1. Introduction

- Background
  - There are many multi-lingual texts in WWW
  - In automatic processing of multi-lingual texts,
    1. Segmentation by language.
    2. Identification of the language of each segment.
- Segmentation is useful:
  - Preprocessing of multi-lingual text analysis
  - Assembling corpora of non-major languages

2. Problem Formulation

- Assumptions
  - For all multilingual texts, every text portion is written in one of the given languages.
  - Amount of learning data is small (up to kilobytes)
  - Only a limited amount of corpus data is available.

- Notation
  - \( X = x_0 \cdots x_{|X|-1} \): i-th character of text \( X \).
  - \( X = [x_0, \cdots, x_{|X|-1}] \): List of text portions, concatenation of \( X_i \)'s equals to \( X \).
  - \( L = [l_0, \cdots, l_{|L|-1}] \): List of languages of corresponding portions.

- Formulation
  - Our formulation is based on the MDL: \( \{dL(X)\} \) description length of \( X \):
    \[
    (\hat{X}, \hat{L}) = \arg \min_{X,L} \sum_{i=0}^{|X|-1} \sum_{j=0}^{|L|-1} dL_i(x_i) \quad \text{with } |X| = |L|
    \]

3. Calculation of Cross-Entropy

Chose 2 each from two representative compression methods.

- Mean of Matching Statistics (MMS)
  - Dictionary-based method [Farach 94], [Juola 97]
  - A simplified method of Lempel-Ziv algorithm [Ziv&Lempel 77]
  - This method uses matching statistics \( \text{len}(Y) \)
  - Modification applied to decrease computational complexity
  
  \[
  MMS_Y(X) = E \left[ \frac{\log|Y|}{\text{len}(Y)} \right] 
  \]

- Prediction by Partial Matching (PPM)
  - Statistical method [Clealy&Witten 84], [Teahan 00]
  - A variation of n-gram model, using Witten-Bell smoothing
  \[
  \text{PPM}_Y(X) = \prod_{i=1}^{|X|} P_Y(x_i|x_{i-1} \cdots x_{i-n})
  \]

4. Segmentation by Dynamic Programming

- Recursive Formulation
  - We use DP to avoid exponential increase of computational complexity.
  - Our formula based on MDL can be decomposed recursively as follows:
    \[
    DP(X, L) = \min_{l \in \{0, \ldots, |L|-1\}} \{DP(x_0 \cdots x_{|X|-1}, L') + dL_i(x_1 \cdots x_{|X|-1}) \}
    \]

- Complexity Reduction
  - Straightforward implementation: \( O(|L|^2|X|^2) \)
  - The complexity decreases to: \( O(|L||X|) \)
  - Reduction Techniques:
    1. Caching
    2. Locality of context dependency
    3. Keep scores of 2 best languages

5. Experiment

- Using Datasets
  - Data (Training 4 / Test 1)
    - UDHR (Universal Declaration of Human Rights) 277
    - Wiki (Wikipedia) 222

- Experimental Setting (for Multi-lingual Text)
  - 5-fold cross-validation (Training Part 4 / Test Part 1)
  - All test data generated artificially from above data
  1. Choose number of segments \( K = 1 \sim 5 \)
  2. Choose K languages are randomly chosen
  3. For each language, take 40 to 160 chars from test part
  4. Shuffle and concatenate K portions.
  5. Each line is generated by varying the \( \gamma \)

- Results (P-R Plot & Max F-score)
  - Results for language detection
    - Very high F-score
  - Results for segmentation
    - When restricting boundaries to occur at any character border
    - When restricting boundaries to occur only at white spaces

Reference


This work was supported by MEXT JSPS KAKENHI Grant Number (2400005, 20240006)