Blue ones: datasets from Microsoft

Category	Task	Dataset Name	Language	Train/Dev/Test Size		Dataset Provider	Comments
Code-Code	Clone Detection	BigCloneBench	Java	900K/416K/416K	CodeBERT	Univ. of Saskatchewan	Predict for a pair of codes
		POJ-104	C/C++	32K/8K/12K		Peking Univ	Retrieve semantically similar codes
	Defect Detection	Defects4J	C	21k/2.7k/2.7k		Univ. of Washington	Identify whether a function is vulnerable
	Cloze Test	CT-all	Python, Java, PHP, JavaScript, Ruby, Go	-/-/176k		MSRA created based on CodeSearchNet	Candidates from whole vocab
		CT-max/min	Python, Java, PHP, JavaScript, Ruby, Go	-/-/2.6k		MSRA created based on CodeSearchNet	Test reasoning ability by choosing between max/min
	Code Completion	PY150	Python	100k/5k/50k	CodeGPT	ETH Zurich, line-level created by MSRA	Support both token-level and line-level code completion, MSRA creates line-completion datasets based on these two token-level datasets.
		GitHub Java Corpus	Java	13k/7k/8k		Univ. of Edinburgh, line-level created by MSRA	
	Code Refinement	Bugs2Fix	Java	98K/12K/12K	Encoder-Decoder	The College of William and Mary	There are small and medium datasets.
	Code Translation	CodeTrans	Java-C#	10K/0.5K/1K		MSRA	Java-to-C# and C#-to-Java
Text-Code	NL Code Search	CodeSearchnet, AdvTest	Python	251K/9.6K/19K	CodeBERT	GitHub + MSR Cambridge, test provided by MSRA	Training/dev data comes from CodeSearchNet, MSRA creates a new test set by replacing function names/variables
		StacQC, WebQueryTest	Python	2.9k/0.9k/1.9k		The Ohio State Univ, test provided by MSRA	Training/dev data comes from StacQA, MSRA creates a new test set where queries are from Bing.
	Text-to-Code Generation	CONCODE	Java	100K/2K/2K	CodeGPT	Univ. of Washington	Inputs are NL documentation and code environment provided by the rest of the class.
Code-Text	Code Summarization	CodeSearchNet*	Python, Java, PHP, JavaScript, Ruby, Go	908K/45K/53K	Encoder-Decoder	Filtered based on CodeSearchNet data	Filter code-text pairs from CodeSearchNet dataset.
Text-Text	Documentation Translation	Microsoft Docs	English-Latvian/Danish/Norwegian/Chinese	53K/4K/4K		MSRA	Microsoft documentation multilingual machine translation