



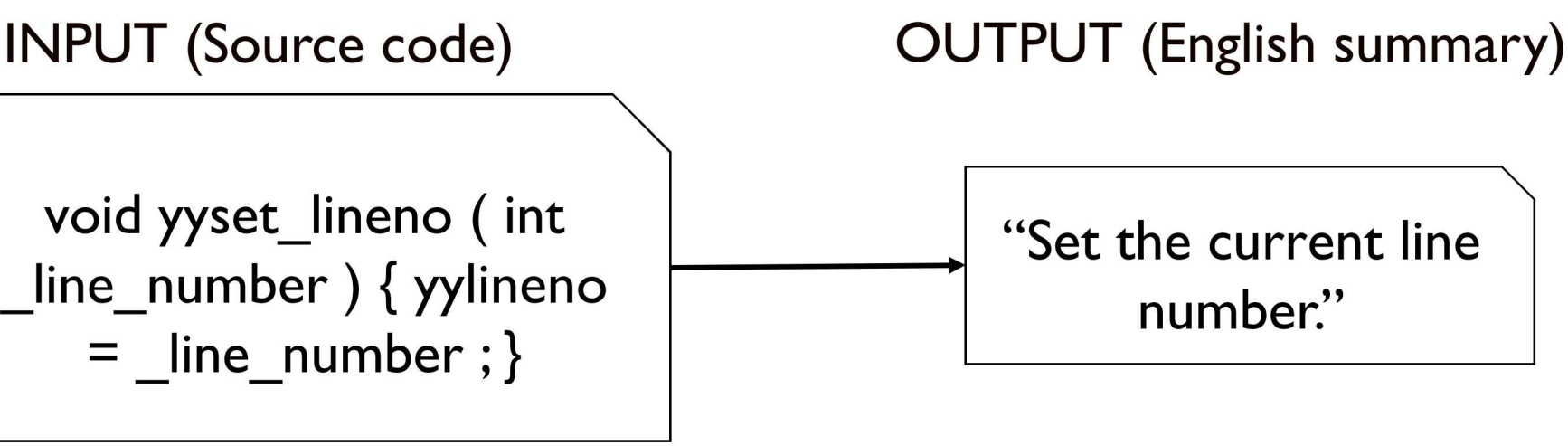
Application of Neural Translation: Source Code to English

Danielle Jacobs

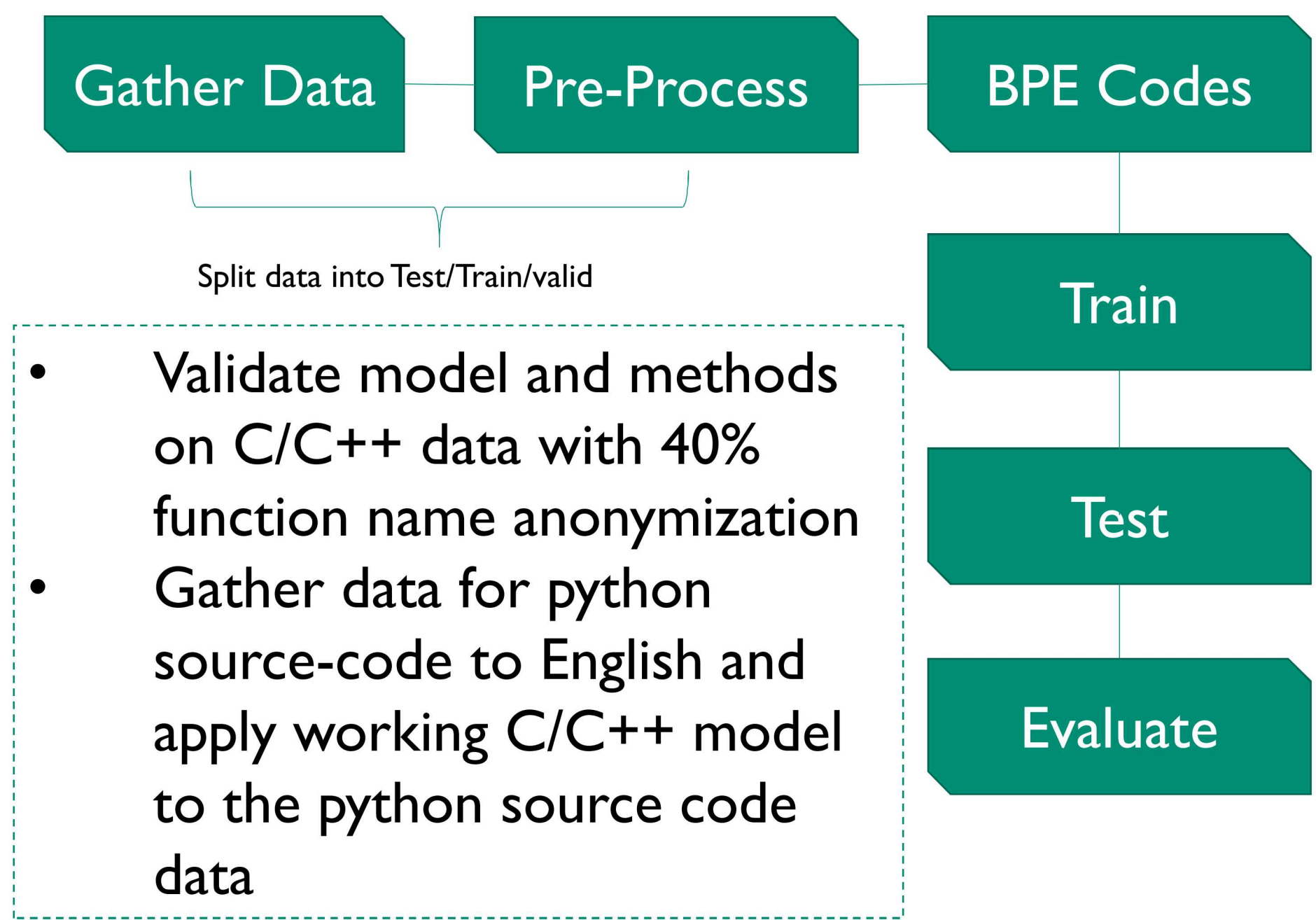
Team: Chris Harrison, Michael Bierma, Kina Kincher-Winoto, Jina Lee, David Kavaler, Sophie Quynn

Motivation

Can we produce English descriptions of functions from source code to quickly understand the code and find similar codes?



Methods



Next Steps:

- Improve preprocessing to optimize the model for python specific source-code
- Gather more data for the python source code to English translation
- Standardize data sources into one SQL DB
- Perform more extensive preprocessing on DB

Evaluation

Source Code (input):
char getField (struct board * target , int x, int y) if (x NUMBERTOKEN y NUMBERTOKEN x target width y target height) return FIELDOUTOFBOUNDS ; return * calcFieldAddress (target , x , y) ;

English Comment (ground truth):
Gets what is on a given field of the board returns FIELDXYZ constant

Model Prediction:
Returns the value of the field at the given coordinates

Source Code (input):
def merge (self , x = None , y = None , ildj = None , kwargs = None , mapping = None) : if mapping is None : mapping = _Mapping (x = x , y = y , ildj = ildj , kwargs = kwargs) elif any (arg is not None for arg in [x , y , ildj , kwargs]) : raise ValueError ("Cannot simultaneously specify mapping and individual " "arguments.") return _Mapping (x = self . _merge (self . x , mapping . x) , y = self . _merge (self . y , mapping . y) , ildj = self . _merge (self . ildj , mapping . ildj) , kwargs = self . _merge (self . kwargs , mapping . kwargs , use_equals = True))

English Comment (ground truth):
Retrieve the given resource .

Model Prediction:
Return a json dictionary representing this model .

Run	Bilingual Evaluation Understudy Score
C/C++ 0% anonymization function name	18.26
C/C++ 40% anonymization function name	4.13
Python 0% anonymization and limited pre-processing	WIP

Models

- Transformer Model uses encoder/decoders with self-attention in translation
- Attention uses vectors and weights to better translate similarities

Data

Language	C/C++	Python
Data Size Prior to Processing	1,056,099	457,445
After Processing	913157	412,125

- Data is significantly smaller for python at this time.
 - Duplicates removed *prior* to preprocessing
 - Functions > 4000 tokens removed
- Tokenization**
- Breaking text chunks into smaller parts
 - BPE** - Byte Pair Encoding