ENHANCED NEURAL MACHINE TRANSLATION FOR MALAY-ENGLISH CODE-SWITCHED TEXT

WONG YIT KHEE

UNIVERSITI TEKNOLOGI MALAYSIA

PSZ 19:16 (Pind. 1/13)

UNIVERSITI TEKNOLOGI MALAYSIA

DECLARATION OF THESIS / UNDERGRADUATE PROJECT REPORT AND COPYRIGHT

Author’s full name : WONG YIT KHEE

Date of Birth : 17TH JAN 1998

Title : ENHANCED NEURAL MACHINE TRANSLATION FOR MALAY-ENGLISH CODE-SWITCHED TEXT

Academic Session : 1/20232024

I declare that this thesis is classified as:

|  |  |  |
| --- | --- | --- |
|  | CONFIDENTIAL | (Contains confidential information under the Official Secret Act 1972)\* |
|  |  |  |
|  | RESTRICTED | (Contains restricted information as specified by the organization where research was done)\* |
|  |  |  |
| ✓ | OPEN ACCESS | I agree that my thesis to be published as online open access (full text) |

1. I acknowledged that Universiti Teknologi Malaysia reserves the right as follows:
2. The thesis is the property of Universiti Teknologi Malaysia
3. The Library of Universiti Teknologi Malaysia has the right to make copies for the purpose of research only.
4. The Library has the right to make copies of the thesis for academic exchange.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  | Certified by: |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  | SIGNATURE OF STUDENT |  | SIGNATURE OF SUPERVISOR |  |
|  |  |  |  |  |
|  | PCS223010 |  | DR. SHARIN HAZLIN BINTI HUSPI |  |
|  | **MATRIC NUMBER** |  | **NAME OF SUPERVISOR** |  |
|  |  |  |  |  |
|  | Date: 03 OCT 2023 |  | Date: 03 OCT 2023 |  |

NOTES : If the thesis is CONFIDENTIAL or RESTRICTED, please attach with the letter from the organization with period and reasons for confidentiality or restriction

“I hereby declare that I have read this proposal and in my

opinion this proposal is sufficient in term of scope and quality for the

award of the degree of Doctor of Philosophy in (Computer Science)”

|  |  |  |
| --- | --- | --- |
| Signature | : | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| Name of Supervisor I | : | SHARIN HAZLIN BINTI HUSPI |
| Date | : | 3 October 2023 |
|  |  |  |
|  |  |  |
| Signature | : | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| Name of Supervisor II | : |  |
| Date | : | 3 October 2023 |
|  |  |  |
|  |  |  |
| Signature | : | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| Name of Supervisor III | : |  |
| Date | : | 3 October 2023 |

**BAHAGIAN A - Pengesahan Kerjasama\***

Adalah disahkan bahawa projek penyelidikan tesis ini telah dilaksanakan melalui

kerjasama antara \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_dengan \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Disahkan oleh:

Tandatangan : Tarikh :

Nama :

Jawatan :

(Cop rasmi)

*\* Jika penyediaan tesis atau projek melibatkan kerjasama.*

**BAHAGIAN B - Untuk Kegunaan Pejabat Sekolah Pengajian Siswazah**

Tesis ini telah diperiksa dan diakui oleh:

|  |  |  |
| --- | --- | --- |
| Nama dan Alamat Pemeriksa Luar | **:** |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
| Nama dan Alamat Pemeriksa Dalam | **:** |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
| Nama Penyelia Lain (jika ada) | **:** |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

Disahkan oleh Timbalan Pendaftar di SPS:

|  |  |  |  |
| --- | --- | --- | --- |
| Tandatangan | : |  | Tarikh : 15JULAI 2018 |
| Nama | : |  | |

ENHANCED NEURAL MACHINE TRANSLATION FOR MALAY-ENGLISH CODE-SWITCHED TEXT

WONG YIT KHEE

A proposal submitted in partial fulfilment of the

requirements for the award of the degree of

Doctor of Philosophy in (Computer Science)

School of Computing

Faculty of Engineering

Universiti Teknologi Malaysia

OCT 2023

DECLARATION

I declare that this proposal entitled *“Enhanced Neural Machine Translation For Malay-English Code-Switched Text”* is the result of my own research except as cited in the references. The proposal has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

|  |  |  |
| --- | --- | --- |
| Signature | : | .................................................... |
| Name | : | WONG YIT KHEE |
| Date | : | 3 October 2023 |
|  |  |  |
|  |  |  |

DEDICATION

This thesis is dedicated to my father, who taught me that the best kind of knowledge to have is that which is learned for its own sake. It is also dedicated to my mother, who taught me that even the largest task can be accomplished if it is done one step at a time.

ACKNOWLEDGEMENT

In the preparation of this thesis, I had the privilege of collaborating with numerous individuals, including esteemed researchers, dedicated academicians, and insightful practitioners. Their contributions have greatly enriched my understanding and shaped my thoughts. I would like to extend my heartfelt gratitude to the following individuals:

First and foremost, I am deeply indebted to my primary thesis supervisor, Dr. Sharin Hazlin binti Huspi. Her unwavering encouragement, expert guidance, constructive critiques, and genuine friendship have been instrumental in the development of this thesis.

I would also like to express my sincere appreciation to my co-supervisor, Dr. Haslina binti Hashim, and Associate Professor Dr. Hishamuddin Jamaluddin. Their invaluable guidance, insightful advice, and unwavering motivation have significantly contributed to the completion of this work.

My profound thanks go to Universiti Teknologi Malaysia (UTM) for generously funding my Ph.D. studies. Additionally, I extend my gratitude to the librarians at UTM, Cardiff University of Wales, and the National University of Singapore for their invaluable assistance in providing relevant literature.

I would like to acknowledge my fellow postgraduate students, whose support has been both encouraging and inspiring. I also extend my appreciation to all my colleagues and individuals who have offered their insights and advice on various occasions. While it is impossible to name everyone in this limited space, your contributions have not gone unnoticed.

Lastly, my deepest gratitude goes to my family members, whose unwavering support and understanding have been a constant source of strength and motivation throughout this journey.

ABSTRACT

Code-switching, the dynamic practice of shifting between two languages within a single discourse, presents significant challenges for accurate machine translation. These challenges made up of the nuances of informal language, diverse grammar structures, and inherent ambiguity. This research proposed a comprehensive three-phase framework, comprises language identification, code-switching type classification, and segment-based translation techniques, to tackle the challenges mentioned of translating Malay-English code-switched texts. The proposed framework will be enhanced with linguistic resources that help resolve ambiguity in language identification and translation process. Therefore, the study meticulously created essential linguistic resources, including an ambiguity word dictionary for language identification, a homonyms dictionary enriched with Part-of-Speech tagging and relevant words, and a Malay-English code-switched parallel corpus. These resources will be seamlessly integrated into the framework to elevate language identification precision and translation fidelity. The research process includes corpus acquisition, dataset curation, and the development of cutting-edge neural machine translation models. This conceptual framework is proposed with the ambition of enabling translation that understands the context, ensuring accurate and semantic-rich translations of code-switched texts.

ABSTRAK

Pertukaran kod, amalan dinamik menukar antara dua bahasa dalam satu perbualan tunggal, memberikan cabaran besar untuk terjemahan mesin yang tepat. Cabaran-cabaran ini terdiri daripada wujudnya bahasa yang tidak formal, struktur tatabahasa yang berlainan, dan keambiguan yang berwujud. Kajian ini mencadangkan satu rangka kerja tiga fasa komprehensif, yang merangkumi pengenalpastian bahasa, pengelasan jenis pertukaran kod, dan teknik terjemahan berdasarkan segmen, untuk mengatasi cabaran-cabaran tersebut dalam terjemahan teks bertukar kod. Rangka kerja yang dicadangkan akan diperkukuhkan dengan sumber linguistik yang membantu menyelesaikan keambiguan dalam pengenalpastian bahasa dan proses terjemahan. Oleh itu, kajian ini dengan teliti mencipta sumber-sumber linguistik penting, termasuk kamus perkataan keambiguan untuk pengenalpastian bahasa, kamus homonim yang diperkayakan dengan penandaan Part-of-Speech (kata-kata yang menunjukkan peranan dan fungsinya dalam struktur sesuatu ayat) dan perkataan yang berkaitan, dan korpus serentak kob bertukar dalam Bahasa Melayu dan Bahasa Inggeris. Sumber-sumber ini akan diintegrasikan ke dalam rangka kerja untuk meningkatkan ketepatan pengenalpastian bahasa dan ketepatan terjemahan. Proses kajian melibatkan pengumpulan korpus, penyediaan dataset, dan pembangunan model terjemahan mesin neural yang terkini. Rangka kerja konseptual ini dicadangkan dengan matlamat membolehkan terjemahan yang memahami konteks, memastikan terjemahan yang tepat dan kaya makna bagi teks bertukar kod.

TABLE OF CONTENTS

|  |  |  |
| --- | --- | --- |
|  | TITLE | PAGE |

[DECLARATION iii](#_Toc147309486)

[DEDICATION iv](#_Toc147309487)

[ACKNOWLEDGEMENT v](#_Toc147309488)

[ABSTRACT vi](#_Toc147309489)

[ABSTRAK vii](#_Toc147309490)

[TABLE OF CONTENTS viii](#_Toc147309491)

[LIST OF TABLES xii](#_Toc147309492)

[LIST OF FIGURES xiii](#_Toc147309493)

[LIST OF ABBREVIATIONS xv](#_Toc147309494)

[LIST OF SYMBOLS xvi](#_Toc147309495)

[LIST OF APPENDICES xvii](#_Toc147309496)

[CHAPTER 1 INTRODUCTION 1](#_Toc147309497)

[1.1 Introduction 1](#_Toc147309498)

[1.2 Problem Background 2](#_Toc147309499)

[1.3 Problem Statement 3](#_Toc147309500)

[1.4 Research Aim 4](#_Toc147309501)

[1.5 Research Objectives 4](#_Toc147309502)

[1.6 Research Scope 5](#_Toc147309503)

[1.7 Expected Contribution 5](#_Toc147309504)

[1.8 Report Organization 6](#_Toc147309505)

[1.9 Chapter Summary 7](#_Toc147309506)

[CHAPTER 2 LITERATURE REVIEW 9](#_Toc147309507)

[2.1 Introduction 9](#_Toc147309508)

[2.2 Linguistic Code-switching 9](#_Toc147309509)

[2.2.1 Malay-English Code-switching 11](#_Toc147309510)

[2.2.1.1 Malay Language 12](#_Toc147309511)

[2.2.1.2 English Language 13](#_Toc147309512)

[2.2.1.3 Language Structure of Malay and English: Similarities and Differences 13](#_Toc147309513)

[2.3 Translation Process 16](#_Toc147309514)

[2.4 Machine Translation 18](#_Toc147309515)

[2.5 Neural Machine Translation 20](#_Toc147309516)

[2.5.1 General Framework 21](#_Toc147309517)

[2.5.1.1 Word Embedding 21](#_Toc147309518)

[2.5.1.2 Encoder-Decoder Framework 22](#_Toc147309519)

[2.5.1.3 Language Identification 22](#_Toc147309520)

[2.5.1.4 Recurrent Neural Network (RNN) 23](#_Toc147309521)

[2.5.2 State-of-art 28](#_Toc147309522)

[2.6 Performance Metric 29](#_Toc147309523)

[2.7 Previous Research 31](#_Toc147309524)

[2.7.1 Code-switching: Data synthesis 31](#_Toc147309525)

[2.7.2 Code-switching: Language model 32](#_Toc147309526)

[2.7.3 Code-switching: NMT language model 34](#_Toc147309527)

[2.8 Challenges 36](#_Toc147309528)

[2.9 Discussion 38](#_Toc147309529)

[2.10 Research Gap 40](#_Toc147309530)

[2.11 Chapter Summary 41](#_Toc147309531)

[CHAPTER 3 RESEARCH METHODOLOGY 43](#_Toc147309532)

[3.1 Introduction 43](#_Toc147309533)

[3.2 Research Framework 43](#_Toc147309534)

[3.2.1 Phase 1: Literature Review and Problem Formulation 51](#_Toc147309535)

[3.2.2 Phase 2: Data Collection and Pre-processing 51](#_Toc147309536)

[3.2.3 Phase 3: Preliminary Study and Model Development 51](#_Toc147309537)

[3.2.4 Phase 4: Result Evaluation and Discussion 52](#_Toc147309538)

[3.2.5 Phase 5: Documentation 52](#_Toc147309539)

[3.3 Tools and Platforms 53](#_Toc147309540)

[3.4 Proposed Solution: Experiment Design 53](#_Toc147309541)

[3.4.1 Research Activities: Data Collection 55](#_Toc147309542)

[3.4.2 Research Activities: Data Preprocessing 58](#_Toc147309543)

[3.4.3 Research Activities: Model Development 59](#_Toc147309544)

[3.5 Chapter Summary 61](#_Toc147309545)

[CHAPTER 4 LANGUAGE IDENTIFICATION 62](#_Toc147309546)

[4.1 Introduction 62](#_Toc147309547)

[4.2 Implementation 62](#_Toc147309548)

[4.2.1 Ambiguous word Identification 63](#_Toc147309549)

[4.2.2 Language Identification Model 63](#_Toc147309550)

[4.3 Performance Evaluation 64](#_Toc147309551)

[4.4 Chapter Summary 64](#_Toc147309552)

[CHAPTER 5 CODE-SWITCHING TYPE IDENTIFICATION 65](#_Toc147309553)

[5.1 Introduction 65](#_Toc147309554)

[5.2 Implementation 65](#_Toc147309555)

[5.2.1 Preparing Code-switch Type Training Data 66](#_Toc147309556)

[5.2.2 Code-Switching Types Identification Model 67](#_Toc147309557)

[5.3 Performance Evaluation 67](#_Toc147309558)

[5.4 Chapter Summary 67](#_Toc147309559)

[CHAPTER 6 SEGMENT-BASED TRANSLATION 69](#_Toc147309560)

[6.1 Introduction 69](#_Toc147309561)

[6.2 Implementation 69](#_Toc147309562)

[6.2.1 Malay-English Parallel Corpus 70](#_Toc147309563)

[6.2.2 Homonyms Dictionary 70](#_Toc147309564)

[6.2.3 Translation Process for Different Code-Switching Types 70](#_Toc147309565)

[6.3 Performance Evaluation 71](#_Toc147309566)

[6.4 Chapter Summary 72](#_Toc147309567)

[CHAPTER 7 CONCLUSION 73](#_Toc147309568)

[7.1 Preliminary Outcomes 73](#_Toc147309569)

[7.2 Achievement of Project Objectives 73](#_Toc147309570)

[7.3 Future Research Steps 74](#_Toc147309571)

[7.4 Conclusion 74](#_Toc147309572)

[REFERENCES 75](#_Toc147309573)

[LIST OF PUBLICATIONS 87](#_Toc147309574)

LIST OF TABLES

|  |  |  |
| --- | --- | --- |
| TABLE NO. | TITLE | PAGE |

[Table 2.1 Malay-English Code-switching Types and Examples. 11](#_Toc147309482)

[Table 2.2 The phonology of Malay and English Language 14](#_Toc147309483)

[Table 2.3 Previous Research 33](#_Toc147309484)

[Table 3.1 Research Design 45](#_Toc147309485)

LIST OF FIGURES

|  |  |  |
| --- | --- | --- |
| FIGURE NO. | TITLE | PAGE |

[Figure 2.1 Malay Phonemic Chart adapted from https://omniglot.com/writing/malay.htm 14](#_Toc147309575)

[Figure 2.2 English Phonemic Chart (Source: https://www.englishclub.com/pronunciation/phonemic-chart.php) 14](#_Toc147309576)

[Figure 2.3 Type of Machine Translation and its techniques (adapted from (Grundkiewicz, 2018)) 19](#_Toc147309577)

[Figure 2.4 General framework of NMT: Encoder-decoder architecture 22](#_Toc147309578)

[Figure 2.5 Simple RNN (source: https://www.v7labs.com/blog/recurrent-neural-networks-guide) 24](#_Toc147309579)

[Figure 2.6 LSTM unit 25](#_Toc147309580)

[Figure 2.7 GRU unit 26](#_Toc147309581)

[Figure 2.8 Alignment sample (Bahdanau et al., 2016) 27](#_Toc147309582)

[Figure 2.9 Encoder-decoder architecture with attention mechanism 27](#_Toc147309583)

[Figure 3.1 Research Framework 44](#_Toc147309584)

[Figure 3.2 Proposed Framework 54](#_Toc147309585)

[Figure 3.3 Data Collecting Flowchart 55](#_Toc147309586)

[Figure 3.4 Covid-19 Health Forum Platform, DoctorOnCall (English Version) 57](#_Toc147309587)

[Figure 3.5 Covid-19 Health Forum Platform, DoctorOnCall (Malay Version) 57](#_Toc147309588)

[Figure 3.6 Data Pre-processing Flowchart 58](#_Toc147309589)

[Figure 3.7 Model Development Flowchart 59](#_Toc147309590)

[Figure 4.1 Language Identification Phase Flowchart 63](#_Toc147309591)

[Figure 5.1 Code-switch Type Identification Phase Flowchart 65](#_Toc147309592)

[Figure 6.1 Segment-Based Translation Phase Flowchart 69](#_Toc147309593)

[Figure 6.2 Performance Evaluation Flowchart 71](#_Toc147309594)

LIST OF ABBREVIATIONS

|  |  |  |
| --- | --- | --- |
| MT | - | Machine Translation |
| RBMT | - | Rule-based Machine Translation |
| SMT | - | Statistical machine translation |
| NMT | - | Neural Machine Translation |
| HMT | - | Hybrid Machine Translation |
| NLP | - | Natural Language Processing |
| UTM | - | Universiti Teknologi Malaysia |
| RNN | - | Recurrent Neural Network |
| LSTM | - | Long Short-Term Memory |
| GRU | - | Gated Recurrent Unit |
| GNMT | - | Google Neural Machine Translation |
| BERT | - | Bidirectional Encoder Representation from Transformers |
| NER | - | Named Entity Recognition |
| BLEU | - | Bilingual Evaluation Understudy |
| TER | - | Translation Edit Rate |
| WER | - | Word Error Rate |
| METEOR | - | Metric for Evaluation of Translation with Explicit ORdering |
| XNLI | - | Cross-lingual Natural Language Inference |
| ACCMT | - | Sentiment Classification on Code-Mixed Text |
| OOV | - | Out-of-vocab |
| AM | - | Main Language |
| AE | - | Embedding Language |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

LIST OF SYMBOLS

|  |  |  |
| --- | --- | --- |
| δ | - | Minimal error |
|  | - | Diameter |
|  | - | Force |
|  | - | Velocity |
|  | - | Pressure |
|  | - | Moment of Inersia |
|  | - | Radius |
|  | - | Reynold Number |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

LIST OF APPENDICES

|  |  |  |
| --- | --- | --- |
| APPENDIX | TITLE | PAGE |

[Appendix A Gantt Chart **Error! Bookmark not defined.**](#_Toc147247578)

[Appendix B Web Scrapping Code 84](#_Toc147247579)

# INTRODUCTION

## Introduction

Text mining, or text data mining is the process of extracting and analyzing important information or knowledge from text data (Tan, Ridge, Labs, & Terrace, 2000). Along with information discovery, text mining is highly commercial profitable due to the invaluable of knowledge discovery from data mining and information obtained from the analysis (Tan et al., 2000). Text mining offers information extraction and patent analysis to identify trends, patterns and relationships for both educational aspect and business aspect (Hotho, Nürnberger, & Paaß, 2005) such as marketing analysis, customer analysis and decision-making process.

As the Internet continues to evolve, text mining for online platforms, such as forums and social networks, which are widely utilized by individuals for self-expression, becomes increasingly significant. These online data resources come in massive volume and provide rich information that is easily accessible. According to Tan et al. (2000), text data is the most natural form of information storing medium, however, its unstructured and fuzzy characteristics induced higher complexity in text mining process than general data mining process (Tan et al., 2000; Hotho et al., 2005). Unstructured data is the type of data without predefined format which further increases the complexity of pattern and information extraction (Hotho et al., 2005).

Besides, ambiguity and noise present in these online text data bring negative impacts on the accuracy and reliability of text mining results. Along with spelling errors, abbreviations, slang, and multiple meanings of words, people tend to switch between their spoken languages while expressing themselves. The adoption of multiple languages in spoken or written form is called code-switching. Code-switching is a process of constructing a sentence verbally with two-or-more languages (Ahmad Mahir & Silahudin Jarjis, 2007). Semi-formal or informal platforms such as social media and online forums frequently encountered written versions of mixed languages. In the context of code-switched sentences, only one language served as the main language while the others are the embedded languages. The main language can be easily identified by its morpheme counts in sentences or discourse (Izyani, 2016).

For cross-lingual or code-switched text mining, translation technique serves as the first line of the analysis process (translation-based approach). The challenges due to the massive volume of the text data have driven the researchers in adopting more sophisticated techniques like machine translation (MT) in the analysis. MT approach evolves from Rule-based machine translation (RBMT) and Statistical machine translation (SMT) and finally evolves to Neural machine translation (NMT) with the advancement of deep learning. Neural Machine Translation (NMT) was proposed with the use of deep neural networks in encoding and decoding the text, and soon developed as another state-of-art for machine translation [3] (Phua, Navaratnam, Kang, & Che, 2022), where the interconnected nodes in neural network allow NMT models to learn from huge volumes of data and adapt to new inputs. However, general NMT models are trained on monolingual parallel corpora of text. Their ability to deal with the code-switching situation is at the least as the models may not be able to correctly identify the different languages being used in a single sentence.

## Problem Background

Malaysians are an example bilingual or multilingual population. Almost every Malaysians would have their mother tongue to be their second language after Malay, with English being their third language. Within this linguistically diverse context, both bilingual and multilingual Malaysians frequently engage in code-switching, seamlessly transitioning between their spoken languages during conversations and self-expression. This practice is particularly prevalent on informal platforms such as social media, where it can lead to the distortion of information, posing challenges for data mining and analysis. The most commonly encountered code-switched language pair is Malay-English.

The inherent differences in language structures between Malay and English present difficulties in comprehending the meaning of code-switched sentences. While both languages generally follow a subject-verb-object word order, subtle variations exist in adjective placement and noun plurality. English introduces unique tenses, singular and plural nouns, whereas Malay expresses time frames and plurality differently. Affixes in both languages further complicate translation, with English indicating tenses and plurality and Malay indicating part-of-speech (POS). The inclusion of non-standard words and the presence of ambiguity further exacerbate translation challenges, as words can possess multiple meanings or entirely differ between languages.

Moreover, standard neural machine translation (NMT) models are typically trained on monolingual parallel corpora of text, making them ill-suited to handle code-switching scenarios, as they may struggle to accurately identify the different languages within a single sentence. The scarcity of parallel datasets, coupled with the low-resource status of Malay, further impedes translation processes. On social media platforms, informal code-switching often introduces non-standard words that are frequently overlooked in normalization efforts. The presence of ambiguity in code-switched texts adds an additional layer of complexity, hampering translation accuracy as translation models may struggle to classify words and accurately interpret their intended meanings.

In light of these challenges, this research proposes a code-switched identification mechanism paired with a corresponding translation mechanism to enhance the performance of NMT in Malay-English code-switch translation scenarios.

## Problem Statement

The scarcity and quality of parallel corpus for Malay-English code-switched texts significantly impact the quality of machine translation. This challenge is compounded by informal language usage, differences in language structure and grammar, and the presence of ambiguity in code-switched texts.

The following research questions served as the guidelines for designing the research:

RQ 1. How can we effectively identify the primary sources of ambiguity that emerge during language identification and translation processes in code-switched scenarios, and systematically categorize and structure the identified sources of ambiguity to enhance the effectiveness of the language identification and translation processes?

RQ 2. How can we detect and classify the distinctive linguistic features and patterns of code-switched segments within Malay-English code-switched text data?

RQ 3. How can a translation process be tailored and optimized to effectively address the unique characteristics of each type of code-switched segment, resulting in more accurate and contextually appropriate translations?

## Research Aim

This research aims to enhance Malay-English code-switched text translation by developing a three-phase framework and essential linguistic resources. The goal is to address language nuances, grammar diversity, and ambiguity in code-switching. By pioneering code-switching type classification and segment-based translation techniques, the study aims to improve translation accuracy and context-awareness, ultimately facilitating effective cross-lingual communication.

## Research Objectives

Objectives are identified and set as the guidelines to ensure the achievement of the goals of this research. Each of them serves as a milestone for every research process. The objectives of the research are:

RO 1. To identify and organize the ambiguity present during language identification and translation process.

RO 2. To proposed an effective identification method that identify potential code-switch segments from code-switch text data.

RO 3. To develop a translation process optimized for each type of code-switched segment.

## Research Scope

Necessary boundaries are introduced to ensure the implementation of this research could achieve the objective identified. These boundaries help guide the design and development process throughout the research. The scopes of the research are:

1. The research will focus on the translation of Malay-English code-switch text.
2. The code-switched dataset will be collected from online health question and answering forum, DoctorOnCall, which consists over 170 general diseases.
3. The language of the question will be in Malay-English mixed mode.
4. The pre-process process and proposed model will be developed by Python language.
5. The effectiveness of the translated text is evaluated by performance metrics review, baseline model comparison and crowd-sourcing evaluation.

## Expected Contribution

Notably, this research is expected to contribute by creating essential linguistic resources, including a Malay-English ambiguity word dictionary with relevant words and a homonyms dictionary with POS tagging and relevant words. These resources are anticipated to be crucial for effectively addressing language ambiguity, a common challenge in code-switched neural machine translation.

Additionally, the work is expected to introduce a code-switching type identification technique. This innovation is designed to handle the inherent ambiguity within Malay-English code-switched content. This contribution is anticipated to significantly enhance the accuracy of translating mixed-language texts by identifying and classifying code-switching types.

The proposed three-phase framework is expected to offer a structured and systematic approach to tackle the specific challenges in translating Malay-English code-switched texts. This framework is designed to address language identification, code-switching type classification, and segment translation. Its expected contribution lies in providing valuable guidance to both researchers and practitioners in the field of machine translation.

In summary, the expected contributions of this research encompass the development of crucial linguistic resources, the introduction of an innovative code-switching type identification technique, and the provision of a structured three-phase framework. These contributions are anticipated to advance the field of Malay-English code-switched neural machine translation by effectively addressing language ambiguity and providing valuable tools for improved translation accuracy.

## Report Organization

This report consists of seven main chapters. In Chapter 1, the information about the research is explained in detail such as the aim, objectives, scope, problem statement and the contribution of this research. Chapter 2 is made up of literature reviews related to the code-switching situation in online platform with a discussion on their trends and challenges that arise in translation quality. In Chapter 3, the research methodology applied and the framework used are explained in detail. The overall design for this experiment is briefly explained by phases in Chapter 4, 5, and 6 accordingly. Besides, the preliminary result obtained from each phase are displayed and analyzed. Finally, a conclusion is made on current progress including the summary of findings and contribution achieved, the limitation and challenges encountered and finally, further research direction are expressed in Chapter 7.

## Chapter Summary

This chapter introduced basic domain knowledge related to this research and particular problems or rooms for improvement for the current solution. The aim and objectives of the research are stated clearly as the research guidelines while scopes are determined to show the right path to continue the research. Preliminary studies and some literature reviews will be introduced in the next chapter.

# LITERATURE REVIEW

## Introduction

This chapter provides a review of the literature on the key concepts and techniques relevant to our research. It begins with an overview of the languages involved including their nature and structure, which is essential for understanding the complexities of our dataset, which is the code-switching situation. The chapter then examines machine translation methods and the various approaches to enhance the translation model, including addressing the issues in word embedding and out-of-vector. By exploring the state-of-the-art, current trends, and discussing the implications of each technique, this chapter provides a comprehensive foundation for our research.

## Linguistic Code-switching

Language was developed along with the ability of people in communicating and comprehend one another (Cummings-Clay& Hostos Community College, 2019). It is a structured system that uses spoken or written language to link people. Understanding a language includes but not limited to produce and comprehend both spoken and written words of the language (Cummings-Clay et al., 2019). A language differs itself from the others by two unique concepts: Grammar rules that create their framework, and the symbols, or so-called lexicon which are the sum total of all words in the language (Cummings-Clay et al., 2019). According to the concept of Syntax, a language is a sequential data, any order disruptions occurred will result in tremendous different or meaningless text. As the sentences “Lily ate the apple” and “The apple ate Lily” own a totally different meaning with a different word order.

Language has evolved as people learned to communicate and understand each other, but it differs over time and in different places. Although the language used in a defined geographical location is commonly the same, there exist countries that speak more than one language. The adoption of multiple languages in spoken or written form is called code-switching. Code-switching is a process of constructing a sentence verbally with two-or more languages (Ahmad Mahir, & Silahudin Jarjis, 2007). Bilingual or multilingual speakers adopt code-switching in both verbal and textual communication for better expressing themselves with the most suitable word in any language. Semi-formal or informal platforms such as social media and online forums frequently encountered written versions of mixed languages. Code-switching is incredibly useful since it enables bilingual or multilingual speakers to communicate using the most appropriate word in any language. In the context of code-switched sentences, only one language served as the main language while the others are the embedded languages. Main language can be easily identified by its morpheme-counts in the sentences or discourse (extended language unit, greater than a sentence) (Izyani, 2016).

Common code-switching types are calques (Hashim, 2017), intra-word code switching (Hidayatullah, Qazi, Lai., & Apong, 2022), inter-sentential code-switching (Kester, 2021) and intra-sentential code-switching (Mohamad Khalil, & Mohd Shahril Firda, 2018). According to Hashim (2017), calque involves the application of other language structure into the sentence without taking syntax's logicalness into account. Intra-word code switching is the combination of affixes of one language to the word of the other language (Hidayatullah et al., 2022). Inter-sentential code-switching includes the language switching in the range of phrasal, sentences or discourse boundaries, while intra-sentential code-switching indicates direct word replacement of another language into the sentence such as nouns, verbs, discourse markers (linking words) and etc (Mohamad Khalil et al., 2018). However, the code-switching situation does not always follow the modes explained. Consequently, modelling the code-switching situation could be more complex.

### Malay-English Code-switching

Malaysia as a country with a diverse population, most Malaysian are bilingual or multilingual. With the existing of 137 living languages, the four major spoken languages in Malaysia are Malay, English, Tamil, and Chinese. The majority of Malaysians would have their mother tongue to be their second language after Malay, with English being their third. Both bilingual and multilingual Malaysians tend to switch between their spoken languages while having conversation or expressing themselves in casual events. Posts on social media and online forums frequently use written versions of mixed languages.

The majority of Malaysians use mixed-language communication that combines Malays with English, Chinese with English, or Chinese with Malay. However, the code-switching language studied in this research is the Malay-English code-switching, which the Malay language is the main language while the English language serves as the embedded language. The following examples illustrated few instances of Malay-English code-switched types and their corresponding standard sentences:

Table 2.1 Malay-English Code-switching Types and Examples.

|  |  |
| --- | --- |
| Type (a): Calque | |
| Malay-English Code-switched | Saya pergi ke kedai, tetapi terlupa dompet **di** rumah. |
| Standard Malay | Saya hendak pergi ke kedai, tetapi terlupa bawa dompet **dari** rumah. |
| Standard English | I am going to the store, but forgot my wallet **at** home. |
| Type (b): Intra-word code-switching | |
| Malay-English Code-switched | Jobnya → job [English] + nya [Malay] |
| Standard Malay | Kerjanya |
| Standard English | His job |
| Type (c): Intra-sentential code-switching | |
| Malay-English Code-switched | I pesakit covid 19 and family sy buat swab test, tiada simptom demam, adakah mereka akan negative? |
| Standard Malay | Saya pesakit COVID-19 dan keluarga saya menjalani ujian saringan swab, tanpa mengalami sebarang gejala demam. Adakah mereka akan mendapat hasil negatif? |
| Standard English | I am a COVID-19 patient and my family underwent a swab test, with no fever symptoms. Will they test negative? |
| Type (d): Inter-sentential code-switching | |
| Malay-English Code-switched | Saya nak pergi ke kedai serbaneka [Malay structure], but forget my wallet at home [English structure]. |
| Standard Malay | Saya hendak pergi ke kedai serbaneka, tetapi terlupa bawa dompet dari rumah. |
| Standard English | I am going to the convenience store, but forgot my wallet at home. |

#### Malay Language

The Malay language is categorized under the Austronesian language family (Phua et al., 2022), spoken by native people living around Straits of Malacca. As the rising of Malacca Straits as popular sea route, the use of Malay language in communication has involved Indonesian islands and it became the lingua franca of Indonesian archipelago. According to National Language Act 1967 and Article 152 of the Federal Constitution of Malaysia, the Malay language, also refers to ‘Bahasa Melayu’ locally, is the national language of Malaysia National Language (Mohamad, Salleh, & Mohamad, 2022). Although other languages are used as mother tongue for various ethnics, Malay language serves as the main language in both formal and education situation.

#### English Language

The English language is categorized under the Indo-European language family, spoken by native people living around medieval England. According to The Most Spoken Languages 2023 (2023), English secures its global lingua franca position with about 1.5 billion English speakers. Due to Malaysia’s colonial history, English language is introduced during British Empire in 18th centuries. Although the Malay language is use in most formal and government events, English language serves as the second language for Malaysian in most professional events such as technology and commerce events. As the learning and application of English language in Malaysia, most Malaysian are Malay-English bilinguals and the English language spoken by Malaysian is ‘corrupted’ with their national language, dialects and other languages such as Chinese and Indian language.

#### Language Structure of Malay and English: Similarities and Differences

All language framework is built up by their own grammar rules and filled by its unique lexicons. Due to the indifferent language structure, Malay-English code-switched sentences induced difficulties for meaning comprehension.

While both having Latin writing system, the Malay and English languages own very different language phonology (sound system) (Figure 2.1 and Figure 2.2) including the quantity of vowels and consonants. Table 2.2 illustrates the quantity different in both languages and their corresponding examples. Besides, Malay and English language record a very distinctive place and manner for articulation. However, the writing system and phonology affect very least if compare to the language grammar in comprehend a code-switch sentence.

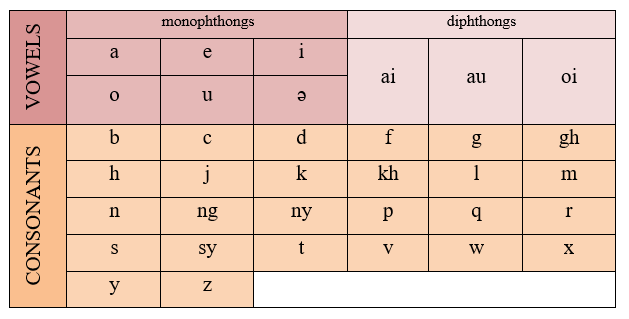


Figure 2.1 Malay Phonemic Chart adapted from https://omniglot.com/writing/malay.htm

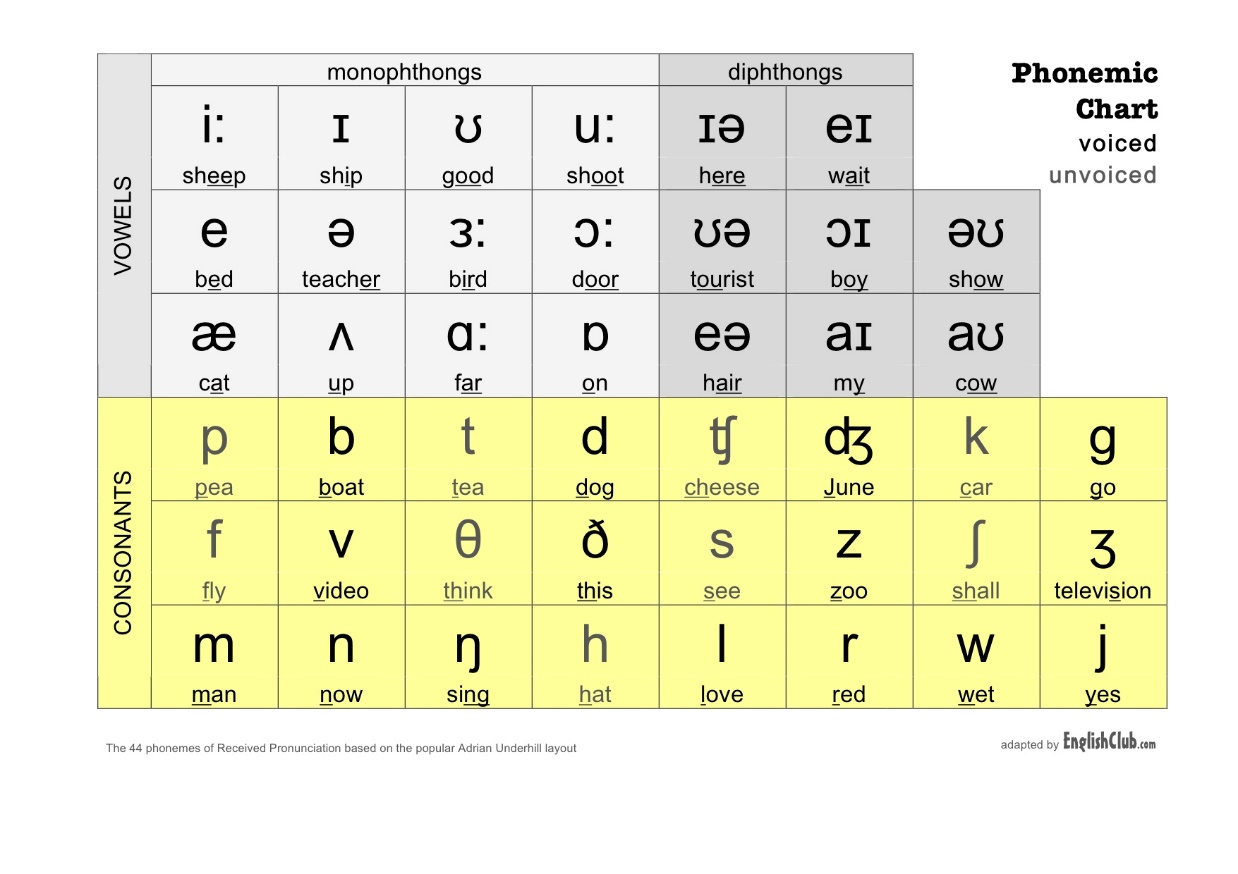


Figure 2.2 English Phonemic Chart (Source: https://www.englishclub.com/pronunciation/phonemic-chart.php)

Table 2.2 The phonology of Malay and English Language

|  |  |  |  |
| --- | --- | --- | --- |
| Language | Quantity | | |
| Vocals | Consonants | Diphthong |
| Malay | 6 | 27 | 3 |
| [a], [e], [i], [ə] | [p], [b], [t], [l] | [ai], [au], [ oi] |
| English | 12 | 24 | 8 |
| [æ], [ɒ], [i:] | [d], [k], [g] | [aɪ], [aʊ], [ɔɪ] |

In the perspective of word order, both Malay language and English language prefer subject-verb-object order. For instances, in Malay, “Saya [subject] pergi [verb] ke pasar [object] kemarin”; in English, ‘I [subject] went to [verb] the market [object] yesterday’. However, English language prefers adjectives before nouns (“Blue shirt”) while Malay language prefers nouns before adjectives (“Baju [Shirt] biru [blue]”). Observing the similar word order, total disrupted translated text is occurred at minimum. However, some small differences may still distort the accurate meaning convey by the original text.

According to (Azmi, Ching, & Norbahyah, et al, 2016), English language differ from Malay language with its unique tenses fields, plural and singular nouns. The tenses for English language involves present, past , past present and future tenses are used to indicate the time frame for an event or situation (Azmi et al., 2016). However, Malay language expresses the time frame with term like ‘telah’ for past event, ‘sedang’ for present event and ‘akan’ for future event, instead of adding suffix to alter the verb. For plural and singular noun, Malay structure differs the plurality of the noun by adding ‘banyak’ (many) or the direct amount to the target noun.

Affixes in both language also harden the process of comprehension and translation of code-switched sentences. For English language, affixes are used to indicate or change the part of tenses or plurality. On top of that, English prefixes such as ‘un’ and ‘dis’ give an opposite meaning to the root word. Malay affixes work differently, they are used to indicate the part-of-speech of the target word. A ‘ke-’ prefix with ‘-an’ suffix often attached to a noun verb. For example, ‘malang’ (misfortune) word itself is an adjective, while ‘ke-malang-an’ (accident) is a noun that indicates an unintended event caused by misfortune. Common translation process would cut off those affixes and obtained their root word to determine the meaning of unknown word. Removal of affixes in both English and Malay words often result in altering the original meaning brought by the sentence. Information loss is expected and inaccurate translation is induced.

## Translation Process

Translation is the process of rendering a text in one particular language to another language(s), keeping the exact information carried by the content. For source and target language that share the same or similar language structure, translation process can be done at its best to produce high-quality translated text. However, languages as the bridges for connecting people and understanding each other owns their very unique structure. Besides the grammar rules and words that build up the language, cultural-related content also impacts the structure of the languages. The indifferent structure between the source and target language induced various gaps in the translation process (Culler, 1976).

General translation process is categorized into two categories, direct translation and indirect translation. Direct translation is applied when the two language structures and concepts and similar. Techniques used in direct translation are borrowing, calque, and literal. Indirect translation is applied when two languages own distinct cultures or structures. The transposition technique tackles the issue of shifting from one grammatical rule to another. The modulation technique decodes the original perspective of a source text to another in the target language. For instance, the last stage of a building holds the same meaning as the top floor of the building. Most frequently used technique should be the equivalence technique, such that linguistics preserves the original meaning of source text by finding equivalent words in the target language. Besides adapting the translated words according to the cultures of the target language, expressing the information at the other point of the document is the compensation technique used in indirect translation. Expansion technique adds words into translated text to preserve the original meaning while reduction technique removes words that produce redundant information.

Current translation methods include human translation and machine translation. Human translation relies on human translators throughout the entire translation process. It encompasses various key translation processes starting with analysing and comprehension of the intended meaning or the context of the source text. Translators then carefully select appropriate translation techniques according to the context or cultural needs. The translation process coverts the source text into targeted language while ensuring the preservation of the original meaning and fluency of the translated text. Finally, a thorough review and post-editing process is carried out to ensure the quality and accuracy of the translation. Manual translation process requires linguistics to conduct long and complex process to process high-quality translation. Correction or enhancement (if needed) is made during post-editing process as the final polishing steps to produce a natural and correct translated text. This translation method performs better in translating content that possess proper tone for audience such as cultural-sensitive content.

Machine translation, as one of the natural language processing (NLP) application, involves machine (software) in the translation process. The translation process is conducted based on programming rules, statistical rules or using neural networks. It has emerged as a valuable tool for speeding up the translation of large volumes of text and promoting cross-language communication. Although machine translation translates faster, the resulting text often needs further tuning from human translators especially when translating marketing or creative text. Lacking of logic sense in machine translation cause the machine translation to produce a translation without considering and understanding the meaning embedded in the source text. Human translators take part in pre-editing or post-editing process to enhance and polish the translated text. Current machine translation techniques still a long way off from producing high-quality translation as great as human translator.

However, human translation and machine translation hit differently in their respective aspect. Trading off with preserving all meaningful contents, human translators spent more time in the translation process to produce high-quality translated text. Considering the time spent and their professionality, the cost hiring a profession human translator is very high. In contrast, machine translation produces the translated text in a relatively short time. If the translation does not involve professional domains that require special understanding in them, machine translator is way more cost-saving if compared to hire a profession human translator for a long period. Machine translation models saved both time and cost in long term expenditure while having memory for key term to be reused.

## Machine Translation

Machine translation is an automated method that performs translation using computer algorithm and technology. The translation process is conducted based on programming rules, statistical rules or using neural networks that analyse the patterns and structure of languages. Machine translation offers high-speed the translation along with large volumes of text data and consistency in translation style and terminology.

In the year of 1949, machine translation concept was introduced by Warren Weaver. Referring to Hutchins report (Hutchins, 2004), automated translation process was successfully introduced on 7th January 1954 by the Georgetown-IBM experiment. The model was built from a dictionary-based model and performed the translation word by word. Due to the issue of very indifferent structure of each language discussed, the resulting translation back then is very poor. Less attention was paid to the development of the field of machine translation. Common machine translation types (as illustrated in Figure 2.1) are Rule-based machine translation (RBMT) and Statistical machine translation (SMT) and finally evolves to Neural machine translation (NMT).

Machine Translation

Rule-based Machine Translation (RBMT)

1950-1980

Statistical Machine Translation (SMT)

1990-2015

Neural Machine Translation (NMT)

2014-Present

Hybrid Machine Translation (HMT)

Regularization

Model Ensembling

Re-scoring with right-left models

Back-translation

types

techniques

Figure 2.3 Type of Machine Translation and its techniques (adapted from (Grundkiewicz, 2018))

In the middle of 1960’s, researchers worked on direct translation method and some started to investigate rule-based machine translation model (RBMT). A RBMT involves running a set of predefined rules from beginning and executing translation processes based on the rules given. Defining and programming these rules are complex and need to be done manually, pretraining the machine is therefore a challenging process. A RBMT model investigates the linguistic information of both source and target languages and establish the rules based on the information retrieved. Generally, common rules generated includes the word order, sentence structure and phraseology by referring to the grammar structures of the languages. With adequate amount of information gained, the model maps source text with an equivalent of translated text in target language that best-described the source text.

A Statistical Machine Translation (SMT) model as one of the state-of-art of MT (Phua et al., 2022), went through massive amount of existing contents and identified the translation patterns existed in the inputs (Yeong, Tan, & Mohammad, 2016). These translational patterns aid in helping the models to identify the translation style when encountered similar contexts in the future. This machine translation model required massive volume of training data for a great-performing translation model, thus, securing high translation quality for resource-rich language. It is excelled and cost-worthy in fields that are more technical. SMT models evolve from word-based translation, to phrase-based translation, hierarchical phrase-based translation and later syntax-based translation. A Hybrid Machine Translation (HMT) type model is a mix of RBMT and SMT. With its translation memory, HMT achieve a better quality in translation result. Finally, Neural Translation Machine evolved current machine translation by employing artificial neural networks to comprehend the pattern and relationships of the languages.

## Neural Machine Translation

With the advancement of deep learning, neural network was introduced into Natural Language Processing (NLP) field for human language analysing and processing. The advancing neural network in NLP technology help to process the relative huge amount of raw language data. Neural Machine Translation (NMT) was proposed around the year of 2014 with the used of deep neural networks in encoding and decoding the text, and soon developed as another state-of-art for machine translation (Phua et al., 2022). Neural networks are a series of nodes that interconnected with each other like human brain nerve. Input data pass through the networks, creating information along the nodes and finally produces output result. These interconnected nodes allow NMT models to learn from huge volume of data and adapt to new inputs.

NMT is the current most-promising machine translation service. Along the development of NMT, researchers claim that neural networks capable of semantic and syntactic captures while translate without hard segmented the source part. NMT offers automated translation by having end-to-end learning that encode and decode the source text. In another word, NMT accept the input sentence as a whole and use the whole sentence to produce the output. It is believed to be able to resolve the problems arise from traditional translation systems and produce high quality translational works. This self-learning model is highly accurate, fast learning and highly flexible. However, neural network is also well-known with its dependency on the data size. Its data-hunger characteristic limited existing study to high-resource language such as English or rule-based and machine learning-based modeling (Fu et al., 2021).

### General Framework

With the aid of neural network, NMT is able to perform translation process by learning from a huge amount of sample text data. The training text data are curated into the form of parallel corpus or word embedding. A parallel corpus consists the pairs of original and desired translated text, while the word embedding is a technique that represents words as vectors, capturing their similarity, syntatic and semantics. However, machines struggle to comprehend human language directly and instead rely on numerical representations. To process text data, human language needs to be transformed into numerical vectors or matrices, which are the fundamental unit of NMT (Phua et al., 2022). NMT utilizes a word embedding to convert the word sequence into a vector.

#### Word Embedding

Word embedding is the word-level vector representation techniques that captures the word similarity, syntatic and semantics representation. It serves as a dictionary for translation model to comprehend the input text. Since neural network only process numerical input, words mapped with relationship information are transformed into vector form. Popular word embedding methods include word2vec and GloVe. A word embedding feed to a neural network is the vocabulary size of that network, as the network will recognized all the words (in vector form) presence in the embedding. Experiments (Goulden, Nation, & Read, 1990) showed the vocabulary size of a native English speaker is around 16000 to 20000 and a non-native speaker is around 14000 (Elgort, 2013; Nguyen & Nation, 2011). Hazenberg and Hulstijn (1996) specially claimed that university students have to master around 10000 to 110000 words to understand a university text while D'Anna, Zechmeister, & Hall (1991) claimed that university students would have a vocabulary size less than 20000.

Character embedding is the character-level vector representation techniques that splits the word into characters. This embedding helps capture the morphologicak features of a word and aids in out-of-vocabulary issues (Hidayatullah et al., 2022). Previous works of Mave et al. (2019), Shekhar et al. (2020) and Gupta et al. (2021), both character embedding and word embedding are applied to train their models.

#### Encoder-Decoder Framework

General framework for NMT translation process is the encoder-decoder architecture (as illustrated in Figure 2.4), introduced by Sutskever et al. (2014). This architecture takes a recurrent neural network (RNN) as encoder to encode the text into fixed-length vectors and as decoder to decode the fixed-length vectors to text. According to Cho et al. (2014), this architecture learns the translation by discovering the conditional distribution over the original text and the translated text. They also emphasize that the length of the translated text need not necessarily match the length of original text. Although models perform relatively well on short sentences without unknown words, the performance of model degrades rapidly as the length of sentence and the number of unknown words increase. Current NMT will apply any variants of RNN for both encoder and decoder.

I ate a slice of cake

English Sentence

Saya telah makan sepotong kek.

Malay Sentence

Neural Network

Neural Network

0.34

0.48

0.13

0.98

0.43

ENCODER

DECODER

Figure 2.4 General framework of NMT: Encoder-decoder architecture

#### Language Identification

Language identification is a crucial step in various natural language processing tasks, including code-switching translation due to the brevity of social media messages and unconventional spelling (Jaech et al., 2016). Language identification plays a crucial role in machine translation by helping to determine the source language of the input text, which is necessary for accurate translation, as different language pairs may require different translation models and techniques (Jain, & Bhat, 2014). Language identification can also assist in handling code-switching scenarios, where multiple languages are used within the same text. By accurately identifying the language, machine translation systems can apply language-specific rules, linguistic features, and translation models to improve translation quality (Jain et al., 2014).

The Computational Approaches to Linguistic Code-Switching (CALCS) community has organized workshops and shared tasks to encourage research on code-mixed text and language identification (Chen et al., 2022). With the aim to improve the understanding and processing of code-switched data, the identification of code-switched languages has been the focus of competitions. For instance, the First Shared Task on Language Identification in Code-Switched Data and the Second Shared Task on Language Identification in Code-Switched Data focused on word-level code-switching identification in Arabic, Spanish-English, and other language pairs (Chen et al., 2022). These shared tasks provide valuable benchmarks for evaluating language identification systems and comparing their performance (Chen et al., 2022).

In conclusion, language identification is a crucial step in decoding code-switched languages to ensure accurate and contextually appropriate translations (Jain et al., 2014). It helps in selecting the appropriate translation resources, such as bilingual dictionaries and parallel corpora, specific to the identified language pair. Neural Network Architecture

#### Recurrent Neural Network (RNN)

Dealing with sequential information where data change with time, RNN is most used to solve common language translation due to its characteristics of self-looping over time. RNN differs itself from other neural network by its ability to memorize the information from input data. The information obtained from previous input are applied to influence the next output. The type of unequal size many-to-many RNN were normally applied to Machine Translation. Although RNN is designed to model sequence at any length and the model size does not increase with input size, the computation grows slower due to its recurrent nature. Thus, the training process would take extra time compared to another neural network.

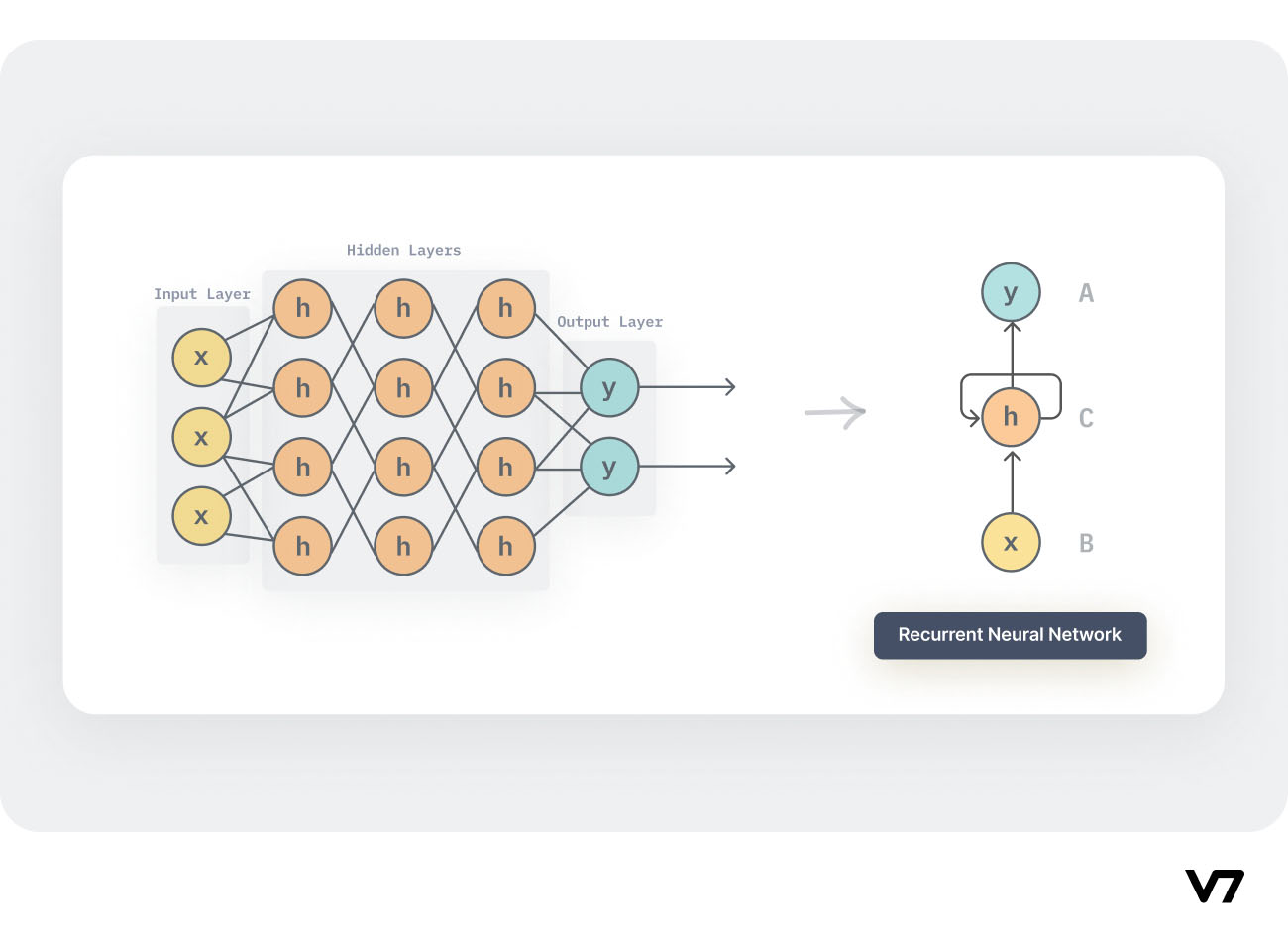


Figure 2.5 Simple RNN (source: https://www.v7labs.com/blog/recurrent-neural-networks-guide)

Besides, RNN remain its weight parameter as the same value across its multiple layers, but still the value will be updated by backpropagation and gradient descent during training process. Weights, w, one of the parameters in neural network modelling, is a numerical value that indicates the importance of the corresponding input. An input with higher weight will bring more impact to the model prediction. The training process takes the error value(s) measured by identifying the difference between predicted and real output. Both backpropagation and gradient descent are aimed to updates the parameters to minimize the error (e) calculated by a cost function. Backpropagation travels backward through time (sequential data) and calculating the negative gradient of cost function for each weight.

A gradient is the changes of the weights (dw). A positive gradient indicates the rising of error value while a negative gradient shows the error is going down. Gradient descent algorithm uses gradient to identified the most appropriate weight by navigating down the cost function. When the cost function is at its minimum point, the error is also in its minimum value. The down-navigation of cost function involved the changes in direction, gradient and learning rate, α. The direction is determined by the gradient while gradient is obtained from backpropagation. The learning rate should be at a rational value. A high rate would fluctuate the result and impact the observation of minimum point while learning too slow eats up computational power and time. The algorithm can be calculated by the equation 1.1.

|  |  |
| --- | --- |
|  | (1.1) |

While having gradient as an important part to minimize errors occurred, the exploding gradients and vanishing gradients as two major problems suffered by RNN models (Phua et al., 2022). Exploding gradients is the multiplication of the gradients with value greater than 1.0 repeatedly, while vanishing gradients is the multiplication of the gradients with value close to 0.0 repeatedly. Zero gradient updates the weight until a very insignificant value causes the algorithm to stop learning. A very larger weight resulted from exploding gradient impact the stability of the model. Minimizing model complexity such as decreasing the number of hidden layers is one of the solutions to these issues.

Long short-term memory (LSTM) is introduced by Sepp Hochreiter and Juergen Schmidhuberat 1997 (Hochreiter, & Schmidhuber, 1997). LSTM increments the RNN structure with gates that immune from the gradient issues and long-term memory units that address the inability of RNN in keeping track of longer term-dependency. Utilizing the input gate, output gate, and forget gate, information flow needed is controlled to aid the prediction. Term-dependency can be better explained with the ability of the model to memorize previous context, that will be helpful in predicting the future content. LSTM experiments by Cho, Merrienboer, Bahdanau, & Bengio (2014) show its ability to deal with longer sentence, best deal with 15-20 words sentence.

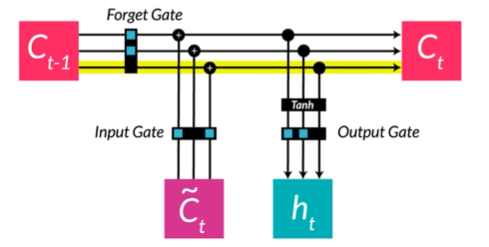


Figure 2.6 LSTM unit

Gated recurrent unit (GRU) as another RNN variant is also a solution to the short-term memory issues. Different from LSTM, GRU adopts two gates to control the information flows, including a reset gate and an update gate. GRU can be trained to keep track of long-term information by update gate and remove irrelevant information by reset gate. A GRU with lesser gates than LSTM performs better with smaller dataset that LSTM.

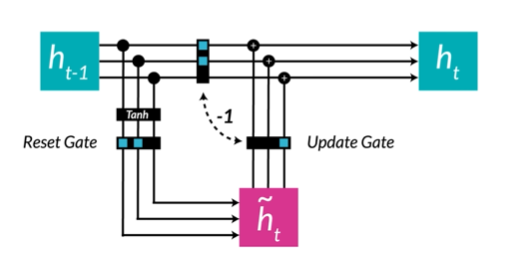


Figure 2.7 GRU unit

However, feed-forward typed recurrent neural network only considers previous input (word) when translating the current word. From the perspectives of language grammar, the structure or the usage of some words also depend on the next word(s). There goes the introduction of bidirectional LSTM (Bi-LSTM) and Bidirectional GRU (Bi-GRU). Bi-LSTM is made up of bi-directional RNN (BRNN) and LSTM Gating. Although BRNN was introduced by Schuster et al at 1993 (Schuster & Paliwal, 1997), it only gained popularity very recently with the advanced in the field of deep learning. Generally, the BRNN will process the input forward and backward to obtain the overall dependency. In another words, the ‘future’ data are used with past and current data to predict the current state.

General encoder-decoder approach fixed the input length by encoding the input into a fixed length vector. Bahdanau, Cho, & Bengio (2016) expressed the limitation of a fixed length encoder on coping with longer sentences. They claimed that the model should not be limited to be focused on the previous or next information (input word). They proposed the model that learn from various alignments to identify the relevant-word(s) to be focused on when translation take place by a special unit named attention mechanism. Based on an alignment plot (illustrated in Figure 2.6), the whiter the alignment box, the more focus is put on to the word when translating current word. Some translation only aligned to the original word itself, like having ‘public’ to ‘awam’. While some translation would need multiple alignments with different focus level to produce the translated word. Observing Figure 2.6, a diagonal line shows translation is done in sequence while an anti-diagonal line indicates the words are flipped during the translation process. This technique advanced previous encoder-decoder mechanism by 30 words. Google translate adopted this method but apply 8 layers of LSTM in both encoder and decoder (Wu, Schuster, Chen, et al. 2016).

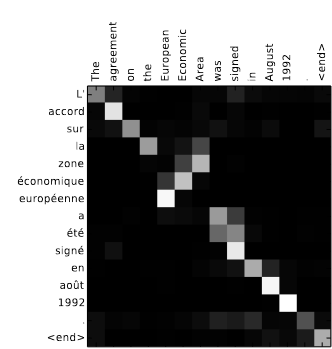


Figure 2.8 Alignment sample (Bahdanau et al., 2016)

I ate a slice of cake

English Sentence

Saya telah makan sepotong kek.

Malay Sentence

BRNN

*encoder*

RNN *decoder*

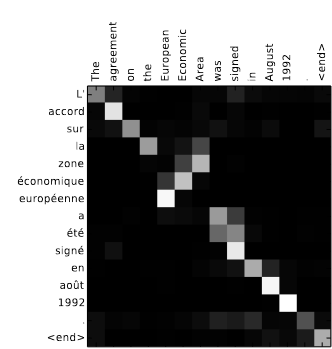
0.34

0.48

0.13

0.98

0.43



Attention

Figure 2.9 Encoder-decoder architecture with attention mechanism

### State-of-art

Language translation models based on encoder-decoder machine translation have revolutionized the field of natural language processing. These models have achieved remarkable success in various applications, including Google neural machine translation (GNMT), encoder-decoder with attention, and Bidirectional Encoder Representation from Transformers (BERT). In this section, a comprehensive overview of these language translation models will be elaborated with their key features, advantages, and contributions to the field.

Google switched to Google Neural Machine Translation (GNMT) System in 2016 with the claims that a lesser engineering and design efforts were required while a leap and bounce in speed and accuracy (Wu et al., 2016). This model utilizes a sequence-to-sequence architecture with attention mechanisms. GNMT has been trained on a large-scale dataset and has demonstrated impressive translation performance, outperforming previous statistical machine translation systems. It has also been integrated into Google Translate, enabling users to access high-quality translations across multiple languages.

In 2017, the concept of "Attention is all you need" was introduced (Vaswani et al., 2017). This concept of attention mechanisms hit differently in the evolution of deep learning. Attention mechanism with encoder-decoder structure build up the whole model architecture. Taking the whole sentence as input to produce a single output is a form of sequence-to-sequence problem. This model uses recurrent neural networks in both encoder and decoder to predict sequence-to-sequence problems. An encoder accepts input and encapsulates it into internal vector(s) while decoder convert the encoded vector(s) into desired output sequence. The attention mechanism allows the decoder to focus on different parts of the input sequence during the decoding process, improving the translation quality. An input is mapped by the encoder to produce an embedding that includes positional encoding and, most crucially, an attention mechanism. A feed-forward neural network is then used by the decoder to produce an output. The flexibility of the attention mechanism allows for the introduction of other parameters to produce additional features or relationships between the data. This model has been widely adopted and has become a standard approach in machine translation tasks

Considering Bidirectional Encoder Representation from Transformers (BERT) as one of the state of art in NLP process, it excels in named entity recognition (NER), language identification and machine translation (Devlin, Chang, Lee, & Toutanova, 2019) in term of high BLEU score and high accuracy. The BERT model that built based on the attention is all you need model can comprehend the meaning of complicated human languages. However, unlike traditional encoder-decoder models, BERT is a language representation model that learns contextualized word embeddings by training on large amounts of unlabeled text data and possess the ability to bi-directionally understand the text. Its ability to capture bidirectional context has significantly improved the translation quality and accuracy. BERT has achieved state-of-the-art performance in various natural language processing tasks, including machine translation.

Language translation models based on encoder-decoder machine translation, such as encoder-decoder with attention, BERT, and GNMT, have significantly advanced the field of machine translation. These models have introduced innovative techniques, such as attention mechanisms and bidirectional context modeling, which have greatly improved translation quality and accuracy. The success of these models can be attributed to their ability to effectively capture and utilize contextual information in the translation process. As research in this field continues to evolve, it is expected that further advancements will be made, leading to even more accurate and reliable translation models.

## Performance Metric

**BLEU score** (Mathur, Baldwin, & Cohn, 2020). The Bilingual Evaluation Understudy (BLEU) metric assesses the similarity of machine-translated text to a set of reference translations by comparing their n-grams. It is used to automatically evaluate machine-translated texts. BLEU calculates precision, which is the ratio of the number of n-grams in the machine translation that appear in the reference translations, to the total number of n-grams in the machine translation. It also incorporates a brevity penalty to account for translations that are shorter than the reference translations. BLEU scores range from 0 to 1, with higher scores indicating better translation quality. However, the evaluation of BLEU and other automatic metrics can be sensitive to the translations used for assessment, leading to potentially misleading conclusions about their efficacy.

**TER score** (Mathur et al., 2020). The Translation Edit Rate (TER) is a machine translation error metric used to measure the number of corrections required to modify the output of a system to one of the samples. It measures the edit distance between the machine translation output and the human reference translation. The edit distance is calculated based on the number of edits required to transform the machine translation into the reference translation. Edits can include insertions, deletions, and substitutions of words or phrases. TER aims to capture the differences in word order, word choice, and grammatical errors between the machine translation and the reference translation. A lower TER score indicates a higher similarity between the machine translation and the reference translation, indicating better translation quality. TER has been widely used in machine translation research and evaluation, especially for comparing different translation systems and tracking their performance over time. TER is one of the traditional metrics used in machine translation evaluation, but studies have shown that it has limitations in capturing semantic similarity between translations and reference translations (Lee, Lee, Moon, Park, Seo, Eo, Koo & Lim, 2023).

**Word Error Rate (WER)** (Lee et al., 2023).WER is an automatic evaluation metric used to measure the performance of machine translation systems. It calculates the percentage of words in the machine translation output that differ from the reference translation. WER is computed by dividing the total number of word-level substitutions, deletions, and insertions by the total number of words in the reference translation. WER is widely used in machine translation research to assess the accuracy and quality of translations. However, studies have shown that WER, along with other traditional metrics like BLEU and TER, may not capture the semantic similarity between machine translation outputs and human reference translations effectively.

**METEOR** (Lee et al., 2023). METEOR (Metric for Evaluation of Translation with Explicit ORdering) is an automatic evaluation metric designed for machine translation systems. It measures the quality of machine translations by comparing them to human reference translations. METEOR considers various aspects of translation quality, including precision, recall, and alignment of words and phrases. It uses a combination of unigram matching, stemming, and synonym matching to calculate the similarity between the machine translation and the reference translation. METEOR also considers the order of words and phrases, giving higher scores to translations that have a similar word order to the reference translation. Additionally, METEOR incorporates a penalty for over-generation or under-generation of words in the machine translation. The final METEOR score is a weighted combination of these different components, providing a comprehensive evaluation of translation quality. METEOR has been widely used in machine translation research and evaluation, and it has shown to have a higher correlation with human judgments compared to other metrics like BLEU.

## Previous Research

### Code-switching: Data synthesis

Research by Song et al. (2019) investigates a data augmentation method for code-switching in NMT, where source phrases are replaced with their target translations, allowing the model to learn lexicon translations by copying source-side target words. The author claims this approach consistently improves translation of constrained words without negatively impacting unconstrained words.

For low-resource language, Cross-lingual Natural Language Inference (XNLI) and BiLSTM model was used across several low resource languages (Jesin et al., 2022; Lal et al., 2019; Mukherjee et al., 2019). In dealing with low-resource dataset, generation of synthetic data were done through parse tree (Samanta, Ganguly, & Chakrabarti, 2019) and phrase tables (Yang, Christopher, Huang, 2020). Kuwanto, Akyürek, Tourni, Li, Jones, & Wijaya (2021) proposed comparable data (parallel corpus) mining by using bilingual dictionary to address the issues of data lacking in truly low-resource language. The authors extract online comparable sentence pairs by obtaining comparable data from the bilingual dictionary. More synthetic processes will be explored in the continuous literature review.

### Code-switching: Language model

The application of computational study in code-switching text is a recently developing topic. Even though computational code-switching was initially introduced in the early 1980s, research on it made very modest progress. In late 2010 and early 2020, several language models were developed for code-switch context including bidirectional long short-term memory (BiLSTM) for Maori-English pair code-switched detection (Jesin et al., 2022), BERT for Persian-English pair sentiment analysis (Sabri, Edalat, & Bahrak, 2021), Character-LSTM for English-Hindi pair sentiment analysis (Lal, Kumar, Dhar, Shrivastava, & Koehn, 2019) and multilingual BERT that support more than 100 languages (Devlin, Chang, Lee, & Toutanova, 2019).

Most of the previous research involved the investigation in the field of language identification (Jesin et al., 2022; Lal et. al., 2019), code-switch detection (Jesin et al., 2022; Vissamsetti, Pravallika, Jadhav, & Kommanti, 2022), sentiment analysis (Sabri et al., 2021; Lal et. al., 2019; Mukherjee et. al., 2019, Vissamsetti et al., 2022; Adilazuarda et al., 2022; Patwa, Aguilar, Kar, Pandey, PYKL, et al, 2020), automatic speech recognition (Weller, Sperber, Pires, Setiawan, Gollan, Telaar & Paulik, 2022; Lee, et al., 2020; Winata, Madotto, Wu, & Fung, 2019) and etc.

Table 2.3 Previous Research

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Research | Application | Language pair | Proposed Solution | Limitation |
| (Jesin et al., 2022) | (i) Language Detection  (ii) Code-switch Detection | Maori-English | BiLSTM with embeddings models; | Low-resource language will affect the quality of the bilingual embedding |
| (Adilazuarda et al., 2022) | (i) Sentiment analysis  (ii) Emotion classification | Sundanese-English;  Javanese-English;  Malay-English | IndoRobusta with 3 tuning strategy:  i. code-mixed only  ii. two-steps  iii. joint training | Tuning with code-mix data only impact the model performance on monolingual translation; Low quality generated code-mix sentence brings negative impact to the model. |
| (Sabri et al., 2021) | Sentiment analysis | Persian-English | (i) Yandex and dictionary-based translation techniques (ii) BERT pretrained embeddings & translation model | Ensemble three models, hard to implement as real-time based. |
| (Kasmuri et al., 2020) | Sentence segregation | Malay-English Code-Switch | Rule-based segregation technique | Adding rules to address more languages increase the complexity of the model. |
| (Zabha et al., 2019) | Sentiment Analysis | Malay-English | Lexicon Approach | Not promising result due to non-standard words in dataset. |
| (Lal et al, 2019) | Sentiment Analysis | English-Hindi | CMSA:  (i) CNN model for sub-word level representations  (ii) Dual Encoder Network  (iii) Character-LSTM | The use of relatively small dataset. |
| (Mukherjee et al., 2019) | (i) Sentiment classifier;  (ii) Word embedding generation | Hindi-English | Sentiment Classification on Code-Mixed Text (ACCMT) | Lack of suitable word embedding |

### Code-switching: NMT language model

Currently, no Malay-English code-switching translation system has been specifically investigated. Most of the previous Malay-English code-switching research involved the investigation in the field of emotion classification (Adilazuarda et al., 2022), sentiment analysis (Adilazuarda et al., 2022; Zabha et al., 2019), and sentence segregation (Kasmuri et al., 2020). However, various NMT researches had been carried on other language pairs including Vietnamese-English (Nguyen, Mayeux & Yuan, 2023), Spanish-English/French-English (Xu, & Yvon, 2021), Hindi-English (Appicharla, Gupta, Ekbal, & Bhattacharyya, 2021; Jadhav et al., 2022) and English-German/English-French/Chinese-to-English (Son, et al., 2019).

A paper suggested a novel bilingual attention language model (BALM) to model both the monolingual as well as the cross-lingual (ZH-EN) sequential dependency. (Lee & Li, 2020). The proposed model uses a quasi-translation objective to learn the bilingual context from a parallel corpus. The attention mechanism encodes the mixed context well. While aligning words in different languages without the needs of positional embedding or aligner, it can generate code-switched sentences too. The authors claimed that the model achieves state-of-the-art performance on the SEAME code-switch database by reducing the perplexity of 20.5% over the best-reported result. This work differs from the other model with its bilingual attention mechanism that learn from bilingual parallel data without the need for code-switched sentences.

Jadhav et al. (2022) proposed a pipeline of 4-stages translation process to decode the Hinglish (Hindi-English) code-switched text without the use of parallel corpus. Language tagging stage helps in segmentize the text to language sub-units and back-transliteration from roman Hindi to Hindi is done to Hindi language sub-units while NMT is done to embedded English sub-units. Final stage re-joins the sub-units into monolingual Hindi text. The adding of language identification stage allows their model (0.737) advanced GNMT (0.496) by 0.241 BLUE score.

A bunch of works improve the code-switching translation models with the language identification mechanism. Instead of neural network, Mave, Maharjan & Solorio (2018) proposed CRF machine learning model used for sequence labelling tasks for language identification in code-switched social media text. They run experiments with different combinations of hand-crafted features and find that the CRF model outperforms the neural network-based models for both Spanish-English and Hindi-English datasets. The paper investigates different code-switching metrics to understand the code-switching patterns in the Hindi-English and Spanish-English language pairs. The three metrics used in the paper to understand the code-switching patterns are the M-Index, H-Index and C-Index.

A paper in 2021 proposes switch-GLAT, a non-autoregressive multilingual machine translation model with a code-switch decoder that employs token-level language tags which can be used to generate code-switched translations and performs code-switch backtranslation (Song, Zhou, Qian, Xu, Cheng, Wang, & Li, 2021). Song et al. (2021) claimed that the language tags on both the source and target sides can help the model learn to align the input and output sequences correctly. Their proposed parallel decoder architecture outperforms the multilingual Transformer with as much as 0.74 BLEU improvement and 6.2x faster decoding speed in inference.

The research conducted by Gupta, Raghuwanshi, & Tripathi in 2021 introduced an innovative approach for language identification. Their methodology, based on a deep learning model utilizing Bidirectional Long Short-Term Memory (BLSTM), effectively captures information at both the word and context levels. This model determines the language origin of a word within a sequence by considering the words that precede it in that sequence. Remarkably, the proposed model achieved an impressive accuracy rate of 91.2% for identifying languages in code-mixed text within the context of social media. The authors expressed optimism about the future potential of this methodology, particularly in the realm of machine translation.

From the review above, most researches showed improvement in translation process with unique language identification process. Only some novel model enhancement researches were found, include a bilingual attention language model (BALM) to model both the monolingual as well as the cross-lingual (ZH-EN) sequential dependency (Lee et al., 2020), switch-GLAT, a non-autoregressive multilingual machine translation model with a code-switch decoder that employs token-level language tags which can be used to generate code-switched translations and performs code-switch back-translation (Song et al., 2021) and a pipeline of 4-stages translation process to decode the Hinglish (Hindi-English) code-switched text without the use of parallel corpus (Jadhav et al., 2022). However, according to Chen, et al. (2020), current NMT technology is insufficient to cope with code-switching as most NLP models are developed to address monolingual situations.

## Challenges

**Code switching**. With the rising of data mining and analysing, code-switching phenomenon poses formidable challenges to NLP processes. Primarily, code-switching itself is suffers from a lack of sufficient parallel dataset (Gahoi, Duneja, Padhi, Mangale, et al., 2022) due to complexity of collect and synthesis of code-switched text, while most machine translation would require a vast amount of parallel corpus in its training process (Jadhav, Kanade, Waghmare, Chandok, & Jarali, 2022). With various dialect and unrecorded written words, Malay language drew very limit attention from NLP fields and is identified as a low-resource language (Liang, Li, Fu, & Lin, 2020). A low-resource language often refers to those language that owns very little labelled resource while some would borrow labelled dataset from languages that share high language similarity, the homologous languages (Fu et al., 2021). Indonesian language as one of the homologous language for Malay language is the most ‘borrowed’ datasets. However, no existing Malay-English code-switched parallel corpus is found.

Due to the casual use of code-switching language, the text is expected to be informal. When an international institute intends to conduct automated analysis on current reviews or trends in a specific field on the Internet, the presence of code-switching language on informal platforms, such as social media, can lead to distorted information for data mining and data analysis process. According to Siti Nor Allia, & Sabrina (2020) and Hidayatullah et al. (2022), non-standard words commonly present in the code-switched text as both of them used in informal events. Siti Nor Allia et al. (2020) special mentioned the normalization of Malays non-standard words also received minimum attention. Various improvement should be further done on all aspects as current rules have limit ability in normalizing the entire Malays vocabularies (Siti Nor Allia et al., 2020). Non-standard words include those with irregular spelling, abbreviation, exaggerated words and characters combinations that mix numbers or special character with alphabets.

Referring to section 2.2.1.3, different languages often have variations in syntax, grammar rules, and word order. When code-switching occurs, the syntactic structure and grammar may change abruptly, making it difficult for NMT models to maintain coherence and produce accurate translations. Besides differences presences in the language structure, ambiguity further impacts the translation of code-switched texts (Hidayatullah et al., 2022). The existence of homonyms, a single word that bring more than one meaning harden the comprehension process, let alone the word that exists in both language but with totally different meaning. For instances, the word ‘air’ means water in the Malay language while it means oxygen in the English language; the word ‘main’ means play in the Malays language but leading in the English language. A translation model would suffer confusion to classify the word as an English word or Malay word. A total disrupted information would be translated if the translation model failed to interpret the true meaning indicate in the text. Additionally, the translation model was misled about the language and meaning of a word by intra-word code-switching.

**Machine Translation**.These algorithms are trained on vast amounts of bilingual data to learn patterns and associations between words and phrases in different languages. It is inevitable that the raw input text data would affect the quality and performance of a machine translation model. A raw input data often come with impurities, which indicates the non-standard word used. Non-standard words refer to the use of slang, dialects, short-form, misspelling, and etc. Besides, text that do not follow proper grammatical rules falls also to the category of non-standard words. These non-standard words impact the ability of the model to interpret and translate the whole input text badly, resulting in an incomplete or misleading translation.

Based on the references, it can be concluded that deeper neural networks, such as those used in the Transformer model (Vaswani et al., 2017), have shown improved performance in modeling complex problems related to grammar and semantics in languages. However, label-rich languages such as English add non-linearity to sequence model or Deep neural network. Deeper network helps better model complex problems: grammar and semantic of the langauge. However, the challenge of reasoning with sequences remains, as the time dimension adds complexity to identifying the data shape at each step (Sutskever, 2014).

Another current gaps in NMT in the code-switching area is the lack of a unified approach to handling code-switching (Yang, Hu, Han, Huang, & Ju, 2020). This means that different NMT models may handle code-switching in different ways, which can lead to inconsistencies in the output (Yang et al.,2020). One approach to addressing this challenge is the use of pre-training techniques. For example, the CSP (Code-Switching Pre-training) method has been proposed to enhance the ability of fine-tuned NMT models in handling code-switching input (Yang et al., 2020). By pre-training the model on code-switched data, CSP aims to improve the model's understanding and generation of code-switched sentences. Another approach is to incorporate alignment-based methods. For instance, a study used an alignment-based approach to train an NMT model on a gold corpus augmented with a synthetic code-mixed parallel corpus (Appicharla et al., 2021). This approach does not explicitly mark tokens for code-switching but relies on the alignment of the parallel corpus to capture code-switching patterns. By incorporating these approaches, NMT models can better handle code-switching input and improve the quality of translation in code-switching scenarios. However, Malay-English code-switch pairs suffers from insufficient code-switch parallel corpus data.

## Discussion

Neural machine translation (NMT) as a potential solution to overcome the limitation of NLP model to comprehend the natural language. With the promise to address many of the shortcomings of traditional phrase-based translation systems, NMT is an end-to-end learning approach for automated translation. Unfortunately, both in training and translation inference, NMT systems are known to be computationally expensive. The majority of NMT systems also struggle with non-standard words. These problems bring challenges for NMT to be used in real-world deployments and services, where accuracy and speed are crucial. Even said, GNMT is still capable of making serious mistakes that a human translator would never commit, such as dropping words, mistranslating proper names or uncommon terminology, and translating sentences without taking the surrounding paragraph or page into account. We still have a great deal of work to do to better serve our users.

Machine translation accepts text data as the input to produce the translated text in target language. It is inevitable that the raw input text data would affect the quality and performance of a machine translation model. A raw input data often come with impurities, which indicates the non-standard word used. Non-standard words refer to the use of slang, dialects, short-form, misspelling, and etc. Besides, text that do not follow proper grammatical rules falls also to the category of non-standard words. These non-standard words impact the ability of the model to interpret and translate the whole input text badly, resulting in an incomplete or misleading translation. Normalization process is an urged to standardized the input text data for a better translation quality. Normalization rules provide guidance for translation model to standardized the input text data. It is a set of rules that records possible polymorphism of the standard text and its corresponding standard word. Referring to the rules, machine translation model could learn to interpret and understand the non-standard word, and produce a better translation.

Next, normalization process as an important aspect of machine translation to improve translation quality. Reviewing on previous studies, most of studies in Malay-English mixed-language mostly focus at sentiment analysis and data construction. Unnormalized data such as slang words, short form words, and misspelling still impact badly in model's overall performance. Kasmuri et al. (2019) claimed that segregation of the code-switch sentences and monolingual sentences is needed before a model can differenciate and analyze the code-switch sentences. Besides detecting Malay-English code-switched text by using Google (GNMT) and Microsoft Azure’s cloud-based service, previous studies showed BERT perform good in detecting this code-switched text and show relatively high-promising result, therefore, BERT is selected as the baseline model of this study. In addition, NMT model that require a massive amount of training data still serve as the top issue for Malay-English mixed-language text analysis.

According to (Siti Nor Allia et al., 2020), the normalization of Malays non-standard words received minimum attention. Various improvement is should be further done on all aspects as current rules have limit ability in normalizing the entire Malays vocabularies (Siti Nor Allia et al, 2020). Normalization of non-standard word by Siti Nor Allia et. al. (2020) collected over 2848 non-standard words and manually mapped to their corresponding standard words. According to them, building up the repository of normalized words is the most challenging. Further review will be done to understand the field better.

## Research Gap

1. The availability of parallel corpus dataset for Malay-English code-switched text.
2. Current language identification or dictionary-based approach works very limited with the presence of code-switching.
3. The ambiguity and irregular phonetic typing from code-switching bring bad impact on the syntax and sentiment analysis.
4. Most NLP techniques are designed for monolingual situation.
5. Impurities in raw input text data would affect the quality and performance of a machine translation model..

## Chapter Summary

This chapter showed useful information obtained from some significant studies related to the important elements in the research. Current challenges and trends in the field are stated clearly. The details of the research framework will be explained in the next chapter.

# RESEARCH METHODOLOGY

## Introduction

Preceding studies related to code-switch text analysis had been reviewed in the previous chapter. From the review, it is suggested that the informal wordings in source text should be addressed to improve the performance of code-switch translation model. The limited resource in Malay-English code-switch text is believed to be addressed by both dataset-tuning and model-tuning. In this chapter, detailed research methodology will be discussed and explained. The framework of this research is separated into 5 phases corresponding to objectives listed in chapter 1. A framework flowchart is demonstrated to explain each phase in details. The dataset used in this research and the performance measurements are also discussed in this chapter.

## Research Framework

This research will be conducted by following a five-phase framework. In the first phase, literature review and problem formulation were conducted to identify the research area and objectives. Data collection and pre-processing process will be conducted in phase 2 to produce a high-quality Malay-English code-switch dataset. Preliminary study on model development and enhancement are the third phases in the framework to develop an effective enhancement process in code-switch neural machine translation. In the fourth phase, the result evaluation and the discussion process are conducted to validate and verify the performance and reliability of the translated text for further analysis. The framework end with documentation as the final process. A flowchart (Figure 3.1) is displayed to illustrate the research framework and Table 3.1 displayed the detailed research design. Gantt chart is attached in Appendix A.

**Tools / Techniques**

**Output**

Malay-English

Code-switch Text

Homonyms

Dictionary;

Optimized

translation

method

for each

code-switch

type

Ambiguity

Dictionary;

Language

identification

method

Research Area

& Planning

Beautifulsoup

web scrapping

Code-switch type identification method

RO1

RO 2

**Phase 1:**

**Literature Review & Problem Identification**

* Resource Collection
* Literature Review
* Problem Identification
* Research Planning

**Phase 2:**

**Data Collection & Pre-processing**

* Data Collection (Web-Scrapping)
* Data Pre-processing

**Phase 3a:**

**Language Identification**

* Develop Malay-English ambiguous word dictionary
* Design and development language identification model

**Phase 3b:**

**Code-switched Type Identification**

* Join words with identical language tag
* Tag segments with label
* Classify segments into code-switchedtypes

**Phase 3c:**

**Segment-Based Translation**

* Develop Malay-English homonyms dictionary
* Optimize translation process for each code-switch type
* Performance Evaluation
  + - Baseline Model Comparison
    - Testing with other low-resource datasets
* Crowd-sourcing Evaluation
* Crowd-sourcing Evaluation

**Phase 4:**

**Result Evaluation & Discussion**

* Output Analysis & Validation
* Discussion

**Phase 5:**

**Documentation**

* Thesis Writing

RO1

Figure 3.1 Research Framework

Table 3.1 Research Design

| Research Objective | Phase | Theme | Research Question | Resources needed | Methodology / Activity | Performance Measurement | Baseline for Comparison | Phase Output |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| - | 1 | Literature Review | - | - | 1. Resource collection  2. Review articles from reliable resources such as Scopus or Web Of Science. | - | - | Domain Knowledge |
| - | Problem Identification | - | - | 1. Problem Formulation  2. Research Planning | - | - | Research Aims and Problem Area |
|  |  |  |  |  |  |  |  |  |
| - | 2 | Data Collection & Pre-processing | - | Online forum; Web scrapping technique: Beautiful-soup | 1. Data Collection by web scrapping  2. Data pre-processing | - | - | Malay-English Code-switch Questions |
|  |  |  |  |  |  |  |  |  |
| To identify and organize the ambiguity present during language identification and translation process. | 3(a) | Language Identification | How can we effectively identify the primary sources of ambiguity that emerge during language identification and translation processes in code-switched scenarios, and systematically categorize and structure the identified sources of ambiguity to enhance the effectiveness of the language identification and translation processes? | Malay-English Code-switch Dataset;  Ambiguity word list | 1. Preliminary Study  2. Develop Malay-English ambiguous word dictionary  3. Design and development language identification model | - | - | Malay-English Ambiguity Word Dictionary;  Language Identification  Method |
|  |  |  |  |  |  |  |  |  |
| To proposed an effective identification method that identify potential code-switch segments from code-switch text data | 3(b) | Code-switch Type Identification | How can we detect and classify the distinctive linguistic features and patterns of code-switched segments within Malay-English code-switched text data? | Code-switch data with language tag | 1. Join words with identical language tag  2. Tag segments with label  3. Classify segments into code-switched types | - | - | Code-switch type identification method |
|  |  |  |  |  |  |  |  |  |
| To develop a translation process optimized for each type of code-switched segment. | 3(c) | Segmentize Translation | How can a translation process be tailored and optimized to effectively address the unique characteristics of each type of code-switched segment, resulting in more accurate and contextually appropriate translations? | Dataset with code-switch type tag;  Malay-English homonyms word list | 1. Develop Malay-English homonyms dictionary  2. Optimize translation process for each code-switch type  3. Performance Evaluation  - Baseline Model Comparison  - Testing with other low-resource datasets  4. Crowd-sourcing Evaluation | BLEU metric, accuracy, and n-grams | BERT, GNMT | Translated code-switch text;  Optimized translation method for each code-switch type |
|  |  |  |  |  |  |  |  |  |
| - | 4 | Result Evaluation & Discussion | - | - | 1. Analyse the evaluation result  2. Discuss its validity | - | - | Conclusion on contribution of proposed model. |
|  |  |  |  |  |  |  |  |  |
| - | 5 | Documentation | - | - | Systematically document all relevant information into the thesis | - | - | Thesis |

### Phase 1: Literature Review and Problem Formulation

Comprehend literature review in the field of code-switch language translation, language identification, and neural machine translation will be first conducted to gain more information and insight for the study. The problem formulation stage involves defining the research questions, setting research objectives, and formulating research framework. It helps to ensure the direction and scope of the research project. The purpose of this phase is to gain a deeper understanding of the research area, identify gaps in the existing literature, and formulate research questions and hypotheses that can be addressed in the subsequent phases of the research project.

### Phase 2: Data Collection and Pre-processing

Data collection and pre-processing are critical stages in the development of a machine learning-based system that aligns with the research objectives and problem statement. These stages will be carried out concurrently with the preliminary study to ensure that the dataset produced fits the proposed framework. Pre-processing steps are essential for creating a clean, error-free, and consistent dataset. The process involves tasks such as data extraction, cleaning, transformation, integration, and normalization. The quality of the dataset greatly affects the performance of the machine learning-based system. Hence, this phase is crucial for producing a high-quality dataset that aligns with the research objectives and problem statement.

### Phase 3: Preliminary Study and Model Development

This phase involves an in-depth study of the current state-of-the-art models and techniques used in the field to identify potential gaps and limitations. Based on this study, a new model will be proposed and developed to address the identified gaps and limitations. The proposed model will be designed and optimized to achieve high accuracy, efficiency, and performance. Explorations of various machine learning models will be done for each of the three processes, namely language identification, code-switch type identification and finally segment-based translation.

Preliminary study on the existing model and parameter-tuning will be conducted next to get an overview and enhance the proposed framework. After having all necessary information ready, model development process will be conducted in rigorous manner. The development process includes translation model of code-switch text, tuning of the translation model and baseline model development. Experiments and investigation will be done on the development phase including suitability of algorithm used, parameter tuning and hyper-parameter tuning. Evaluation including baseline will be conducted on the model proposed to ensure its effectiveness and efficiency of the proposed model in achieving the research objectives and addressing the problem statement.

### Phase 4: Result Evaluation and Discussion

In this phase, the results of these experiments in previous phase will be meticulously analysed and discussed in detail. The discussion will highlight the strengths and limitations of the proposed framework. Potential limitations of the model will be identified and suggestions for improvement will be provided. Additionally, this phase provides an opportunity to compare our proposed model with existing models in the literature and identify any significant improvements achieved. This phase is critical to validate the effectiveness and performance of the proposed model and to provide insights for future research directions.

### Phase 5: Documentation

The Documentation phase encompasses the finalization and organization of the entire research process, including the creation of technical reports, academic papers, and an open-source code repository. A complete and detailed thesis will document every aspect of the research, from problem formulation to result evaluation and discussion. It will include a clear and concise description of the proposed framework, the data collection and pre-processing steps, the model development and enhancement, and the result evaluation and discussion. Additionally, it will provide comprehensive information on the datasets used, the experimental setup, and the evaluation metrics employed, to ensure reproducibility and facilitate future research in the field.

## Tools and Platforms

For the data collection, pre-processing process and the development of the proposed model, Google Colab with python kernel 3.6 will be utilized. Google colab is the best platform for conducting initial experiments as it is a free online cloud-based Jupiter Notebook environment with zero setup needed. Besides, it provides the various computing resources including GPU and TPU services for deep learning model trainings. Document accessing and storing is extremely convenient as it can connect to both google drive and local drive.

## Proposed Solution: Experiment Design

The preliminary study of this research has indicated limited development in a proper yet unified code-switch type identification mechanism for Malay-English code-switch situation. To address the problems stated in the previous section, we have planned to experiment and investigate the combination of language identification, code-switch type identification and translation with deep learning; by using a Recurrent Neural Networks (RNN) model to address the Malay-English code-switched translation. The proposed framework will be conducted by following a 3-phased framework (as illustrated in Figure 3.2). The first phase will explore a word-level language identification mechanism. The following phase investigates code-switch type identification towards Malay-English code-switched text. The final phase will focus on the neural machine translation process, where exploration will be conducted around the appropriate translation process for each code-switch type. RNN served as the neural translation model to assist in the translation process.

Malay-English Code Switch Parallel Corpus

Phase 1: Language Identification

Phase 2: Code-switched Type Identification

Phase 3: Segment Translation

**Language Identification**

*word-level identification*

{"Cs-text": " Job-nya hari ini adalah menjaga kedai runcit di end of

the road.",

"Malay": "Kerjanya pada hari ini adalah menjaga kedai runcit di

hujung jalan.",

"English": "His job today is shop-keeping the grocery store at the end

of the road."}

**Code-switch Type Identification**

Split text into segment by language

Classify the segment into code-switch type

[Job-nya]IW [adalah menjaga kedai

runcit di]AM [end of the road]AE.

[Job-nya]M-E [adalah menjaga kedai

runcit di]M5 [end of the road]E3.

**Translated Text**

Malay or English

Homonyms Dictionary with POS Tagging and relevant words

{"anggur": {"meaning-1": {"NN": "grape"},

{"related": {"M": "manis", "E": "sweet"}},

{"meaning-2": {"NN": "vines"},

{"related-1": {"M": "panjat", "E": "climb"},

"related-2": {"M": "tumbuhan", "E": "plant"}}}

**Segment Translation**

Calque

(C)

Intra-word

(IW)

Intra-sentential

(AM|AE)

*Process only the embedded segment(s)*

Inter-sentential (EM|EE)

*Process only the segment not in desired language*

**RNN Machine Translation**

* *with multi-head attention focusing on cs tag, POS tag and relevant words*

**POS Tagging**

**Ambiguous Word Filtering**

Ambiguous Language Dictionary

{"air": {"Malay", {"related-1": {"M": "minum", "E": "drink"},

"related-2": {"M": "haus", "E": "thirsty"},

"related-3": {"M": "jernih", "E": "clear"}}},

{"English", {"related-1": {"M": "nafas", "E": "breath"},

"related-2": {"M": "segar", "E": "fresh"}}}}

*train*

**Malay-English Code-switch Text**

*train*

*train*

Job-nya\M-E adalah\M menjaga\M kedai\M

runcit \M di\M end\E of\E the\E road\E.

Proposed Framework

Training

Figure 3.2 Proposed Framework

### Research Activities: Data Collection

Export collected data as csv

Start

Build web-scrapping to fit with desired page

End

Scrap all dataset from DoctorOnCall

Figure 3.3 Data Collecting Flowchart

As mentioned in section 1.2, the Malay-English code-switch parallel corpus is not found in current open access database. Exploration on various online forum and social media platform has been done for obtaining Malay-English code-switch text. Online blog is not included as most of the blogs were written in formal language. Code-switching condition is rare in the platform, thus, other platforms are explored. For online famous forum like Lowyat and Cari Forum, Out-of-Vocab (OOV) issue is at high risk as the information posted in the forum is too wide, even in a categorized sub-forum. Their thread-form reply also raises issues for web-scrapping and pre-processing process. Social media platform, such as Twitter, Facebook and Instagram, suffers from relatively-short text found. Curate text data based on language specifications is a challenges in scrapping from these platform. From the lessons learnt, online forum that focus on a certain domain is explored.

When focusing to general practitioner forum, most forums are charged for one-to-one personal conversation. No data is exposed to public. Finally, DoctorOnCall as an open online forum, allows users to ask questions and receive responses from a team of qualified doctors who have years of experience treating patients and are registered with the Malaysia Ministry of Health. The data collection process for this project will focus on gathering general practitioner-related questions from the DoctorOnCall health platform. This platform is available in both English (as shown in Figure 3.4) and Malay versions (as shown in Figure 3.5) with each version containing more than 1000 questions. Condition of code-switching is observed from many users of the Malay version when posing their queries.

To extract the data from the platform, web scraping tools will be used initially, and then the collected datasets will undergo a pre-processing stage to eliminate noise and ensure a smooth and successful research process. Data Scraping from DoctorOnCall Forum is done with BeautifulSoup technique provided by Python bs4 library. Queries from the forum were saved to CSV file grouped by the disease name. A total of 171 disease categories is found from the forum, over 52970 queries. After collecting the text from webpage, the collected text will be gone through a series of pre-processing process to produce a complete set of parallel Malay-English code-switched corpus. Since not all scrapped text are guaranteed as Malay-English code-switched text, quality checked will be conducted to filtered out the monolingual rows (purely written in Malay language). Once all CSV files are ready with code-switched text, the files will be integrated as one.

Graphical user interface, text, application, chat or text message

Description automatically generated

Figure 3.4 Covid-19 Health Forum Platform, DoctorOnCall[[1]](#footnote-1) (English Version)

Graphical user interface, text, application, email

Description automatically generated

Figure 3.5 Covid-19 Health Forum Platform, DoctorOnCall[[2]](#footnote-2) (Malay Version)

### Research Activities: Data Preprocessing

Emoji Conversion

Phase Contraction Dictionary

Further Pre-processing

Export as clean data

Regex Identification

Convert to Lowercase

Removal of Punctuation, Digits and Stopword

Start

Integrated Code-switch Dataset

End

Figure 3.6 Data Pre-processing Flowchart

Data preprocessing serves as part of the data analytic process to resolve various problems that arise from raw data. Raw data may be inconsistent and incomplete in form of their content. Directly applying raw data to the analysis process would affect the quality of the analysis as ‘Rubbish in, Rubbish out”. Thus, this technique would transform the dataset into a more understandable format.

Once all CSV files are ready with code-switched text, all text will be switched to lower cases for avoiding mixed case typing variation and sparsity issues(Phua, Navaratnam, Kang & Che, 2022). Observing the data start with greetings word and ends with appreciation. Although these expressions contribute less in conveying the message, the final decision was to keep the information as over-cleaned dataset might result in overfitting issues. Next, emoji conversion process will be conducted to convert emojis to their corresponding text representations. For example, ":)" should be converted to "laugh". Regex examination for repetition will be done by using regular expressions to remove repetitions of letters in tokens. For instance, "goooood" should be converted to "good”. Creating a contraction dictionary to map unusual words or contractions to their more standard forms will be considered when further exploration is done. For example, "sry" should be mapped to "sorry." URLs, mentions, and hashtags removal is less as the source of the dataset is a simple forum. Removal of punctuation, digits, and stopwords will be done with Natural Language Toolkit (NLTK) library. Further pre-processing processes involve tasks such as data extraction, cleaning, transformation, integration, and normalization.

### Research Activities: Model Development

Segment-Based Translation

Code-switch Type Identification

Start

Language Identification

End

Figure 3.7 Model Development Flowchart

In the first phase, the language identification process will be trained with a Malay-English ambiguous word dictionary. The dictionary is a collection of all ambiguous words that exist in both Malay and English language lexical and potentially cause confusion for the model to identify their language. Besides their meanings in each language, the dictionary will be supplied with a list of related words in both language that used frequently with the ambiguous words to aid the model in selecting appropriate language tag. The trained language identification process will attach each word of input text data with a language tag, where M is for Malay language, E is for English language and M-E is for potentially intra-word code-switched words.

The code-switched type identification process will be conducted in phase 2. The input text is first split into segments by joining the continuous word with identical language tag. Each segment will be tagged with the format of [language][words count], for instance, M5 indicates a segment of 5 Malay words. The words with M-E tag will be kept for the next step. Next, the segments will be classified into code-switch types according to the pattern observed. M-E tagged words will be identified as intra-word (IW) type; Intra-sentential type is the situation of observing only two language segments and they are separated by a punctuation mark, for instance, [M6], [E3]. The greater word count segment is the main language (AM) while the smaller count is the embedding language (AE); Alternating segments with different language tags will be identified as intra-sentential type. The main (EM) and embedding (EE) language will be resolved by observing the total word count of each segment tag. A single language segment indicates Calque (C).

In phase 3, the segment translation process will be carried out. The translation pro-cess is differed by the code-switching type. An in-depth exploration of the appropriate translation process for each code-switch type will be the focus. RNN model, which is selected to develop the translation model, will be trained on Malay-English code-switch parallel corpus and homonyms dictionary with POS tagging and relevant words. Malay-English homonyms dictionary with POS tagging is a collection of homonyms words that may have several related meanings or completely distinct meanings which induced confusion in translation process. This issue is often solved by computing a translation norm that takes the translated frequency as primary criteria which may not be accurate in all conditions. Along with all of their meanings, this dictionary will be supplied with lists of related words frequently found with them in the hopes of enhancing the decoding of the appropriate meaning in translation process.

### Research Activities: Performance Evaluation

## Chapter Summary

This chapter elaborates on the outline of the proposed research. The proposed research will be conducted according to the five phases described (section 3.2). General research design and implementation are explained in Chapter 3.2. Detailed explanation on model development that separate into three stages would be elaborated in the next few chapters respectively.

# LANGUAGE IDENTIFICATION

## Introduction

The primary objective of the Language Identification Research Phase is to develop a robust and accurate language identification process specifically tailored for Malay-English code-switched texts. This phase is crucial in determining the language of each word or segment within the input text, laying the foundation for subsequent translation processes.

## Implementation

Design & Develop Language Identification Model

Performance Evaluation

Train Model with Ambiguity Dictionary

Build Ambiguity Dictionary

Start

End

Figure 4.1 Language Identification Phase Flowchart

This phase starts with building of ambiguity word dictionary, language identification model and end with performance evaluation. The expected output of current phase is an ambiguity dictionary and optimized language identification method.

### Ambiguous word Identification

The research phase involves the creation of an Ambiguity Word Dictionary, a critical linguistic resource. This dictionary contains words or terms commonly found in both Malay and English but with different meanings. For example, the ‘main’ word indicates ‘play’ in Malay while indicates ‘primary or important’ if identified as English word. Each word is tagged with its appropriate language label (Malay or English) to facilitate precise identification. Instead of their meanings in each language, the dictionary will be supplied with a list of related words in both language that used frequently with the ambiguous words to aid the model in selecting appropriate language tag. In short, the dictionary is a collection of all ambiguous words that exist in both Malay and English language lexical and potentially cause confusion for the model to identify their language. The proposed format of the dictionary is illustrated as below:

{"air":

{"Malay", {"related-1": {"M": "minum", "E": "drink"},

"related-2": {"M": "haus", "E": "thirsty"},

"related-3": {"M": "jernih", "E": "clear"}}},

{"English", {"related-1": {"M": "nafas", "E": "breath"},

"related-2": {"M": "segar", "E": "fresh"}}}}

### Language Identification Model

The proposed language identification model will be utilised deep neural network and operate at the word level. It involves training with parallel translation input of both Malay and English and the ambiguity word dictionary proposed. The trained language identification process will attach each word of input text data with a language tag, where M is for Malay language, E is for English language and M-E is for potentially intra-word code-switched words. The expected output data is proposed as below:

Job-nya\M-E adalah\M menjaga\M kedai\M runcit \M di\M end\E of\E the\E road\E.

## Performance Evaluation

Language identification models will be evaluated using various performance metrics such as equal error rate, balanced accuracy, F1 score, and false positive rate. These metrics will be applied at the level of an entire corpus, providing an overview of the system's performance. The evaluation aimed to measure the accuracy and reliability of the proposed language identification systems in identifying languages code-switch text data.

## Chapter Summary

Accurate language identification is the foundational step for effective code-switched translation. It enables the system to understand which language is being used at any given point in the text, facilitating precise translation and reducing ambiguity. The development of the Ambiguity Word Dictionary and the inclusion of relevant words are expected to significantly enhance the accuracy of language identification in code-switched texts, making it a vital contribution to the research.

# CODE-SWITCHING TYPE IDENTIFICATION

## Introduction

The primary objective of the Code-Switching Identification Research Phase is to develop a technique for accurately identifying and classifying code-switching patterns within Malay-English code-switched texts. Code-switching involves alternating between two or more languages within a single discourse, and understanding these patterns is crucial for precise translation.

## Implementation

Design & Develop Classification Model

Performance Evaluation

Train the Model with Data Prepared

Build Code-switch Type Dataset

Start

End

Figure 5.1 Code-switch Type Identification Phase Flowchart

This phase starts with building of code-switched type training data, CS-type classification model and end with performance evaluation. The expected output of current phase is a code-switch type classification dataset and optimized code-switched type classification method.

### Preparing Code-switch Type Training Data

Based on the code-switch type identified in section 2.2.1, a training data with cs-type tagging will be constructed for model training process. The word-level language tagged code-switched text from chapter 4 is utilized. The text will be first split into segments by joining the continuous word with identical language tag. Each segment will be tagged with the format of [language][words count], for instance, M5 indicates a segment of 5 continuous Malay words. The words with M-E tag will be kept for the next step. The expected output is as follow:

[Job-nya]M-E [adalah menjaga kedai runcit di]M5 [end of the road]E3.

Next, the segments will be classified into code-switch types according to the pattern observed. M-E tagged words will be identified as intra-word (IW) type; Intra-sentential type is the situation of observing only two language segments and they are separated by a punctuation mark, for instance, [M6], [E3]. The word count within each segment is used to determine the main (dominant) language and the embedding (secondary) language in intra-sentential and inter-sentential switches. The greater word count segment is the main language (AM) while the smaller count is the embedding language (AE); Alternating segments with different language tags will be identified as intra-sentential type. The main (EM) and embedding (EE) language will be resolved by observing the total word count of each segment tag. A single language segment indicates Calque (C). The expected output is as follow:

[Job-nya]IW [adalah menjaga kedai runcit di]AM [end of the road]AE.

### Code-Switching Types Identification Model

The code-switch identification model involves pattern recognition techniques to identify code-switching instances. This includes analysing the arrangement of language segments within sentences or discourse. Instead of traditional rules-based approach, deep neural network is proposed to learn the code-switching pattern from the prepared training data.

## Performance Evaluation

Besides general metrics, such as accuracy, precision and recall. The rate of separability (RS) is proposed to evaluate the separability of data in classification models.

## Chapter Summary

Accurate code-switching identification is pivotal for understanding the structure of code-switched texts and ensuring precise translation. This phase's contributions, in terms of classifying code-switching types, play a vital role in reducing ambiguity and guiding subsequent translation processes, ultimately enhancing the overall quality of Malay-English code-switched neural machine translation.

# SEGMENT-BASED TRANSLATION

## Introduction

The primary objective of the Segment-Based Translation Research Phase is to enable precise and context-aware translation of Malay-English code-switched texts. This phase focuses on translating segmented portions of the text based on language and code-switching type, ensuring accurate and meaningful translations.

## Implementation

Build Homonyms Dictionary

Performance Evaluation

Train the Model with Data Prepared & Homonyms Dictionary

Build Code-switch Parallel Corpus

Start

End

Design & Develop NMT Model

Figure 6.1 Segment-Based Translation Phase Flowchart

This phase starts with building of Malay-English code-switch parallel, homonyms dictionary, optimized translation model for each cs-type, and end with performance evaluation. The expected output of current phase is a Malay-English code-switch parallel corpus, homonyms dictionary and optimized translation method for translating Malay-English code-switched texts segment by segment. Each segment is translated with consideration of its language, code-switching type, and context, resulting in precise and context-aware translations that maintain the original meaning of the text.

### Malay-English Parallel Corpus

### Homonyms Dictionary

Malay-English homonyms dictionary with POS tagging is a collection of homonyms words that may have several related meanings or completely distinct meanings which induced confusion in translation process. This issue is often solved by computing a translation norm that takes the translated frequency as primary criteria which may not be accurate in all conditions. Along with all of their meanings, this dictionary will be supplied with lists of related words frequently found with them in the hopes of enhancing the decoding of the appropriate meaning in translation process.

### Translation Process for Different Code-Switching Types

In phase 3, the segment translation process will be carried out. The translation process is differed by the code-switching type. The input text is divided into segments based on language and code-switching type. Segments are tagged with labels indicating the language and word count, such as "M5" for a segment with 5 Malay words or "[M6], [E3]" for an inter-sentential switch between 6 Malay words and 3 English words.

An in-depth exploration of the appropriate translation process for each code-switch type will be the focus. Different translation models are employed for translating segments based on their language, code-switching type, and context. These models are designed to handle intra-sentential switches (AM-AE), inter-sentential switches (EM-EE), and code-switched segments involving ambiguity (M-E). Various pretrained language models will be explored and fine-tune with Malay-English code-switch parallel corpus and homonyms dictionary with POS tagging and relevant words.

## Performance Evaluation

Baseline Model Comparison

Crowd-sourcing Evaluation

Application to other low-resource language pair

Performance Metrics Evaluation

Start

End

Figure 6.2 Performance Evaluation Flowchart

The study will use BLEU metric, accuracy, and n-grams as performance metrics, where it will provide a comprehensive evaluation of the proposed translation model's performance, accuracy, and effectiveness. Each metrics will be computed for both baseline models and proposed model to evaluate their translation quality by comparing the translated text with original reference translation text. After self-evaluation, the proposed translation model will be further evaluated through a two-stage process. The first stage involves comparing the model's performance with the baseline BERT and Indorobusta models, followed by application to other low-resource datasets. The second stage involves crowd-sourcing evaluation, where a questionnaire containing translations from the proposed model and the GNMT will be distributed to the crowd to obtain and analyse their preferences.

## Chapter Summary

This phase plays a crucial role in the overall code-switched neural machine translation process. By segmenting and translating the text in a context-aware manner, it ensures that the translation accurately reflects the intended meaning of the code-switched content. This contributes to reducing ambiguity and enhancing the quality of Malay-English code-switched translation, ultimately improving cross-lingual communication.

# CONCLUSION

## Preliminary Outcomes

After conducting preliminary studies, the result obtained is discussed in this chapter. Current achievements and further works on this research are also elaborated in this chapter

## Achievement of Project Objectives

Phase one of the research had been successfully conducted on literature reviews on past papers are done while proper analysis has been described to justify relevant and useful information from the papers. Phase two of the research is partially accomplished by having the data collection and simple preprocessing process done. Further pre-processing process is needed to be conducted to further explore the text data.

In the progress to accomplish objective one, “to identify and organize the ambiguity present during language identification and translation process”, literature reviews on past papers are done on language identification. Language structure analysis and open-access Malay and English dictionary has been explored to identify the list of ambiguity word exists in both Malay and English lexicon. This study has provided useful insight for understanding the basic concept in developing a machine translation mechanism for Malay-English code-switch text.

## Future Research Steps

Future works to be conducted for this research will emphasize the design and development process including:

1. Malay-English Ambiguity Dictionary with related words.
2. Explore the effect of ambiguity dictionary to the performance of language identification.
3. Explore more on code-switch type identification including the evaluating metrics and efficient algorithms.
4. Homonyms Dictionary with POS tags and relevant words.
5. Explore the effect of homonyms dictionary to the performance of machine translation.
6. Explore more about the structure and design of machine translation to improve the translation accuracy of code-switch text.
7. Apply the proposed solution to other code-switched language-pairs in order to further evaluate the solution

## Conclusion

Current NMT models cope less effectively with code-switching situations. A conceptual framework is proposed in this paper to address the challenges that arise in translating Malay-English code-switched texts. The framework aims to improve translation accuracy by handling the linguistic characteristics, informal language usage, structural differences, and inherent ambiguity prevalent in code-switched texts. The incorporation of advanced techniques such as language identification, code-switching type identification, and segment translation, built a promising solution for neural machine translation in the code-switching context. This research will produce all data needed in the proposed framework including Malay-English parallel corpus, Malay-English ambiguous word dictionary and Malay-English homonyms dictionary with POS tagging and relevant words. Future work can build upon this framework to refine and expand our understanding of code-switched translation and its applications.

REFERENCES

Adilazuarda, M. F., Cahyawijaya, S., Winata, G. I., Fung, P., & Purwarianti, A. (2022). IndoRobusta: Towards Robustness Against Diverse Code-Mixed Indonesian Local Languages. *In Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, 25-34.

Ahmad Mahir, N., & Silahudin Jarjis, M. K. (2007). The Use of Malay Malaysian English in Malaysian English: Key Considerations. *In: The Second Biennial International Conference on Teaching and Learning of English in Asia : Exploring New Frontiers (TELiA2)*, 1-9. Retrieved from https://repo.uum.edu.my/id/eprint/3263

Alon, Lavie., Mark, A., Przybocki. (2009). Introduction to the special issue on Automated Metrics for Machine Translation Evaluation. *Machine Translation, 23*, 69-70.

Appicharla, R., Gupta, K. K., Ekbal, A., & Bhattacharyya, P. (2021). IITP-MT at CALCS2021: English to Hinglish neural machine translation using unsupervised synthetic code-mixed parallel corpus. *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, 31-35.

Azmi, M. N. L., Ching, L. T. P., Jamaludin, N. B., Ramli, M. N. H. B., Razali, M. H. B., Amram, M. A. Y. B., & Jayakumar, K. A. (2016). The comparison and contrasts between English and Malay languages. *English Review, 4*(2), 209-218.

Bahdanau, D., Cho, K., & Bengio, Y. (2016). Neural Machine Translation by Jointly Learning to Align and Translate. *ICLR 2015.* doi:10.48550/arXiv.1409.0473

Chen, S., Aguilar, G., Srinivasan, A., Diab, M., & Solorio, T. (2020). CALCS 2021 Shared Task: Machine Translation for Code-Switched Data. *arXiv*. Retrieved from https://doi.org/10.48550/arXiv.2202.09625

Cho, K., van Merrienboer, B., Bahdanau, D., & Bengio, Y. (2014). On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. *Proceedings of {SSST}-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, 103-111. doi:10.3115/v1/W14-4012

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)1724-1734.* doi:10.3115/v1/D14-1179

Culler, J. (1976). *Structuralist poetics: structuralism, linguistics, and the study of literature.* The University of Chicago Press.

Cummings-Clay, D. & Hostos Community College. (2019). Chapter 8.2 Introduction to Language. In D. &. Cummings-Clay, *Child Development.* LibreTexts.

D'Anna, C. A., Zechmeister, E. B., & Hall, J. W. (1991). Toward a meaningful definition of vocabulary size. *Journal of Reading Behavior: A Journal of Literacy, 23*(1), 109-122.

Devlin, J., Chang, W. M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *In Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171-4186.

Dinu, G., Mathur, P., Federico, M., & Al-Onaizan, Y. (2019). Training Neural Machine Translation to Apply Terminology Constraints. . *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3063–3068.

Elgort, I. (2013). Effects of L1 definitions and cognate status of test items on the Vocabulary Size Test. *Language Testing, 30*(2), 253-272.

Fu, Y., Lin, N., Yang, Z., & Jiang, S. (2021). An Open-Source Dataset and A Multi-Task Model for Malay Named Entity Recognition. *Fu, Y., Lin, N., Yang, Z., & Jiang, S. (2021). An Open-Source Dataset and A Multi-Task Model for Malay Named Entity Recognition. ArXiv, abs/2109.01293.* doi:10.48550/arXiv.2109.01293

Gahoi, A., Duneja, J., Padhi, A., Mangale, S., Rajput, S., Kamble, T., Sharma, D., & Varma, V. (2022). Gui at MixMT 2022: English-Hinglish: An MT approach for translation of code mixed data. *Conference on Machine Translation.* doi:10.48550/arXiv.2210.12215

Goulden, R., Nation, P., & Read, J. (1990). How large can a receptive vocabulary be? *Applied Linguistic, 11*(4), 314-363.

Grundkiewicz, R. &. (July, 2018). Neural Machine Translation Techniques for Named Entity Transliteration. *In Proceedings of the Seventh Named Entities Workshop*, 89-94. doi:10.18653/v1/W18-2413

Gupta, Y., Raghuwanshi, G., & Tripathi, A. (2021). A New Methodology for Language Identification in Social Media Code-Mixed Text . *Advanced Machine Learning Technologies and Applications*, 243-254. doi:10.1007/978-981-15-3383-9\_22

Hashim, A. (2017). Crosslinguistic Influence In the Written English Of Malay Undergraduates. *Journal of Modern Languages, 12*(1), 60-76.

Hazenberg, S., & Hulstijn, J. H. (1996). Defining a minimal receptive second-language vocabulary for on-native university students: An empirical investigation. *Applied Linguistics, 17*, 145-163. doi:10.1093/applin/17.2.145

Hidayatullah, A. F., Qazi, A., Lai, D. T. C., & Apong, R. A. (2022). A Systematic Review on Language Identification of Code-Mixed Text: Techniques, Data Availability, Challenges, and Framework Development. *IEEE Acess, 10*, 122812-122831. doi:10.1109/ACCESS.2022.3223703

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memoery. *Neural computation, 9*(8), 1735-1780.

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A Brief Survey of Text Mining. *Journal for Language Technology and Computational Linguistics, 20*(1), 19-62. doi:10.21248/jlcl.20.2005.68

Hutchins, J. (2004). The Georgetown-IBM Experiment Demonstrated in January 1954. *Machine Translation: From Real Users to Research*.

Izyani, N. R. (2016). *INTRASENTENTIAL CODE-SWITCHING AMONG MALAYS IN MALAYSIA ON FACEBOOK.* Retrieved from Academia.

Jadhav, I., Kanade, A., Waghmare, V., Chandok, S. S., & Jarali, A. (2022). Code-Mixed Hinglish to English Language Translation Framework. *2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, (pp. 684–688). doi:10.1109/ICSCDS53736.2022.9760834

Jaech, A., Mulcaire, G., Hathi, S., Ostendorf, M., & Smith, N. (2016). Hierarchical character-word models for language identification. Retrieved from https://doi.org/10.18653/v1/w16-6212

Jain, N., & Bhat, R. A. (2014). Language Identification in Code-Switching Scenario. *Proceedings of the First Workshop on Computational Approaches to Code Switching*, 87-93. doi:10.3115/v1/W14-3910

Jesin, J., Vithya, Y., Isabella, S., Catherine, W., Peter, J.K., Peter-Lucas, J., & Keoni, M. (2022). Language Models for Code-switch Detection of te reo M¯aori and English in a Low-resource Setting. *Findings of the Association for Computational Linguistics: NAACL 2022* , 650-660.

Kasmuri, E., & Basiron, H. (2020). Segregation of Code-Switching Sentences using Rule-Based Technique. *Int. J. Advance Soft Compu*. Retrieved from http://www.i-csrs.org/Volumes/ijasca/4.%20p49\_64\_Segregation%20of%20Code-Switching%20Sentences.pdf

Kester, E. (04 February, 2021). *Code-Switching and Code-Mixing – What You Need to Know*. Retrieved from Bilinguistics: Better Speech Therapy: https://bilinguistics.com/code-switching-and-code-mixing/

Kuwanto, G., Akyürek, A. F., Tourni, I. C., Li, S., Jones, A. G., & Wijaya, D. (2021). Low-Resource Machine Translation Training Curriculum Fit for Low-Resource Languages. *Arvix Preprint*. doi:http://arxiv.org/abs/2103.13272

Lal, Y.K., Kumar, V., Dhar, M., Shrivastava, M., & Koehn, P. (2019). De-mixing sentiment from code-mixed text. *In Proc. of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, 371-377.

Lee, G., & Li, H. (2020). Modeling Code-Switch Languages Using Bilingual Parallel Corpus. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 860–870). Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.80

Lee, S., Lee, J., Moon, H., Park, C., Seo, J., Eo, S., Koo, S., & Lim, H. (2023). A Survey on Evaluation Metrics for Machine Translation. *Mathematics, 11*(4), 1006. doi:10.3390/math11041006

Liang, S.Y., Li, S.S., Fu, S.H., & Lin, N.K. (2020). An Overview of Natural Language Processing for Indonesian and Malay. *Pattern Recognition and Artificial Intelligence, 33*(6), 530-541. doi:10.16451/j.cnki.issn1003-6059.202006006

Mave, D., Maharjan, S., & Solorio, T. (2018). Language Identification and Analysis of Code-Switched Social Media Text. *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, (pp. 51-61). doi:10.18653/v1/W18-3206

Mohamad Khalil, S., & Mohd Shahril Firda, M. (2018). INTER-SENTENTIAL AND INTRA-SENTENTIAL CODE SWITCHING IN PARLIAMENTARY DEBATE. *International Journal of Modern Languages and Applied Linguistics, 2*(4), 29-35. doi:10.24191/ijmal.v2i4.7691

Mohamad, A. N. A., Salleh, A. R., & Mohamad, W. N. A. (2022). Malay Language as the State Emblem of Malaysia. *E-Jurnal Bahasa Dan Linguistik (e-JBL), 4*(2), 96-108. doi:doi.org/10.53840/ejbl.v4i2.121

Mukherjee, S., Prasan, V., Nediyanchath, A., Shah, M., & Kumar, N. (2019). Robust deep learning based sentiment classification of code-mixed text. . *In Proc. of the 16th International Conference on Natural Language Processing*, 124-129.

Nguyen, L. T. C., & Nation, I. S. P. (2011). A bilingual vocabulary size test of English for Vietnamese learners. *RELC Journal, 42*(1), 86-99.

Nguyen, L., Mayeux, O., & Yuan, Z. (2023). Code-switching input for machine translation: A case study of Vietnamese–English data. *International Journal of Multilingualism*, 1-22. doi:10.1080/14790718.2023.2224013

Nitika, Mathur., Timothy, Baldwin., Trevor, Cohn. (2020). Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* doi:Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics.

Patwa, P., Aguilar, G., Kar, S., Pandey, S., PYKL, S., Gambäck, B., Chakraborty, T., Solorio, T., & Das, A. (2020). SemEval-2020 Task 9: Overview of Sentiment Analysis of Code-Mixed Tweets. . *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, 774-790.

Phua, Y. T., Navaratnam, S., Kang, C.-M., & Che, W.-S. (2022). Sequence-to-sequence neural machine translation for English-Malay. *IAES International Journal of Artificial Intelligence (IJ-AI), 11*(2), 658-665. doi:10.11591/ijai.v11.i2.pp658-665

Sabri, N., Edalat, A., & Bahrak, B. (2021). Sentiment analysis of persian-english code-mixed texts. *26th International Computer Conference, Computer Society of Iran, CSICC 2021*.

Samanta, B., Ganguly, N., & Chakrabarti, S. (2019). Improved sentiment detection via label transfer from monolingual to synthetic code-switched text. *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3528-3537.

Schuster, M. & Paliwal, K.K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing, 45*(11), 2673-2681. doi:10.1109/78.650093

Siti Nor Allia, N. A., & Sabrina, T. (2020). Rule-based Text Normalization for Malay Social Media Texts. *International Journal of Advanced Computer Science and Applications, 11*(10), 156-162. doi:10.14569/IJACSA.2020.0111021

Song, K., Zhang, Y., Yu, H., Luo, W., Wang, K., & Zhang, M. (2019). Code-Switching for En-hancing NMT with Pre-Specified Translation. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Lanugage Technologies, 1*, 449-459.

Song, Z., Zhou, H., Qian, L., Xu, J., Cheng, S., Wang, M., & Li, L. (2021). switch-GLAT: Multilingual parallel machine translation via code-switch decoder. *ArXiv preprint* . doi:2104.08687

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. *ArVix*. doi: /abs/1409.3215

Tan, A.-H., Ridge, K., Labs, D., & Terrace, H. (2000). Text Mining: The state of the art and the challenges.

*The Most Spoken Languages 2023*. (2023). Retrieved from Statistics and Data: https://statisticsanddata.org/data/the-most-spoken-languages-2023/#:~:text=According%20to%20new%20studies%2C%20English,over%20the%20past%20few%20years.

UKEssays. (November, 2018). *History of the Malay Language*. Retrieved from UK Essays Expert: https://www.ukessays.com/essays/history/history-of-malay-language-history-essay.php?vref=1

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 5998-6008.

Vissamsetti, S. H., Pravallika, K. S. S., Jadhav, S. C., & Kommanti, H. B. (2022). Offensiveness detection in hinglish code-switched language. . *Communications in computer and information science*, 124-135.

Weller, O., Sperber, M., Pires, T., Setiawan, H., Gollan, C., Telaar, D., & Paulik, M. (2022). End-to-End Speech Translation for Code Switched Speech. . *arXiv*. Retrieved from https://arxiv.org/abs/2204.05076v1.

Winata, G. I., Madotto, A., Wu, C.-S., & Fung, P. (2019). Code-Switched Language Models Using Neural Based Synthetic Data from Parallel Sentences. *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 271–280.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., … Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *ArXiv*. doi:10.48550/arXiv.1609.08144

Xu, J., & Yvon, F. (2021). Can You Traducir This? Machine Translation for Code-Switched Input. *arXiv*. Retrieved from https://doi.org/10.48550/arXiv.2105.04846

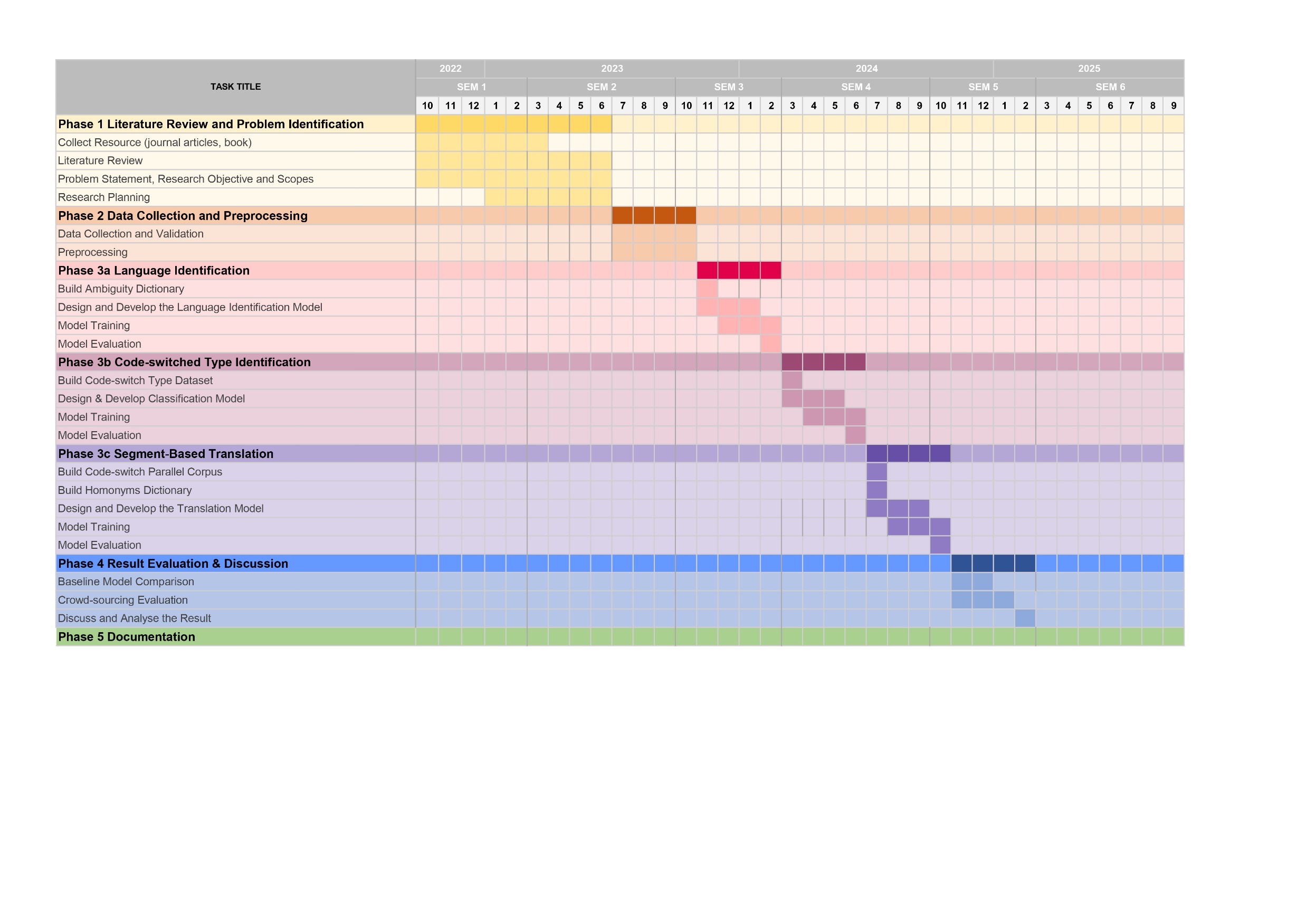
Yang, Y., Christopher, M., Huang, A. (2020). Finbert: A pretrained language model for financial communication.

Yang, Z., Hu, B., Han, A., Huang, S., & Ju, Q. (2020). Csp:code-switching pre-training for neural machine translation. Retrieved from https://doi.org/10.18653/v1/2020.emnlp-main.208

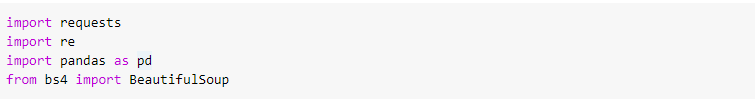
Yeong, Y.-L., Tan, T.-P., & Mohammad, S. K. (2016). Using Dictionary and Lemmatizer to Improve Low Resource English-Malay Statistical Machine Translation System. *Procedia Computer Science, 81*, 243-249. doi:10.1016/j.procs.2016.04.056

Zabha, N. I., Ayop, Z., Anawar, S., Hamid, E., & Zainal, Z. (2019). Developing cross-lingual sentiment analysis of Malay Twitter data using lexicon-based approach. *International Journal of Advanced Computer Science and Applications, 10*(1). doi:10.14569/IJACSA.2019.0100146

Appendix A Gantt Chart



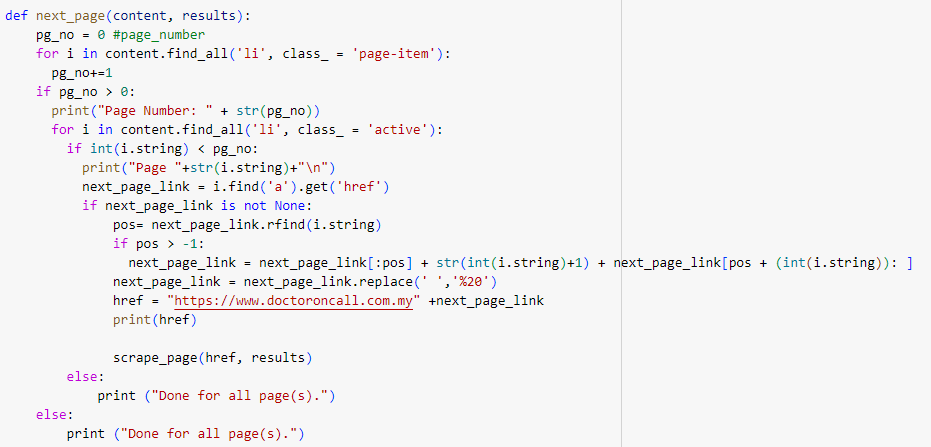
Appendix A Web Scrapping Code Snippet





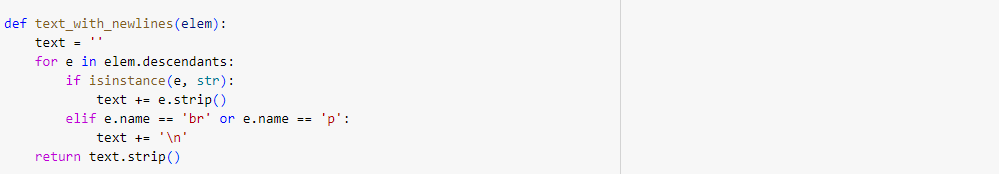










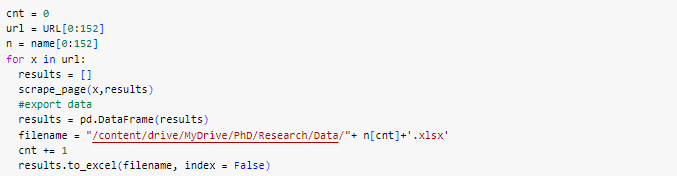












LIST OF PUBLICATIONS

1. https://www.doctoroncall.com.my/tanya/en [↑](#footnote-ref-1)
2. https://www.doctoroncall.com.my/tanya/ [↑](#footnote-ref-2)