Code2Text Challenge: Text Generation in Source Code Libraries

Kyle Richardson†  Sina Zarrieß‡  Jonas Kuhn†

† Institute for Natural Language Processing, University of Stuttgart, Germany
{kyle,jonas}@ims.uni-stuttgart.de

‡ Dialogue Systems Group // CITEC, Bielefeld University, Germany
sina.zariess@uni-bielefeld.de

September 6, 2017
The challenge: Can we generate text from code?

Java Standard Library

* Returns the greater of two long values
* @param a an argument
* @param b another argument
* @return the larger of a and b
* @see java.lang.Long#MAX_VALUE
*/

public static Long max(long a, long b)

- A natural parallel corpus: High-level descriptions (red) of internal software functionality inside software projects (blue/black).
The challenge: Can we generate text from code?

Java Standard Library

* Returns the greater of two long values
* @param a an argument
* @param b another argument
* @return the larger of a and b
* @see java.lang.Long#MAX_VALUE
*/
public static Long max(long a, long b)

▶ A natural parallel corpus: High-level descriptions (red) of internal software functionality inside software projects (blue/black).

▶ The task: code $\rightarrow$ text
The challenge: Can we generate text from code?

Java Standard Library

* Returns the greater of two long values
* @param a an argument
* @param b another argument
* @return the larger of a and b
* @see java.lang.Long#MAX_VALUE
*/

public static Long max(long a, long b)

▶ A natural parallel corpus: High-level descriptions (red) of internal software functionality inside software projects (blue/black).

▶ The task: code → text

▶ Why? New challenges (next slide), interesting applications, explosion of interest in NLP (centrally featured in Mirella Lapata’s ACL 2017 keynote Translating from Multiple Modalities to Text and Back).
# Code2Text: Challenges for Data-to-Text Generation

| Text                                                                 | MR                                      |
|----------------------------------------------------------------------|-----------------------------------------|
| Compares this Calendar to the specified Object                      | boolean util.Calendar.equals(Object obj) |
| Extract the second component of a pair.                             | Data.Tuple.snd :: (a, b) -> b            |
| Computes the arc tangent given y and x.                             | Math.atan2(y,x) → Float                  |
| Purple7 kicks to purple9 (Chen and Mooney (2008))                   | kick(purple7,purple9)                   |
| ..Bibimbap serves French food in the price range cheap.             | name[Bibibap] food[French] priceRange[cheap] |
| (Novikova and Rieser (2016))                                        |                                         |

- **End-to-End Generation**: Existing tasks involve simpler lexicalization problem, narrower domains, lexical overlap in text and MR language, MR designed for generation, realization

- **General Question**: Are code representations rich enough to serve as a KR, can we just rely on input-output, do we need intermediate KR?
Observation 1: Tight coupling between high-level text and code, cheap and easy to extract.

```
* Returns the greater of two long values
* @param a an argument
* @param b another argument
* @return the larger of a and b
* @see java.lang.Long#MAX_VALUE
*/
public static Long max(long a, long b)
```
Parallel Source Code Data: Some Observations

Observation 1: Tight coupling between high-level text and code, cheap and easy to extract.

* Returns the greater of two long values
* @param a an argument
* @param b another argument
* @return the larger of a and b
* @see java.lang.Long#MAX_VALUE
*/

public static Long max(long a, long b)

Function signatures: Provide operationalization of text meaning.
Parallel Source Code Data: Some Observations

- **Observation 1:** Tight coupling between high-level text and code, cheap and easy to extract.

* Returns the greater of two long values
  * @param a an argument
  * @param b another argument
  * @return the larger of a and b
  * @see java.lang.Long#MAX_VALUE
*/

```java
public static Long max(long a, long b)
```

- **Function signatures:** Provide operationalization of text meaning.

  * Returns the greater of two long values
    * Long max(long a, long b)
  * Returns items from coll with random...
    * (random-sample prob coll)
Observation 2: There are many languages, hence many datasets.

```java
public static Long max(long a, long b)
```

```clojure
(defn random-sample
  "Returns items from coll with random probability of prob (0.0 - 1.0)"
  ([prob] ...) ([prob coll] ...))
```

```python
class ZipFile(object):
    """Class to open ... zip files."""
    def write(filename, arcname, ...):
        """Put the bytes from filename into the archive under the name."""
```

```python
--| Mostly functions for reading and showing RealFloat like values
module Numeric
-- | Show non-negative Integral numbers in base 10.
showInt :: Integral a => a -> ShowS
```
Observation 3: Many NLs, hence many multilingual datasets.

```php
namespace ArrayIterator;
/**
 * Appends values as the last element
 *
 * @param value The value to append
 * @see ArrayIterator::next()
 */
public void append(mixed $value)
```

```php
namespace ArrayIterator;
/**
 * Ajoute une valeur comme dernier élément
 *
 * @param value La valeur á ajouter
 * @see ArrayIterator::next()
 */
public void append(mixed $value)
```

```php
namespace ArrayIterator;
/**
 * 値を最後の要素として追加します。
 *
 * @param value 追加する値。
 * @see ArrayIterator::next()
 */
public void append(mixed $value)
```

```php
namespace ArrayIterator;
/**
 * Anade el valor como el último elemento.
 *
 * @param value El valor a anadir.
 * @see ArrayIterator::next()
 */
public void append(mixed $value)
```
| Dataset | #Pairs | #Descr Symbols | #Words Vocab. | Example Pairs (x = text, z = code) |
|---------|--------|----------------|---------------|-----------------------------------|
| Java    | 7,183  | 4,804          | 4,072 82,696 | x: Computes this Calendar to the specified Object. z: boolean util.Calendar.equals(Object obj) |
| Ruby    | 6,885  | 1,849          | 3,803 67,274 | x: Computes the arc tangent given y and x. z: Math.atan2(y,x) → Float |
| PHPen   | 6,611  | 13,943         | 8,308 68,921 | x: Delete an entry in the archive using its name. z: bool ZipArchive::deleteName(string $name) |
| Python  | 3,085  | 429            | 3,991 27,012 | x: Remove the specific filter from this handler. z: logging.Filterer.removeFilter(filter) |
| Elisp   | 2,089  | 1,365          | 1,883 30,248 | x: Returns the total height of the window. z: (window-total-height window round) |
| Haskell | 1,633  | 255            | 1,604 19,242 | x: Extract the second component of a pair. z: Data.Tuple.snd :: (a, b) → b |
| Clojure | 1,739  | –              | 2,569 17,568 | x: Returns a lazy seq of every nth item in coll. z: (core.take-nth n coll) |
| C       | 1,436  | 1,478          | 1,452 12,811 | x: Returns current file position of the stream. z: long int ftell(FILE *stream) |
| Scheme  | 1,301  | 376            | 1,343 15,574 | x: Returns a new port and the given state. z: (make-port port-type state) |

Standard library documentation for 9+ programming languages, 7 natural languages, from Richardson and Kuhn (2017b) for text → code study.
Example resources: Multilingual Standard Library Docs.

| Dataset | # Pairs | #Descr. | Symbols | Words   | Vocab. |
|---------|---------|---------|---------|---------|--------|
| PHP<sub>fr</sub> | 6,155   | 14,058  | 7,922   | 70,800  | 5,904  |
| PHP<sub>es</sub> | 5,823   | 13,285  | 7,571   | 69,882  | 5,790  |
| PHP<sub>ja</sub> | 4,903   | 11,251  | 6,399   | 65,565  | 3,743  |
| PHP<sub>ru</sub> | 2,549   | 6,030   | 3,340   | 23,105  | 4,599  |
| PHP<sub>tr</sub> | 1,822   | 4,414   | 2,725   | 16,033  | 3,553  |
| PHP<sub>de</sub> | 1,538   | 3,733   | 2,417   | 17,460  | 3,209  |

- **Non-English:** PHP documentation collection, French (fr), Spanish (es), Japanese (ja), Russian (ru), Turkish (tr), German (de)
Py27: Open Source Python projects

| Project     | # Pairs | # Symbols | # Words | Vocab. |
|-------------|---------|-----------|---------|--------|
| scapy       | 757     | 1,029     | 7,839   | 1,576  |
| zipline     | 753     | 1,122     | 8,184   | 1,517  |
| biopython    | 2,496   | 2,224     | 20,532  | 2,586  |
| renpy       | 912     | 889       | 10,183  | 1,540  |
| pyglet      | 1,400   | 1,354     | 12,218  | 2,181  |
| kivy        | 820     | 861       | 7,621   | 1,456  |
| pip         | 1,292   | 1,359     | 13,011  | 2,201  |
| twisted     | 5,137   | 3,129     | 49,457  | 4,830  |
| vispy       | 1,094   | 1,026     | 9,744   | 1,740  |
| orange      | 1,392   | 1,125     | 11,596  | 1,761  |
| tensorflow  | 5,724   | 4,321     | 45,006  | 4,672  |
| pandas      | 1,969   | 1,517     | 17,816  | 2,371  |
| sqlalchemy  | 1,737   | 1,374     | 15,606  | 2,039  |
| pyspark     | 1,851   | 1,276     | 18,775  | 2,200  |
| nupic       | 1,663   | 1,533     | 16,750  | 2,135  |
| astropy     | 2,325   | 2,054     | 24,567  | 3,007  |
| sympy       | 5,523   | 3,201     | 52,236  | 4,777  |
| ipython     | 1,034   | 1,115     | 9,114   | 1,771  |
| orator      | 817     | 499       | 6,511   | 670    |
| obspy       | 1,577   | 1,861     | 14,847  | 2,169  |
| rdkit       | 1,006   | 1,380     | 9,758   | 1,739  |
| django      | 2,790   | 2,026     | 31,531  | 3,484  |
| ansible     | 2,124   | 1,884     | 20,677  | 2,593  |
| statsmodels | 2,357   | 2,352     | 21,716  | 2,733  |
| theano      | 1,223   | 1,364     | 12,018  | 2,152  |
| nltk        | 2,383   | 2,324     | 25,823  | 3,151  |
| sklearn     | 1,532   | 1,519     | 13,897  | 2,115  |

- 27 Python projects from Github (Richardson and Kuhn (2017a)), tool for extracting datasets from APIs, FunctionAssistant and code search.
Data is already available for immediate experimentation (around 42 API datasets), task will use publicly available test sets: https://github.com/yakazimir/Code-Datasets, SMT baselines coming.

- **New Test Sets:** build using Function Assistant, by language: Java, Python, and a surprise programming language, **1 Month**

- **Evaluation:** Automatic metrics, human evaluation at IMS, University of Stuttgart for German, English (annotation funding available, extra slides). **summer 2018**

- **Organization:** IMS, University of Stuttgart (central), other possible co-organizers involving people in code to text generation (not yet official).
Evaluation

- **Automatic metrics:** METEOR and BLEU, following common practices for code-to-text (Iyer et al. (2016))

- **Human (on new testing sets):** Comparison between text generated from code and reference text (from actual API), using tools such as Kow and Belz (2012), measure along a number of dimensions: meaning similarity, adequacy, fluency/clarity.
  - **Funding:** Through IMS, special resources and funding for annotation projects.
  - **Participants:** Students from international Masters program (taught in English).
  - **New Testing Sets:** Will consist of around 1,000 data pairs, will be selected for quality.
Belz, A. and Kow, E. (2009). System Building Cost vs. Output Quality in Data-to-Text Generation. In *Proceedings of ENLG*.

Chen, D. L. and Mooney, R. J. (2008). Learning to sportscast: A test of grounded language acquisition. In *Proceedings of ICML-2008*, pages 128–135.

Iyer, S., Kostas, I., Cheung, A., and Zettlemoyer, L. (2016). Summarizing Source Code using a Neural Attention Model. *Proceedings of ACL-2016*.

Kow, E. and Belz, A. (2012). Lg-eval: A toolkit for creating online language evaluation experiments. In *LREC*, pages 4033–4037.

Miceli Barone, A. V. and Sennrich, R. (2017). A parallel corpus of Python functions and documentation strings for automated code documentation and code generation. *arXiv preprint arXiv:1707.02275*.

Novikova, J. and Rieser, V. (2016). The aNALoGuE Challenge: Non Aligned Language GEneration. In *Proceedings of INLG*.

Richardson, K. and Kuhn, J. (2017a). Function Assistant: A Tool for NL Querying of APIs. In *Proceedings of EMNLP-17 (demo)*.

Richardson, K. and Kuhn, J. (2017b). Learning Semantic Correspondences in Technical Documentation. In *Proceedings of ACL-17*.

Wong, Y. W. and Mooney, R. J. (2007). Generation by Inverting a Semantic Parser that Uses Statistical Machine Translation. In *Proceedings of HLT-NAACL*.
Extra: More about the Data

- **Noisy**: Trade off between collecting large amounts of data (for immediate experimentation), and having clean data/good documentation.

- **Preprocessing**: Function Assistant does tokenization involving conversion out of camel case, underscore elimination, ...
  - `myBigFunction(long a, long b) → my big function long a long b`
  - `(my-big-function x y) → my big function x y`

- **Underlying idea**: Not to deal with idealized documentation, but rather documentation as it exists, in its raw form, quick experimentation.

- **Data Combinations**: Participants are free to train models on multiple datasets if it helps.
Extra: Baselines

- **SMT baselines**: Phrase-based SMT, and newer Seq2Seq baselines, have previously served as reliable baselines (Wong and Mooney (2007); Belz and Kow (2009); Miceli Barone and Sennrich (2017)).

- Baseline results will be provided for all datasets (dev and test sets) by training individual models on each dataset.