Towards Zero Knowledge Learning for Cross Language API Mapping

Nghi D. Q. Bui

**Motivation**
- Language migration is important in software development
- API mapping is an indispensable step for the migration
- Existing works hinge on parallel data:
  - Require lots of human effort for data labeling
  - Require lots of parallel data for training

**Our Vision:**
**Effective API mapping with less reliance on parallel data** (towards zero knowledge)

**Key Idea**
- **Our Insight: Domain Adaptation**
  - Represent two languages as vector spaces X and Y
  - Assume X and Y are similar in geometric arrangement
  - Rotate source space X to align with target space Y

**Approach**

**STEP 1: Code Embedding**
- Extract API sequences with types
- Use Word2Vec for API embedding
- Construct two language vector spaces

**STEP 2: Domain Adaptation**
- Adapt the two vector spaces with a mapping matrix
- Learn a mapping matrix W, such that:
  \[
  \arg\min_W ||WX - Y||
  \]

**STEP 3: Query**
- Query for mapping of an API x
- Mapping = nearest neighbor(Wx)

**Adversarial Learning**
- Comprise of two neural networks:
  - **Discriminator D** (parameter \(W_D\)) maximizes likelihood of identifying an origin of an embedding
  - **Generator G** (parameter \(W_P\)) prevents the discriminator from doing so by making WX and Y as similar as possible

**Refinement**
- Frequent APIs have higher mapping accuracy. They can be utilized as knowledge to refine Adversarial Learning
- Create mapping candidates \(X_2\) and \(Y_2\) based on following heuristics:
  - Signature-based mapping candidates
  - Top-K frequent mapping candidates
  - Cosine similarity threshold mapping candidates
- Derive \(W_1\) iteratively from Procrustes Analysis of \(X_2\) and \(Y_2\)

**Evaluation**
- **Data**: Java and C# repositories on Github, 2 million files
- **Baseline**: API2API [4]
- **Metric**: Top-k accuracy (k=1,5,10)
- **Ground truth**: 860 mappings from Java2CSharp as the evaluation dataset, 2 folds validation

**Future work**
- Consider more advanced Adversarial Learning techniques
- Explore on API sequence mapping with zero knowledge
- Investigate on API mapping between languages that are different in geometric arrangement

**References**