Reformulating MaskLID with ILP for Robust Code-Switching Detection

Name: Dwi Prima Handayani Putri

Course: Master 1 DAC – PLDAC

Date: May 22st, 2025

Background

Why is code-switching challenging?

- Frequent in multilingual contexts
- Existing LID models fail at segment-level accuracy

Limitations of current tools:

- Greedy segmentation (MaskLID)
- No global inference
- Fragmented outputs

Objectives

Reformulate MaskLID using global optimization

Improve segmentation consistency

Evaluate performance using GlotLID & OpenLID

Focus on Turkish–English dataset

Dataset

Turkish-English CS dataset from Yirmibeşoğlu & Eryiğit (Detecting Code-Switching between Turkish-English Language Pair)

339 CS sentences

341 mono Turkish sentences

Preprocessing:

- CS: ≥ 40 bytes
- Mono: ≥ 20 bytes

Format: token-level annotations

MaskLID Overview

Heuristic masking strategy

Iterative: mask dominant language to expose others

Relies on token-level logit vectors

No global objective, no continuity guarantee

ILP Reformulation

$$\max \sum_{i \in W} \sum_{j \in L} b_{ij} \cdot x_{ij}$$

Constraints:

- One label per token
- Max 2 active languages
- Minimum byte length per language segment

Constraint

- **Unique language label per token:**

$$\sum_{j \in L} x_{ij} = 1 \quad \forall i \in W$$

- **Language activation constraint:**

$$\sum_{i \in W} x_{ij} \leq |W| \cdot y_j \quad \forall j \in L$$

  This ensures that $y_j = 1$ only if label j is assigned to at least one token.

- **Maximum number of active languages:**

$$\sum_{j \in L} y_j \leq K$$

- **Minimum byte-length per language:**

$$\sum_{i \in W} \mathrm{len}(i) \cdot x_{ij} \geq \tau \cdot y_j \quad \forall j \in L$$

Evaluation Metrics

Exact Match (EM)

A prediction is counted as **EM** if the predicted set of language labels exactly matches the gold set—no missing labels and no extra ones.

- **Monolingual:**
  - Gold: `{tr}` → Predicted: `{tr}`
  - Predicted: `{tr, eng}` → not EM
- **Code-Switched:**
  - Gold: `{tr, eng}` → Predicted: `{tr, eng}`
  - Predicted: `{tr}` → not EM

Evaluation Metrics

Partial Match (PM)

- **Monolingual:**
  PM is counted if the predicted set **includes the correct label**, even if it includes additional incorrect ones.
  - Gold: `{tr}` → Predicted: `{tr, eng}` → PM
  - Predicted: `{eng}` → not PM
- **Code-Switched:**
  PM is counted if the prediction shares **at least one label** with the gold set, **and does not contain any labels outside** the gold set.
  - Gold: `{tr, eng}` → Predicted: `{tr}` → PM
  - Predicted: `{eng}` → PM
  - Predicted: `{tr, fr}` → not PM (this will be counted as FP)

Note: **All EM cases are also counted as PM**.

Evaluation Metrics

False Positives (FP)

A prediction is counted as FP if it includes **any label not present in the gold set** (only counted in Code-Switch Dataset)

- **Code-Switched:**
  - Gold: `{tr, eng}` → Predicted: `{tr, de}` → FP

Results Summary

Dataset	#S	MaskLID (Heuristik)				Optimization (ILP)			
		#EM/#PM ↑		#FP ↓		#EM/#PM ↑		#FP ↓	
		GlotLID	OpenLID	GlotLID	OpenLID	GlotLID	OpenLID	GlotLID	OpenLID
CS Turkish–English	339	**74**/315	**58**/301	19	32	127/133	135/145	199	186
Mono Turkish	341	**331**/338	**323**/336	-	-	109/332	107/332	-	-

Table 1: Evaluation results on CS Turkish–English with and without MaskLID. We report the number of exact matches (EM), partial matches (PM), and false positives (FP).

Analysis

ILP improves EM but increases FP in CS data

Heuristic better for mono data (higher EM, zero FP)

Trade-off: precision vs recall

ILP is sensitive to segment length and ambiguity

Conclusion & Future Work

ILP reformulation enhances robustness for CS detection

Doesn’t require prior knowledge of language pairs

Limitations: no constraint on label contiguity

Future work:

- Add continuity constraints
- Test on more LinCE datasets

Thank You!