

Intent2Code

Nitin Balachandran, Yunheng Han, Zeping He, John Kastner, Callie Kim, Anjali Mittu, Suteerth Vishnu

Research Questions

The goal of our project is to investigate whether an existing natural language to code conversion technique generalizes well to non-English languages

Problem and motivation

- It is often easier to describe code in Natural Language than to immediately know the language specific implementation. Natural Language to code generators could help speed up this process.
- It is interesting to see how existing approaches generalize to natural languages other than English. Since code is not just written by English speakers, it is important for a model to be able to learn from languages other than English.

Datasets

Our baseline is the CoNaLa Dataset. The CoNaLa dataset is an English based dataset consisting of the title of a question from StackOverflow (the intent) and a snippet of code taken one of the answers to the question.

```
{
  "intent": "Normalizing a pandas DataFrame by row",
  "rewritten_intent": "normalize a pandas dataframe `df` by row",
  "snippet": "df.div(df.sum(axis=1), axis=0)",
  "question_id": 18594469
},
{
  "intent": "swap values in a tuple/list inside a list in python?",
  "rewritten_intent": "swap values in a tuple/list inside a list `mylist`",
  "snippet": "map(lambda t: (t[1], t[0]), mylist)",
  "question_id": 13384841
},
}
```

<Examples from CoNaLa Dataset>

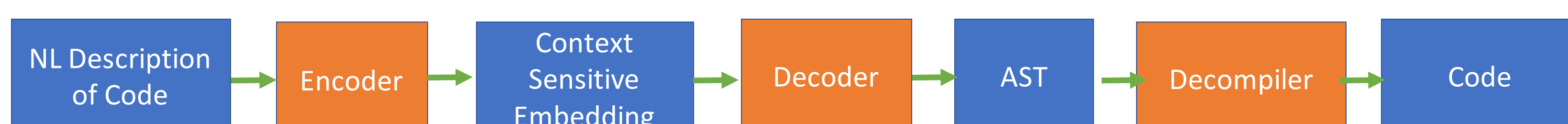
Our Spanish CoNaLa dataset comes from using Google translate to convert the CoNaLa dataset question titles to Spanish.

We gathered a Spanish Language dataset from StackOverflow which is smaller than the English dataset due to the amount of data available in Spanish (500MB Spanish v. 72GB English).

```
{
  "question_id": 1219,
  "parent_answer_post_id": 1218,
  "intent": "\u00bfc\u00f3mo obtener un valor de una lista y almacenarla en una variable con formato unicode?",
  "snippet": "a = [{u'usuario': u'Pepe'}]\nb = a[0][u'usuario']"
},
}
```

<Examples from Spanish Dataset>

Models



- We used both the NL2Code and the TranX model
- Both are based on an Encoder/Decoder model
 - Encoding: bi-directional LSTM
 - Decoder: conditional LSTM with attention
 - Uses a beam search with a size of 15 to pick the best rule from the grammar as output for the AST.

Results

Dataset	Model	BLEU Score
CoNaLa	NL2Code	17.8
CoNaLa	TranX	24.05
CoNaLa Translated to Spanish	TranX	15.4
Mined English CoNaLa	TranX	2.63
Mined Spanish Data	TranX	0
Mined Spanish Data	TranX – Trained on Spanish CoNaLa	2.58

We evaluate the generated code against the reference code using Bilinear Evaluation Understudy (BLEU). We use token-level Bilinear Evaluation Understudy (BLEU) since BLEU produces a higher score when more sub-components of the generated code matches the ground truth code.

```
-----
number and example_id:
298-1024847
original rewritten_intent:
Add key "mynewkey" to dictionary `d` with value "mynewvalue"
original snippet:
d['mynewkey'] = 'mynewvalue'
predicted code:
d['mynewkey'] = 'mynewvalue'
-----
<exact match on CoNaLa, TranX>

-----
number and example_id:
428-3964681
original rewritten_intent:
Find all files in directory "/mydir" with extension ".txt"
original snippet:
for (root, dirs, files) in os.walk('/mydir'):
    for file in files:
        if file.endswith('.txt'):
            pass
predicted code:
os.listdir('/mydir')
-----
<Failed to predict multiple lines on CoNaLa, TranX>

-----
number and example_id:
284-1024847
original rewritten_intent:
Aadir clave "mynewkey" al diccionario `d`
con el valor "mynewvalue"
original snippet:
d['mynewkey'] = 'mynewvalue'
predicted code:
d['mynewkey'] = 'str_2'
-----
<Prediction from Spanish CoNaLa, TranX>

-----
number and example_id:
76-235242
original rewritten_intent:
.write() en una sola linea Python
original snippet:
contenido = archivo.read().strip('\n').split('|')
predicted code:
print(zip(a, zip(*params)))
-----
<Prediction from Mined Spanish, TranX>
```

Key Insights

- It appears this task is hard even for languages that have similar root languages as English, given the lower BLEU score for the Spanish version of this model
- Even with a stronger model in TranX, compared to NL2Code, the BLEU score is still on the lower end which suggests there could be an issue on our end regarding the pre-processing the data.
- We believe the mined Spanish data scored so low because the data was lower quality and had fewer examples.
- To test this, we ran the TranX with the lower quality version of the English CoNaLa data which also scored much lower than the high quality data

Future Work / Additional Ideas

- Evaluate performance when given input in natural languages such as Portuguese and Russian mined from StackOverFlow
- Improve performance of model by refining mined data

Key References:

TranX : A general-purpose Transition-based abstract syntax parser that maps natural language queries into machine executable source code (<https://github.com/pcyin/tranX>)
CoNaLa: A dataset that constructs executable code snippets from natural language (<https://arxiv.org/pdf/1805.08949.pdf>)