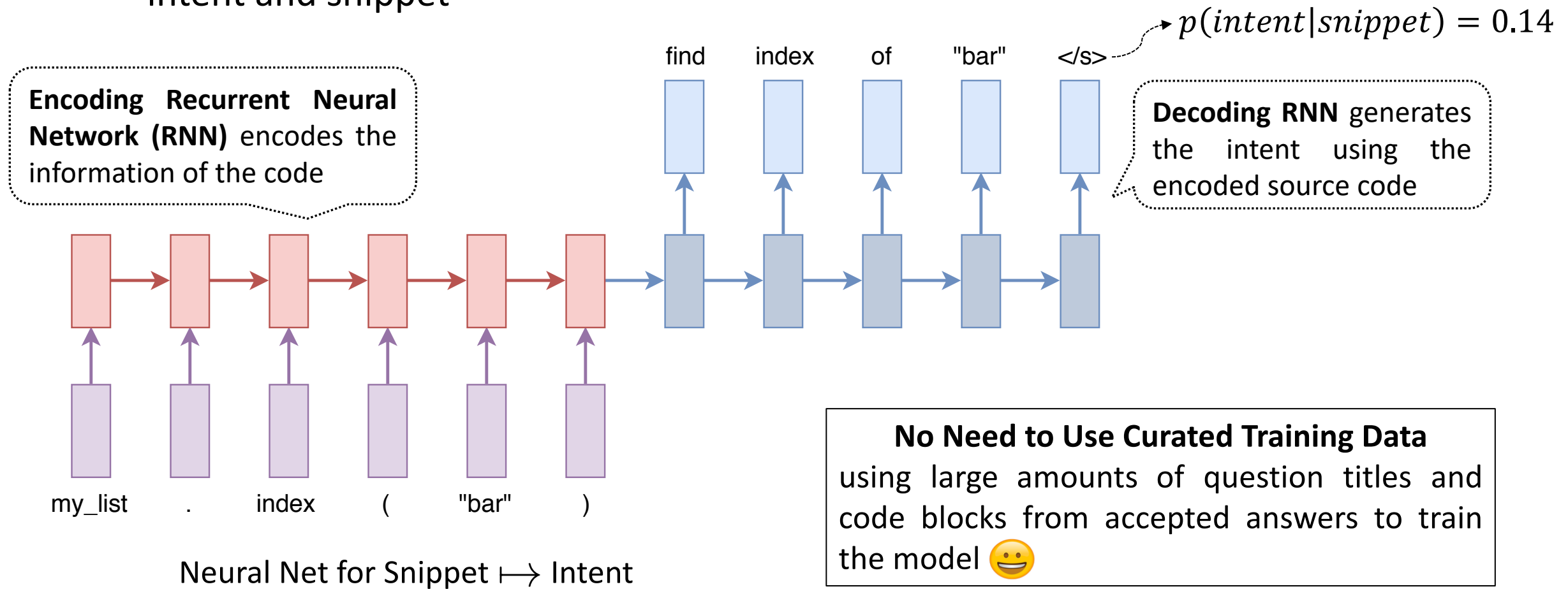
Correspondence Features

⭐ Use state-of-the-art **neural networks** to estimate the (semantic) correspondence between and intent and snippet

Powered by

$$Score\{ \textit{removing duplicates in lists} \Leftrightarrow \texttt{list(set(t))} \}$$

Language Technologies Institute

Neural Correspondence Model between Code and Intent

- **Neural sequence-to-sequence networks** for translation probability between intent and snippet

$$p(intent|snippet) = 0.14$$

find   index   of   "bar"   </s>

Encoding Recurrent Neural Network (RNN) encodes the information of the code

Decoding RNN generates the intent using the encoded source code

my_list   .   index   ("bar"   )

Neural Net for Snippet ⟼ Intent

No Need to Use Curated Training Data
using large amounts of question titles and code blocks from accepted answers to train the model 😀
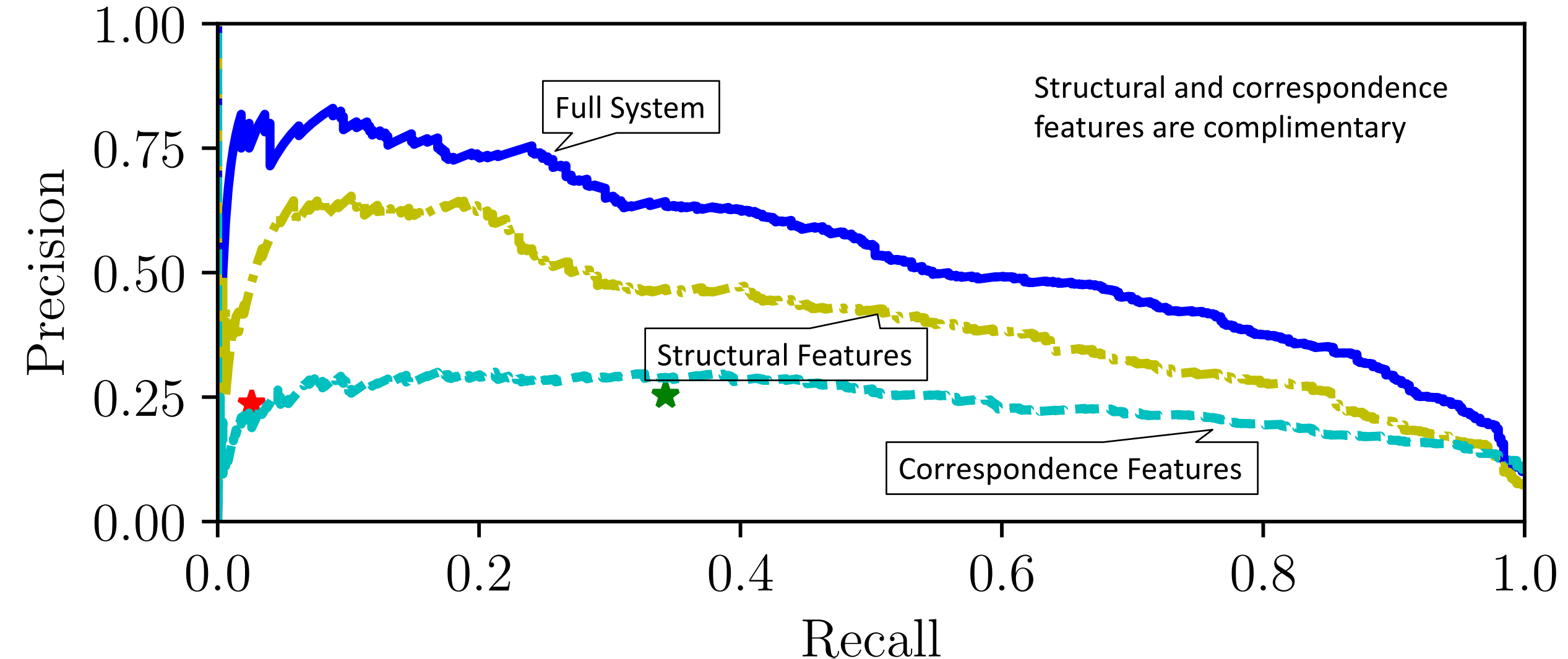
Language Technologies Institute

What Left to be Learned?

- **Structural Features** shown by existing works as indicative

- **Correspondence Features** already pre-trained on massive, readily available data on Stack Overflow (questions and code blocks)!

- We just need **small amount** of manually annotated intent/snippet pairs to tune the ~20 weights in the classifier 😀
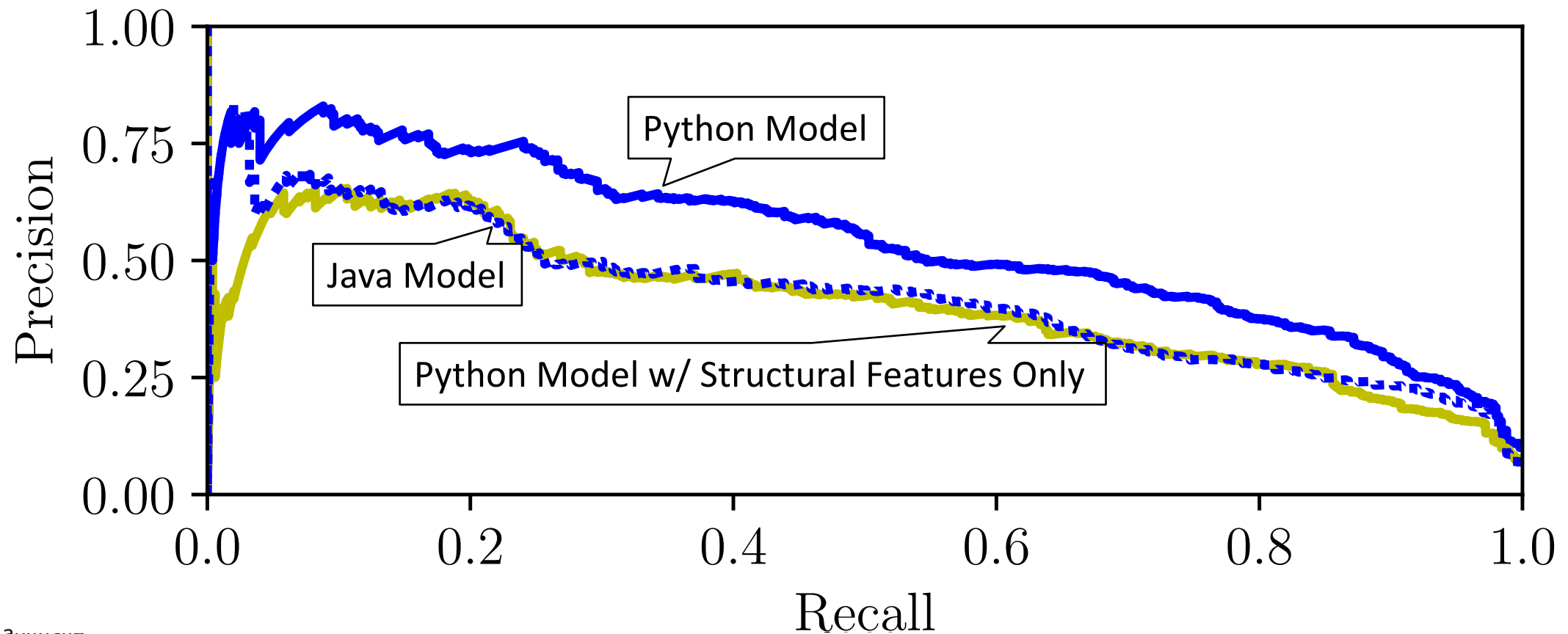
Gold Annotations

527 examples
142 questions

330 examples
100 questions

Language
Technologies
Institute

Cross-Validation Results

Structural and correspondence features are complimentary

Full System

Structural Features

Correspondence Features

Precision

Recall

Language
Technologies
Institute

Must we Annotate each Language?

- The classifier requires small set of gold-standard annotations to train on

- When apply our system to a new language, can we use the existing model trained on another (old) language?

Apply the Java model to mine Python data

Dataset Collection

- Apply the system to Python questions on Stack Overflow, collecting ~600K pairs

- ~2300 (and counting) high-quality annotated intent/code snippet pairs

- **Rewritten Intents** manually annotated, revised intents to reflect the full meaning of the code
  - Add free variable names, arguments to the intent
  - Useful for fine-grained language to code tasks like code generation

Crowdsourced

Intent	copying one file's contents to another
Rewritten Intent	copy the content of file 'file.txt' to file 'file2.txt'
Code Snippet	`shutil.copy('file.txt','file2.txt')`

An example from the annotation dataset

Language
Technologies
Institute

Examples

- Examples covers a wide variety of use cases
  - Built-in data type operation
  - I/O operation
  - Third-party library usage
- Examples are highly expressive and compositional!
  - Pose challenges to existing code/NL models

Intent	dict how to create key or append an element to key
Rewritten Intent	Create a key `key` if it does not exist in dict `dic` and append element `value` to value
Code Snippet	`dic.setdefault(key, []).append(value)`

Intent	How do I check if all elements in a list are the same
Rewritten Intent	check if all elements in list `mylist` are the same
Code Snippet	`len(set(mylist)) == 1`

Intent	Iterate through words of a file in Python
Rewritten Intent	check if all elements in list `mylist` are the same
Code Snippet	`words = open('myfile').read().split()`

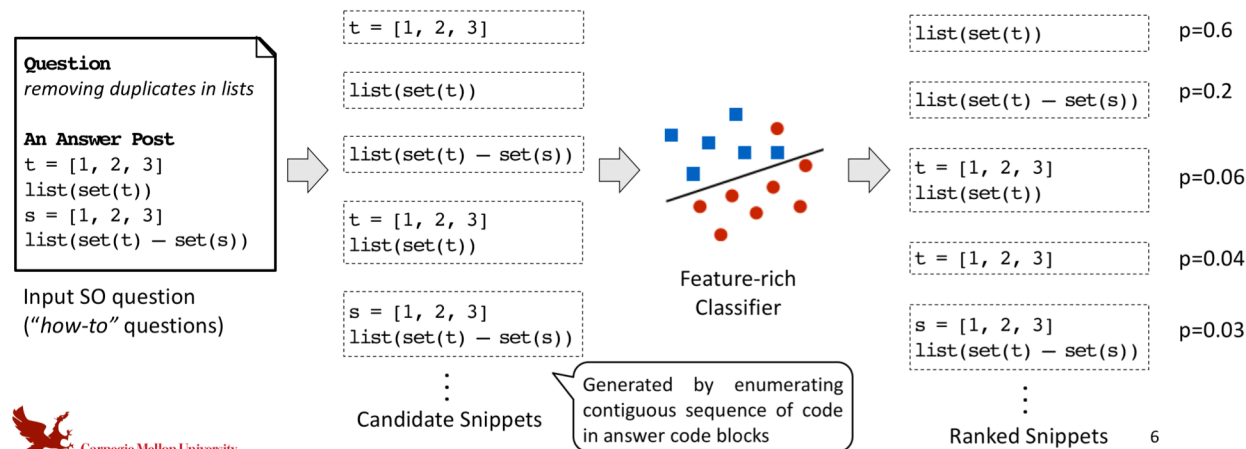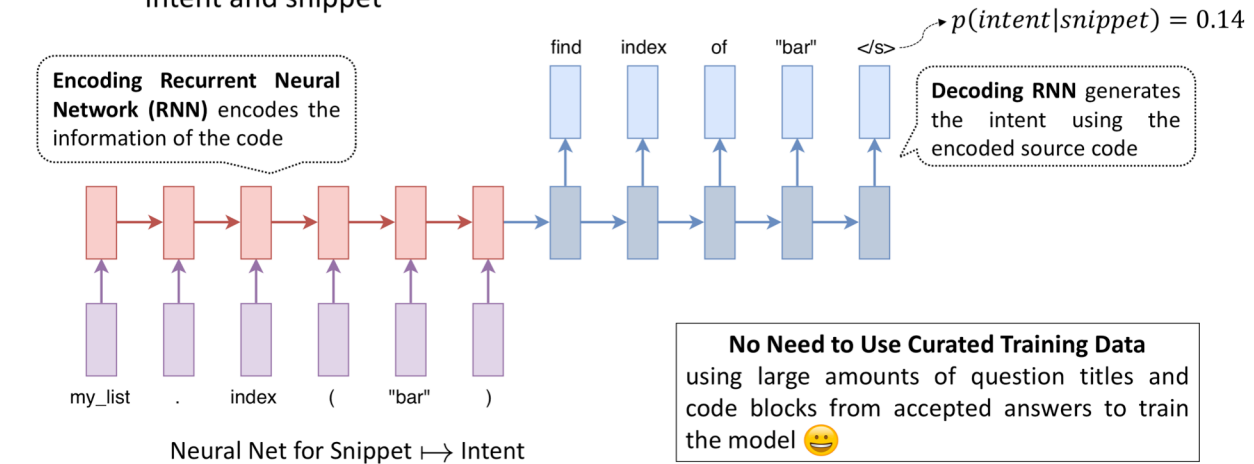|	
Intent	Delete Column in Pandas based on Condition
Rewritten Intent	delete all columns in DataFrame `df` that do not hold a non-zero value in its records
Code Snippet	`df.loc[:, ((df != 0).any(axis=0))]`

Language
Technologies
Institute

System Architecture

- Enumerate all possible candidate intent/snippet pairs in a Stack Overflow page
- Learn a classifier to rank each candidate intent/snippet pair

```
t = [1, 2, 3]
```

```
list(set(t))
```

```
list(set(t) − set(s))
```

```
t = [1, 2, 3]
list(set(t))
```

```
s = [1, 2, 3]
list(set(t) − set(s))
```

Question
removing duplicates in lists

An Answer Post
```
t = [1, 2, 3]
list(set(t))
s = [1, 2, 3]
list(set(t) − set(s))
```

Input SO question
(*"how-to"* questions)

Candidate Snippets

Generated by enumerating contiguous sequence of code in answer code blocks

Feature-rich Classifier

```
list(set(t))
```   p=0.6

```
list(set(t) − set(s))
```   p=0.2

```
t = [1, 2, 3]
list(set(t))
```   p=0.06

```
t = [1, 2, 3]
```   p=0.04

```
s = [1, 2, 3]
list(set(t) − set(s))
```   p=0.03

Ranked Snippets

Carnegie Mellon University
School of Computer Science

6

---

## Neural Correspondence Model between Code and Intent

- **Neural sequence-to-sequence networks** for translation probability between intent and snippet

$p(intent|snippet) = 0.14$

find   index   of   "bar"   </s>

**Encoding Recurrent Neural Network (RNN)** encodes the information of the code

**Decoding RNN** generates the intent using the encoded source code

my_list   .   index   (   "bar"   )

Neural Net for Snippet ⟼ Intent

**No Need to Use Curated Training Data**
using large amounts of question titles and code blocks from accepted answers to train the model 😀

Carnegie Mellon University
School of Computer Science

8

---

## Examples

- Examples covers a wide variety of use cases
  - Built-in data type operation
  - I/O operation
  - Third-party library usage
- Examples are highly expressive and compositional!
  - Pose challenges to existing code/NL models

| | |
|---|---|
| **Intent** | dict how to create key or append an element to key |
| **Rewritten Intent** | Create a key `key` if it does not exist in dict `dic` and append element `value` to value |
| **Code Snippet** | `dic.setdefault(key, []).append(value)` |

---

| | |
|---|---|
| **Intent** | How do I check if all elements in a list are the same |
| **Rewritten Intent** | check if all elements in list `mylist` are the same |
| **Code Snippet** | `len(set(mylist)) == 1` |

---

| | |
|---|---|
| **Intent** | Iterate through words of a file in Python |
| **Rewritten Intent** | check if all elements in list `mylist` are the same |
| **Code Snippet** | `words = open('myfile').read().split()` |

---

| | |
|---|---|
| **Intent** | Delete Column in Pandas based on Condition |
| **Rewritten Intent** | delete all columns in DataFrame `df` that do not hold a non-zero value in its records |
| **Code Snippet** | `df.loc[:, ((df != 0).any(axis=0))]` |

15

# Check our Dataset at
**conala-corpus.github.io**

Carnegie Mellon University
School of Computer Science

# Annotate Gold-standard Dataset

- Our system needs a *small* set of gold-standard intent/snippet data to learn the parameters of the classifier powered by high-level features
- Annotate top-ranked *how-to* questions on Stack Overflow for each language



Gold Annotations

527 examples
142 questions

330 examples
100 questions

Language
Technologies
Institute

# System Deployment

- We deploy our system on the top-50K Python-tagged questions on Stack Overflow, and collected ~600K ranked intent/code-snippet pairs