Detecting and Analyzing Code-Switching Between English and Darija

Belfqueh Hatim,Mouhcine Riany

January 20, 2025

Abstract

This report explores the detection and analysis of code-switching between English and Darija, a Moroccan Arabic dialect, in multilingual texts. Given the unique challenges of Darija, including its lack of standardized writing systems and frequent lexical borrowings from English, our study aims to modify and improve existing methods for detecting code-switching in this under-explored linguistic context. We present a detailed review of recent studies on code-switching detection and propose a framework tailored to the specific characteristics of Darija-English interactions.

Contents

1 Introduction

Code-switching, the practice of alternating between two or more languages within a conversation or text, is a common phenomenon in multilingual communities. This study focuses on detecting and analyzing code-switching between English and Darija, a Moroccan Arabic dialect. Darija-English code-switching presents unique challenges due to the lack of standardized writing, the casual transliteration of words, and the significant lexical borrowing from English. Despite the growing body of research on code-switching detection, particularly in language pairs such as English-Spanish, Mandarin-English, and Turkish-English, the application of these techniques to Darija-English is scarce.

The goal of this study is to adapt and enhance existing code-switching detection methods for the Darija-English language pair. We aim to develop a system that can accurately identify instances of code-switching and differentiate them from lexical borrowings, a key challenge in this context. This report reviews existing literature, highlights gaps in current research, and proposes an approach for improving code-switching detection in Darija-English texts.

2 Related Work

Detecting and modeling code-switching has become a significant area of study, especially in multilingual contexts, including popular language pairs such as English–Spanish, Mandarin–English, and Turkish–English. Although these studies offer valuable insights into various methods for detecting code-switching, direct applications to Darija–English, a Moroccan Arabic dialect paired with English, are still underexplored. Darija poses unique challenges, including the absence of standardized writing systems, casual transliteration practices, and the lack of annotated data. The aim of our research is to adapt and improve existing methods for this linguistically distinctive pair.

2.1 Detailed Annotation and Linguistic Differences

The importance of precise annotation for differentiating between code-switching and other linguistic phenomena is emphasized in "Borrowing or Code-Switching? Annotating for Finer-Grained Distinctions in Language Mixing." This study provides a token-level annotated Twitter corpus that distinguishes between

code-switching and lexical borrowing, specifically in the context of English and Spanish. This distinction is crucial for Darija–English texts, as the frequent use of English lexical borrowings can obscure the boundaries between code-switching and borrowing. We aim to adopt a similar annotation strategy to provide clearer distinctions in Darija–English, which will enhance the accuracy of code-switching detection in this context.

2.2 Supervised Techniques and Model Specialization

Another significant category of research involves the use of supervised methods, which require large labeled datasets for training. The authors of "Code-Switched Language Models Using Dual RNNs and Same-Source Pretraining" developed a model that combines two specialized recurrent neural networks (RNNs) for Mandarin–English code-switching. This approach reduced perplexity by handling the linguistic nuances of each language separately before integrating them. Similarly, "Detecting Code-Switching Between Turkish-English Language Pair" demonstrated the effectiveness of Conditional Random Fields (CRF) and character-level features for Turkish-English code-switching. These results suggest that developing specialized models for each language could improve the performance of code-switching detection in Darija–English, particularly if combined with a quality annotated dataset such as a collection of Darija–English social media posts or tweets.

2.3 Scalable Solutions Requiring Fewer Resources

Given the challenges of collecting large annotated datasets, several studies have explored methods that reduce reliance on labeled data. For instance, "Estimating Code-Switching on Twitter with a Novel Generalized Word-Level Language Detection Technique" presented a novel Hidden Markov Model (HMM) approach, achieving an impressive 96.3% accuracy without manual annotations. This approach, though not tested directly on Darija–English, highlights a potential solution for scenarios where annotated data is scarce. Since Darija–English code-switching often occurs in informal user-generated content, an unsupervised or semi-supervised model could be an efficient starting point, which can later be fine-tuned with small, high-quality labeled datasets.

2.4 Advanced Learning of Representations and Contrastive Methods

In addition to traditional models, recent work has focused on advanced representation learning techniques to enhance the performance of language detection in code-switching tasks. "SuperConText: Supervised Contrastive Learning Framework for Textual Representations" introduced a supervised contrastive learning framework that improves the quality of textual representations by promoting better intra-class compactness and inter-class separation. Although tested primarily on English datasets, the methodology offers significant promise for Darija–English code-switching detection, where language boundaries can be less clearly defined due to varying transliterations and frequent borrowing. By leveraging contrastive learning, we could enhance the model's ability to distinguish between Darija and English in mixed-language contexts.

2.5 Conclusion

The body of research on code-switching detection spans a wide range of techniques, from detailed manual annotation to unsupervised learning and advanced representation frameworks. The studies reviewed offer numerous valuable insights for adapting these methods to Darija–English code-switching. By combining detailed annotation strategies, supervised model specialization, scalable unsupervised approaches, and advanced learning methods, our research aims to create a flexible and effective system tailored to the specific challenges of Darija–English code-switching.

3 Methodology

To tackle the unique challenges presented by Darija–English code-switching, we propose a multi-step methodology that integrates various detection techniques and adapts them for this specific language pair. This includes the following steps:

3.1 Data Collection and Preprocessing

3.1.1 BERT Part

- **Data Sources:** The data was collected from a code switching dataset containing phrases having both darija and english words and a dataset from the url https://huggingface.co/datasets/imomayiz/darija-english/resolve/main/sentence containing seperate darija and english phrases
- **Data Cleaning:** The data was preprocessed to remove non-textual elements such as URLs, hashtags, and mentions. Standardization of transliterations and tokenization was performed for consistency.
- **Labeling:** Each word was labeled as either dar or eng.
- **Data Augmentation:** Techniques such as back-translation and paraphrasing were applied to increase dataset size, especially for the code-switching class.

3.1.2 BiLSTM Part

- **Data Sources:** The same dataset used for BERT was utilized for training the BiLSTM model.
- **Data Preparation:** The data was split into sequences of tokenized words, and labels were one-hot encoded.
- **GloVe Embeddings:** Pre-trained GloVe embeddings (with 400,000 words) were used to map words to dense vector representations.

3.2 Model Construction

3.2.1 BERT Part

The multilingual BERT (mBERT) model was used for language modeling. The model was fine-tuned for text classification tasks by adjusting the last few layers and adding a dropout layer to prevent overfitting.

- **Fine-Tuning:** Fine-tuning was done by adjusting the last layers of mBERT, with a learning rate of 5e-5.
- **Batch Size:** A batch size of 16 was used during training.

- **Training Time:** The model was trained for 3 to 5 epochs.

3.2.2 BiLSTM Part

The BiLSTM model was built with GloVe embeddings. It consisted of two bidirectional LSTM layers, followed by dense layers and a dropout layer.

- **Embedding Layer:** Initialized with pre-trained GloVe embeddings.
- **Bidirectional LSTM Layers:** Two layers of BiLSTM were stacked to capture both forward and backward dependencies.
- **Dense Layer:** A dense layer with softmax activation was added to output the classification.
- **Dropout Layer:** A dropout rate of 0.3 was applied to prevent overfitting.

Model Summary: The BiLSTM model had a total of 1,000,000 parameters, with most being non-trainable (due to the GloVe embeddings).

3.3 Model Training

3.3.1 BERT Part

The model was trained using the AdamW optimizer with a learning rate of 5e-5. The model was fine-tuned for 3-5 epochs.

- **Training Results:** The model achieved an overall accuracy of 99
- **Loss:** The loss steadily decreased after training.

3.3.2 BiLSTM Part

The BiLSTM model was trained for 10 epochs using the Adam optimizer with a learning rate of 1e-3.

- **Epoch 1:** 84.50% accuracy with loss 0.412.
- **Epoch 10:** Accuracy of 98.82% with loss of 0.04.
- **Training Time:** Each epoch took about 75 seconds, with the total training process lasting approximately 750 seconds.

3.4 Evaluation Metrics

3.4.1 BERT Part

The BERT model was evaluated using accuracy, precision, recall, f1-score, and confusion matrix.

Classification Report:

Class	Precision	Recall	F1-Score
Darija	0.99	1.00	1.00
English	0.99	0.99	0.99
Code-Switching	0.99	0.99	0.99

Confusion Matrix:

$$\begin{bmatrix} 537 & 0 & 1 \\ 3 & 520 & 2 \\ 0 & 4 & 365 \end{bmatrix}$$

3.4.2 BiLSTM Part

The BiLSTM model was also evaluated using the same metrics, achieving an accuracy of 99.30% and a loss of 0.0285.

Classification Report:

Class	Precision	Recall	F1-Score
Darija	1.00	1.00	1.00
English	0.99	0.99	0.99
Code-Switching	1.00	0.98	0.99

Confusion Matrix:

$$\begin{bmatrix} 538 & 0 & 0 \\ 2 & 522 & 1 \\ 0 & 7 & 362 \end{bmatrix}$$

3.5 Conclusion

Both the BERT and BiLSTM models demonstrated effective classification capabilities for multilingual text, particularly in Darija and English code-switching. While BERT showed slightly better performance in understanding code-switching contexts, BiLSTM was particularly strong at capturing sequential dependencies within the text. Future improvements could involve expanding the dataset and further fine-tuning the models.