Paradigm Shift in Source Code Translation Approaches: An AI use in Code Translation Tasks

Vikram Bhutani, Farshad Ghassemi Toosi, Jim Buckley

1. EVOLUTION SOURCE CODE TRANSLATION APPROACHES:

Approaches used in Source Code Translation

- Statistical Based Approaches
  - Stratarchical Machine Translation (SMT)
  - Hidden Markov Modelling in SMT
  - PBSMT (Phrase Based Statistical Machine Translation)

- Deep Learning - Machine Translation
  - Neural Machine Translation (NMT)
  - TransCoder by Meta\(\text{Facebook}\)
  - CodeBERT
  - CodeCG
  - CodeLLAMA

2. CATEGORIZATION OF SOURCE CODE TRANSLATION TECHNIQUES:

Overview of Source Code Translation Approaches

Source Code Translation Approaches: A Taxonomy
Categorized into semi-automated or fully automated based on our SLR findings
Subcategories include, Statistical Based (SMT), AI based (NMT)
Further AI-based techniques are classified as supervised, semi-supervised and unsupervised

3. EVALUATION OF MACHINE TRANSLATED SOURCE CODE:

Evaluation of Source Code Quality using Static Analysis Tools

Motivation
No universal metric exists for Translated Code Quality (CodeBLEU, FE) to identify quality of automated output from AI-based tools.
Static Analysis Tools use on translated code.
Checking for rules (maintainability, exceptions, buffer flow etc.)
Checking for Dataflow, Syntax errors, Model Checking, Compilation issues.
Can be added as part of full E2E flow for source code translation framework.

4. GENAI (GENERATIVE AI) USE IN CREATION OF BENCHMARK FOR TRAINING AND EVALUATION:

Solving Data Scarcity issues in Source code Translation

Real Data can be:
Expensive and time consuming
Has privacy issues
Has Safety issues
Should be updated real time with feedbacks.
Data cleanup can be tedious.

Synthetic Data Generation:
Uses Gen AI Approach (GAN, Cyclic GAN)
Easy to generate (match original dist)
Safe to use
No privacy issues
Successful in Image generation.
Source code data gen is under-explored