2021 International Congress of Advanced Technology and Engineering (ICOTEN) | 978-1-6654-1224-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICOTEN52080.2021.9493503

A Conceptual Framework For Malay-English Mixed-language Question Answering System

Hui Ting Lim
School of Computing, Faculty of Enginering
Universiti Teknologi Malaysia
Johor, Malaysia
htlim4@graduate.utm.my

Sharin Hazlin Huspi
School of Computing, Faculty of Enginering
Universiti Teknologi Malaysia
Johor, Malaysia
sharin@utm.my

Roliana Ibrahim
School of Computing, Faculty of Enginering
Universiti Teknologi Malaysia
Johor, Malaysia
roliana@utm.my

***Abstract*—Mixed language has turned into a current trend of language which refers to combining two or more languages either in spoken or written form. It has been widely used in social media forums to improve communication and for a greater range of expression. The current question answering (QA) system only supports monolingual queries, which restricts the capability of multilingual users to have a natural interaction with the system. In recent years, there has been a rise of interest in multilingual QA systems where translation models merged with machine learning algorithms in question classification are the commonly used solution. However, using words from other languages in a single sentence has led to the problem of the inability to identify code-switch from the monolingual sentence; this has also caused the problem of limited captured language context from machine translation processed mistranslated questions. The informal mixed-language representation that disobeys the natural linguistic rule in particular languages provides a challenge for automated QA systems, as the systems would need to translate and extract answers for the given questions. Additionally, lack of public resources such as Chunker, POS Tagger, and WordNet for mixed-language, especially for Malay-English, leads to low performance of the translation and classification model. Furthermore, the use of machine learning algorithms in question classification requires a large number of structured training data to ensure performance. This is impracticable in the Malay-English mixed-language domain since the availability of the mixed-language dataset is still an issue. To solve these problems, we aim to propose a framework consisting of the combination of enhanced translation models with deep learning; by using Convolutional Neural Networks (CNN) to address the Malay-English mixed-language question classification to generate the best answer. The first part will study the machine translation model, where word-level language identification and text normalization towards Malay-English mixed-language questions will be developed. The second part will focus on the deep learning algorithm, where we will explore CNN as the classification model to assist in the translated questions to provide the best answer. Thus, in this paper, a framework consisting of an enhanced translation model for Malay-English, and also an end-to-end mixed-language question answering system for the Malay-English Q&A system, is presented. This research will provide a significant contribution to a multilingual forum platform and also to intelligent Q&A systems (chatbots).**

Keywords—Text Analytics, Code-Switching, Malay-English Translation, Text Normalization, Deep Learning, Question-Answering System

I. Introduction

Our world is becoming increasingly multilingual, and this can be observed obviously especially after colonization in one country. Malaysia is a multilingual society with at least hundreds of languages that are spoken by more than a million native speakers. However, Malay is the official language in Malaysia. Whereas English is a second language that is known by mostly all the Malaysians especially the younger generation because it is used by many schools and universities as a medium of instruction [1]. Therefore, many bilingual or trilingual people in Malaysia have the knowledge of Malay and English in addition to their mother tongue. With the different levels of proficiency and zones of comfort in English and Malay, it is unavoidable that mixed language will be used in such a multilingual setting. Gradually, mixed language has turned into the current trend of language which refers to combining two or more languages whether in spoken or written form.

The Malay-English mixed-language scenario is growing extensively in social media such as Facebook, Twitter, Blogs, and even in forum platforms [2]. It is common to see that social media users show a preference for using mixed language in one sentence in expressing their opinion, comments, and questions. With the availability of easy internet access to people, social media involvement has been gradually increasing year by year. Based on the survey by Statista [3], there are 29 million internet users in Malaysia and this amount is forecast to reach up to 33.5 million in 2025. This has resulted in Malay-English mixed language content on various media types such as websites and chats, which has increased significantly. These internet users publish content on different topics which often deal with personal views, discussions, or inquiry on recent events, services, technology, tourism, and health. The mixed-language texts written in social media are usually informal in nature and do not follow the natural linguistic rules in the particular languages. Therefore, Malay-English mixed-language poses challenges and difficulties for Natural Language Processing (NLP). In addition to Malay language, it also currently still lacks public resources [4].

It can be seen that current search engines as well as information retrieval systems have developed more and more over the years and have become advanced; from enabling telegraphic user queries which return top relevant URLs, to being able to support questions as queries that can be expressed in Natural Language and return a precise and accurate answer to the users [5]. The evolution of information retrieval systems has resulted in the existing systems being called as Question Answering (QA) systems. It brings a much convenient and natural way to get the information needed quickly and precisely. While the current QA systems can process a query in natural language, they only support monolingual questions such as in English, German, or French. This severely restricts the capability of multilingual users to have a natural interaction with the QA system. This can be seen especially when expressing the query involved in health, technology, and scientific terms. The critical scenario has

978-1-6654-1224-7/21/$31.00 ©2021 IEEE

been observed from the community question answering as well as the community health platform for COVID-19. Majority of Malaysians pose their inquiry in Malay-English mixed-language form, specifically for terms in COVID-19 such as “swab test”, “positive”, “negative”, “red zone” and so on.

Currently, there is no Malay-English mixed-language question answering that has been specifically investigated. Most of the previous studies in Malay-English mixed-language text analysis were intended to sentiment analysis [4, 6-8], where the mixed language was commonly used in social media as a popular platform for sentiment analysis study. However, social media should not be referred to only Facebook, Instagram, and YouTube, but it also includes the forum platforms such as Quora, Yahoo! Answers, and Stack Overflow. These forum platforms as well as Question Answering platforms allow the user to ask questions and get a precise answer from other users. It is found that a health forum platform for COVID-19, namely, DoctorOnCall[1], allows Malaysian users to ask questions and get answers from a professional doctor team. Through observation, it is seen that the majority of Malaysian users use Malay-English mixed-language for their inquiry, and they then need to wait for the reply. As such, despite the fact that English is the principal language for social media communication, there is a necessity to create and develop mechanisms for other languages [9].

There have been several mixed-language question answering systems investigated by other researchers in other mixed-languages such as Hindi-English, Bengali-English and Spanish-English. In a mixed-language question answering system, question classification is a crucial part, where the user's mixed-language query needs to be analyzed to predict the answer type [5]. The machine learning method proposed by Raghavi et al. [5] and Anand et al. [10] poses the disadvantage that a large number of training data is needed. Nevertheless, the availability of mixed-language datasets is the major problem that is not yet solved in mixed-language computational study. Although there exists a python package for cross-lingual word similarity, the problem of certain Malay words not existing in the WordNet dataset lead to inaccurate performance [7]. Apart from that, the usage of short form words, slang words, misspelling is still an issue that is not yet solved in Malay-English mixed-language text analysis [11].

Hence, in this paper, we have presented a framework of a question answering system that can process a Malay-English mixed-language question in order to generate the answer for the users. The framework is the combination of word-level language identification, text normalization, and lexical translation with deep learning; by using a Convolutional Neural Networks (CNN) to address the Malay-English mixed-language question classification in order to generate the best answer. The following are the key contributions of this research:

- Building an automatic Covid-19 Community Question Answering Framework that can process Malay-English mixed-language queries from users.
- Proposed translation model based on word-level language identification and text normalization can be used for different areas of Malay-English mixed-language study such as sentiment analysis and chatbots.
- Producing the rule set for normalizing the noise words such as short form words and slang words.
- Producing the Malay-English mixed-language dataset for question answering systems in the Covid-19 health domain.

The organization of this paper is as follows. In Section II, the review of related works and research motivation are discussed. Next, the proposed framework is demonstrated in Section III. Lastly, our work of this paper is concluded and summarized in the Section IV.

II. Review And Motivation

As mentioned above, mixed language is a current trend of language which refers to combining two or more languages whether in spoken or written form. This phenomenon usually happens in multilingual society countries, especially in Southeast Asia, in place like Malaysia, Singapore, Indonesia, and Brunei. There are two new kinds of languages generated from the mixing, called code-switching [12-14], and code-mixing [5, 13, 15].

According to Thara and Poornachandran [12], the term code-switching can refer to the occurrence of changing between two or more languages in a single conversation. While for the term code-mixing, Raghavi and Shrivastava [5] defined it as the mixing of different types of languages in a single utterance. Sometimes, it is hard to differentiate between code-mixing and code-switching, and they can be used interchangeably especially for study areas in syntax, morphology, and other formal aspect languages [15]. Code-mixing has its own extremely specific definitions in linguistics, education theory, or communication study areas. There exist much more complicated and deeper definitions for code-mixing terms. Therefore, in this research, the term code-switching will be used as a cover for all types of mixed-language question texts. Even this terminology is still an issue among the researchers, but it is undoubtedly that all types of mixed languages create difficulties for computational systems developed with monolingual data.

Many multilingual communities have appeared across the globe because of colonization, which has led to the emergence of code-switching [16]. The use of code-switching is extremely prevalent in social media such as Facebook, Twitter, WhatsApp, and even different domains of forum platforms. The crucial examples are Hindi-English [5,9,12,17-19] in India, Spanish-English [20-22] in the United States of America, Malay-English [1,2,11,22] in Malaysia, Cantonese-English [23] in Hong Kong, and Arabic-English in Egypt. A number of studies have found that the possible reasons for code-switching are ascribed to the speeding up of communication, inadequate use of appropriate words in the native language, and emphasis on specific words or phrases emphasis. Based on a study by Choudhury et al. [24], the user generated contents in the form of code-switching in these areas are as much as 20%. The informal and special styles of mixed-language texts possess linguistic challenges in computational processing and in understanding of the text. Therefore, it is getting essential to handle code switching content as part of the Natural Language Processing (NLP) systems and applications for these regions.

Throughout the study and review of previous works, each language pair had their respective problems. Several researchers have studied the computational models and NLP

[1] https://www.doctoroncall.com.my/tanya/en

techniques to process code-switching data. The mixed-language applications comprise of information retrieval, language identification, sentiment analysis, data construction and so on. There is a large volume of published studies on the Hindi-English and Bengali-English question answering systems. However, there has been little discussion about the Malay-English mixed-language. Most studies in Malay-English mixed-language have only been carried in sentiment analysis and data construction. No report has been found so far using the Malay-English mixed-language dataset that investigate question answering.

One of the mixed-language applications' features is information retrieval as well as question answering. Information retrieval is a system that will return a list of top relevant URLs for a user query, for example, Google search engine, Yahoo, Bing etc. Whereas question answering was introduced to improve the traditional systems of information retrieval, where users are allowed to get a short, precise, and accurate answer for their query. Both information retrieval and question answering support and process the question queries expressed in natural language. Current information retrieval and question answering systems have only enabled interactions in monolinguals such as English, German, and French. For the purpose of enhancing the information retrieval and question answering to be more human-like and user friendly, several studies investigating mixed-language queries have been carried out on the information retrieval and question answering application.

It has been suggested that question classification is the most significant process that can be used to handle mixed-language queries in information retrieval and question answering applications. Khayathi et al. [5] proposed a basic Support Machine Vector (SVM) to classify English-Hindi code-mixed questions. They utilized the word-level resources such as language identification, transliteration, and lexical translation, where the accuracy of classification reached up to 63%. However, the SVM relies on a large number of training data and this causes failure to differentiate and identify two words that have similar meaning, for instance, mountain and hill.

The research study by Anand et al. [10] was also in code-mixed question classification. They proposed a Bog-of-Word (BoW) model and Recurrent Neural Network (RNN) to classify the Bengali-English code-mixed questions automatically. The use of RNN also has the same drawback as Khayathi et al.'s [5] study, where the training corpus is small, and thus the inaccessibility of unsupervised data caused the performance of RNN (73.88%) to be lower than that of BoW (80.55%). In contrast, the study by Banerjee et al. [25] indicated the use of feature engineering with Convolutional Neural Network (CNN) to classify the Bengali-English questions. The issues with the scarcity of training data occurred in the Khayathi et al. [5] and Anand et al.'s [10] studies was solved. The authors obtained 87.22% accuracy of classification on their proposed framework and successfully enhanced the latest question classification accuracy by around 4%.

Another study by Thara et al. [18] presented the mixed-language question answering framework that processed the question that was mixed between English and 3 different Indian languages: Hindi, Telugu, and Tamil. The authors performed experimental investigation on question classification by comparing two different approaches which are Recurrent Neural Network (RNN) and Hierarchical Attention Network (HAN). The performance of the RNN algorithm reached up to 80.67%, and it was better than HAN (75.70%) in classifying the mixed-language questions. However, interestingly, this is contrary to a study conducted by Anand et al. [9], where the RNN resulted in a worse performance than other approaches that the author used for comparison. The difference between Thara et al. 's [18] works and Anand et al.'s [10] is that Thara and team [18] interpreted the mixed-language question into full English before feeding the query into RNN, while Anand et al. [10] followed a difference process. Table I shows the comparison of previous works with the proposed model.

TABLE I. COMPARISON OF PREVIOUS WORKS WITH THE PROPOSED MODEL.

Authors	Methods	Datasets	Pros	Cons
[5]	Translation + Support Machine Vector (SVM)	English-Hindi	- Additional translation model and adjacent features improve the course-grained class.	- SVM relies on large number of training data. - Cannot differentiate and identify two words that have similarity meaning.
[18]	Translation + RNN / HAN	English-Hindi/ Telugu/Tamil	- Translation of mixed-language questions.	- Simple translation model causes unordered sentences of words.
[25]	CNN + feature engineering	Bengali-English	- Can overcome the scarcity of training data.	- No translation of mixed-language questions.
Our Proposed Model	**Translation + CNN**	**Malay-English**	- **Translation Model with word-level language identification and text normalization.** - **Can overcome the scarcity and the unlabeled nature of training data.**	- **The text normalization may only be specific to the Malay and English.**

Based on the review in Malay-English mixed-language research, the problems that have not been in solved are those of slang words, short form words, and misspelling leading to misclassification and affecting the overall performance of the model. As mentioned in the study by [4], their proposed lexicon-based model for sentiment analysis of Malay twitter obtained low accuracy and precision. This is because the classifier was unable to recognize the keywords represented in slangs or dialect words and also short form words, resulting in misclassification. The low number of clean data obtained by the author is another reason that affected the confidence level of the outcome. Other works done by [7] also have similar issues, where the proposed semantic similarity approaches were based on WordNet for sentiment analysis. In this study, since such Malay words do not appear in the available sources, and Malay is still a low computational resource language, an inaccurate score was obtained. These issues have not only arisen in Malay-English mixed-language texts, but also in other language pairs. The study by [26] had a misclassification problem due to certain Bengali words that have the same spelling as English words. This revealed that pre-processing as well as cleaning of the dataset is a critical task in mixed-language computational study. The non-standard format used in mixed-language texts has a big impact on the performance of the proposed models as shown in the previous studies.

Apart from that, in mixed-language question answering research, the obvious and major problems that appeared in previous studies were regarding the proposed models being based on supervised machine algorithms as well as deep learning requiring a large number of training data to ensure better performance. These include the Support Vector Machine (SVM), Recurrent Neural Network (RNN), and Hierarchical Attention Network (HAN) in question classification, which are the main tasks in mixed-language question answering. However, it is not reliable in the Malay-English mixed-language text analysis context, as it is a low resource language, and limited accessible datasets are available online.

Hence, we decided to explore the Malay-English mixed-language text analysis for a question answering system in a Covid-19 health forum domain. Automated Covid-19 health question answering can bring much benefit and convenience to the community, if they are allowed to inquire for any information by using mixed-language at any time. It is a challenging task in question answering to process the mixed-language questions and extract the correct answers in response to the users. Our proposed models will integrate and combine the advantage of previously proposed techniques and rectify the limitations in order to accomplish the research goals.

A translation model with word-level language identification and text normalization has been suggested to the low computational resource languages problem. Instead of directly using the mixed-language question for classification, we suggested to translate the Malay-English mixed-language questions into English to reduce the data complexity. Besides that, in order to solve the slangs, short forms as well as the ambiguous words in the mixed-language questions, text normalization will be performed before translating question. Whereas the deep learning Convolutional Neural Network (CNN) model will be proposed to overcome the large number of training dataset requirement as stated above. The detailed explanation of the research design and experimental framework has been provided in next section.

III. Proposed Framework

The preliminary study of this research has indicated that there is a growing need to develop a framework for Malay-English question answering systems. In order to solve the problems stated in the previous section, we have planned to experiment and investigate the combination of language identification, text normalization, and lexical translation with deep learning; by using a Convolutional Neural Networks (CNN) model to address the Malay-English mixed-language question classification, in order to generate the best answer. The first part will introduce the machine translation model whereby word-level language identification and text normalization towards Malay-English mixed-language questions will be developed. While the following part will focus on the deep learning algorithm, where we will explore CNN as the classification model to assist in the translated questions to provide the best answer. Fig. 1 shows our proposed framework.

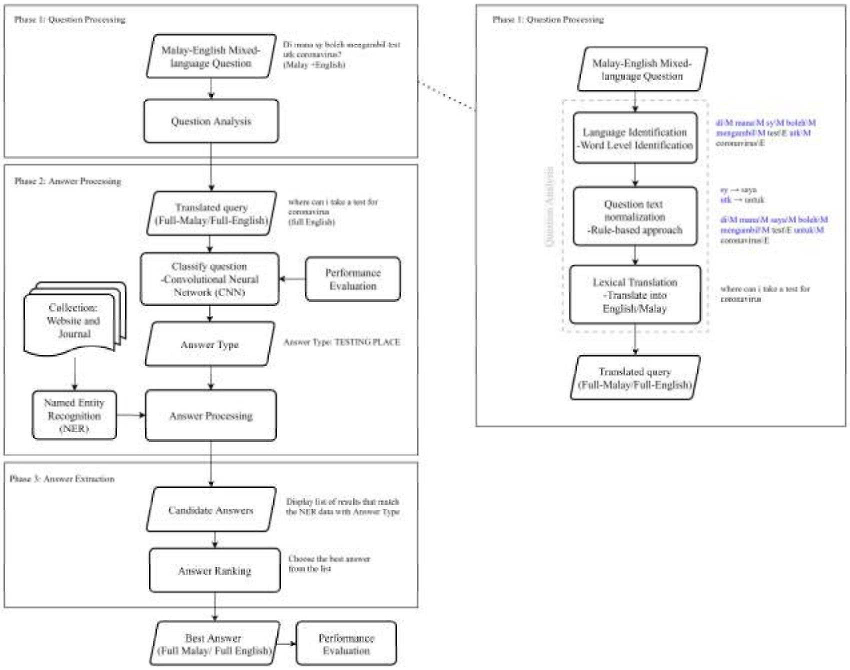
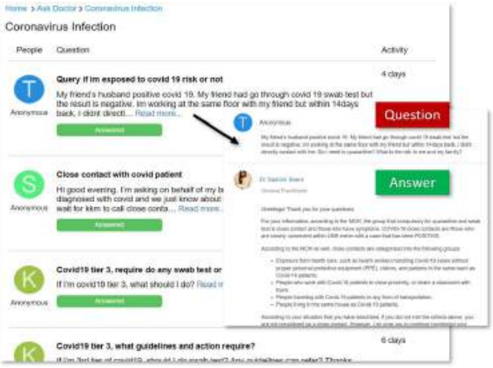

Phase 1: Question Processing
Malay-English Mixed-language Question
Di mana sy boleh mengambil test utk coronavirus? (Malay +English)
Question Analysis
Phase 2: Answer Processing
Translated query (Full-Malay/Full-English)
where can i take a test for coronavirus (full English)
Classify question -Convolutional Neural Network (CNN)
Performance Evaluation
Collection: Website and Journal
Answer Type
Answer Type: TESTING PLACE
Named Entity Recognition (NER)
Answer Processing
Phase 3: Answer Extraction
Candidate Answers
Display list of results that match the NER data with Answer Type
Answer Ranking
Choose the best answer from the list
Best Answer (Full Malay/ Full English)
Performance Evaluation
Phase 1: Question Processing
Malay-English Mixed-language Question
Question Analysis
Language Identification -Word Level Identification
di/M mana/M sy/M boleh/M mengambil/M test/E utk/M coronavirus/E
Question text normalization -Rule-based approach
sy → saya
utk → untuk
di/M mana/M saya/M boleh/M mengambil/M test/E untuk/M coronavirus/E
Lexical Translation -Translate into English/Malay
where can i take a test for coronavirus
Translated query (Full-Malay/Full-English)

Fig. 1. Proposed Framework

A. Data Preparation

The dataset is collected from a Covid-19 health platform named DoctorOnCall. This platform allows the community to ask any questions about Covid-19 as well as other health inquiries. The user inquiry will be answered by a professional doctor team that is certified and registered under the Malaysian Ministry of Health and has years of experience in treating patients. Fig. 2 and Fig. 3 shows the DoctorOnCall health forum platform in English version and Malay version, respectively. There are more than 1000 questions along with an answer available in each version. In the Malay version, it is obversed that the majority of the users practice code-switching between Malay and English when asking questions.

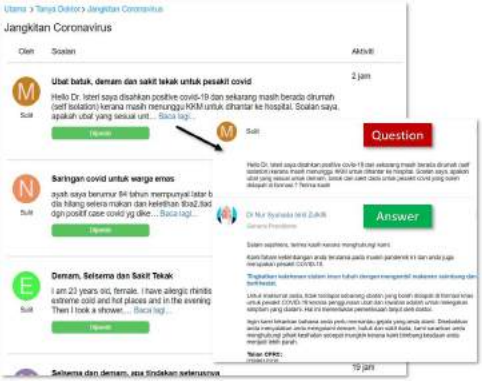


Home > Ask Doctor > Coronavirus Infection

Coronavirus Infection

People	Question	Activity
T Anonymous	**Query if im exposed to covid 19 risk or not** My friend's husband positive covid 19. My friend had go through covid 19 swab test but the result is negative. Im working at the same floor with my friend but within 14days back, I didnt directl... Read more... Answered	4 days
S Anonymous	**Close contact with covid patient** Hi good evening. I'm asking on behalf of my b diagnosed with covid and we just know about wait for kkm to call close conta... Read more... Answered	
K Anonymous	**Covid19 tier 3, require do any swab test or** If I'm covid19 tier 3, what should I do? Read m Answered	
K	**Covid19 tier 3, what guidelines and action require?**	6 days

Fig. 2. Covid-19 Health Forum Platform, DoctorOnCall. (English Version)



Utama > Tanya Doktor > Jangkitan Coronavirus

Jangkitan Coronavirus

Oleh	Soalan	Aktiviti
M Sulit	**Ubat batuk, demam dan sakit tekak untuk pesakit covid** Hello Dr. Isteri saya disahkan positive covid-19 dan sekarang masih berada dirumah (self isolation) kerana masih menunggu KKM untuk dihantar ke hospital. Soalan saya, apakah ubat yang sesuai unt... Baca lagi... Dijawab	2 jam
N Sulit	**Saringan covid untuk warga emas** ayah saya berumur 84 tahun mempunyai latar b dia hilang selera makan dan kelethan tiba2.tiad dgn positif case covid yg dike... Baca lagi... Dijawab	
E Sulit	**Demam, Selsema dan Sakit Tekak** I am 23 years old, female. I have allergic rhinitis extreme cold and hot places and in the evening Then I took a shower... Baca lagi... Dijawab	
	Selsema dan demam, apa tindakan seterusnya	19 jam

Fig. 3. Covid-19 Health Forum Platform, DoctorOnCall. (Malay Version)

Instead of copying and pasting the web contents manually, a web scraping tool called Octoparse[2] is utilized. It is an automated web data extraction tool written in .NET with API integration for grabbing public information from almost any website via HTTP. Generally, the scraping process involves three main steps which are API connecting, parsing, and extraction. The scraping process begins with entering the link of the website from which the data needs to be extracted, where an HTTP request is sent, and this returns the contents of the webpage in a response. After that, the parses offered by this tool will convert the website into a nice tree structure of HTML. This tree structure will then go over by XPath generator to locate and extract the required information precisely. The extracted information is then cleaned and converted into structured data formats such as Excel, CSV, Text, and others.

The full English version of questions and answers as shown in Fig. 4 is planned to be used as a training set, while the Malay version questions as shown in Fig. 5 will be filtered to choose the question that contains Malay-English mixed-language, to be used as testing dataset. Besides the answers set from the DoctorOnCall website, we have also planned to get the answers from trustworthy data sources such as journals from PubMed and FAQs from WHO, as shown in Fig. 6. All the datasets will undergo pre-processing to remove noise to ensure the research procedures can be conducted smoothly and successfully. The dataset will also be evaluated by experts to make sure of the consistency of the questions and answers crawled from the dataset.

Fig. 4. English version question and answer dataset. (Training data)

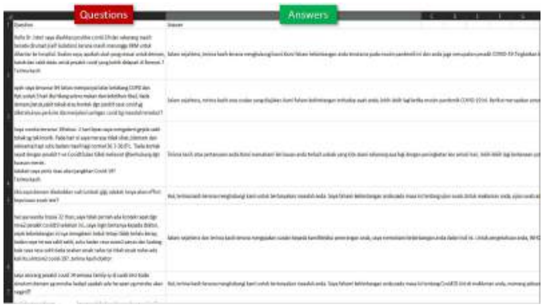
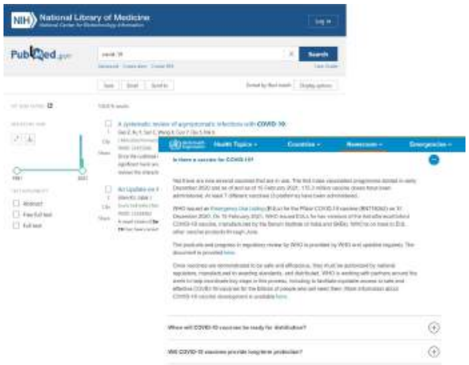


Fig. 5. Malay version questions dataset and answer as references. (Testing data)



NIH National Library of Medicine
Log in
PubMed.gov
Search
Health Topics
Countries
Newsroom
Emergencies
Is there a vaccine for COVID-19?

Fig. 6. Another Answer Data Sources from PubMed and World Health Organization (WHO).

B. Phase 1: Question Processing

The first phase of this research is to analyze and process the user generated questions, where the Malay-English mixed-language questions are translated into monolingual form. The mixed-language question will be represented in full English

[2] https://www.octoparse.com/

form at the end of this phase. This is considered as the main contribution of the project because we found out that Malay language is a low resource language, and thus we planned to utilize a high resource language (English) to overcome this issue. The major steps of design and implementation of the proposed translation module are:

1) Language Identification: : Identify the language of each word in the Malay-English mixed-language question. Before undergoing the translation process, the Malay-English mixed-language question is fed into a language identification module that helps in tokenizing a mixed-language sentence. Due to the aforementioned lack of POS tagger, Chunkers and Parser in Malay language, language identification is conducted based on word-level. The output of this step is a sequence of mixed-language words annotated with their corresponding language. An example of the language identification output is shown below, where each bracket denotes the token of the words, where ms refers to Malay and en refers to English:

“(di)$_{ms}$ (mana)$_{ms}$ (sy)$_{ms}$ (boleh)$_{ms}$ (mengambil)$_{ms}$ (test)$_{en}$ (utk)$_{ms}$ (coronavirus)$_{en}$”

2) Text Normalization: Normalize and standardize the detected Malay words as well as English words that are expressed in short forms, slang forms or misspelling, into the correct words. The normalization is conducted by using a rule-based approach and the set of normalization rules are modified based on previous work done by [11]. Table II shows the normalization rules by [11].

TABLE II. NORMALIZATION RULES BY [11].

Process	Rules	Example	
		Before	*After*
Change word into lowercase	PQR → xyz	Ujian	ujian
Elimate the word -nya and ny	PQRnya → PQR	sedapnya	sedap
	PQRny → PQR	panasny	panas
Split the words la, lah and lh.	PQRla → PQR la	marila	mari la
	PQRlah → PQR lah	terimalah	terima lah
	PQRlh → PQR lh	sayalh	saya lh
Elimate the words with 2 or hyphens	PQR2 → PQR	pokok2	pokok
	PQR^2 → PQR	ayam^2	ayam
	PQR-PQR → PQR	itik-itik	itik
Elimate duplicate characters	PPQQRR → PQR	takkkkk	tak
Split words that have simialr terms into two groups	PQPQPQPQ → PQPQ	hahaha	haha
Divide combine words with two different meaning	PQPR → PQ PR	takpayah	tak payah
Transform typographical words into actual terms	PR → PQR	pyh	payah
	PQRx → PQR	cantix	cantik

3) Lexical Translation: Translate the Malay-English mixed-language question into a monolingual question in full English form as shown as below.

“where can i take a test for coronavirus”

The output of this phase is the translated mixed-language question (Translated from mixed-language to English). It will be used in Phase 2 for classification to predict the answer type and retrieve the list of candidate answers for the question.

C. Phase 2: Answer Processing

In this phase, the mixed-language question classification is done by using deep learning by using the Convolutional Neural Network (CNN). Convolution Neural Network (CNN) is rarely applied in Natural Language Processing (NLP) as it is not completely intuitive. Through the initial literature review, many authors argued that SVM and RNN have low accuracy in classifying the mixed-language questions due to the small number of datasets available. This problem was overcome by Barnerjee and his team [25] through the combination of the feature engineering with CNN towards Bengali-English mixed-language questions. Hence, this research is investigating whether CNN without additional feature engineering, as Malay is a low resource language, can overcome the scarcity of Malay-English mixed-language questions.

The translated question is fed into an Embedding Layer (the first layer of the Convolutional Neural Network) where each word in the translated mixed-language question (output of previous phase) is converted into a set of the vector called Word Embedding. It then undergoes several different layers of featurization and lastly the display output is the predicted answer type such as Origin, Symptoms, Prevention, Transmission, Treatment and so on for the given mixed-language question, which will be used for retrieving a list of candidate answers in next phase, as shown in Fig. 7. The approach in this phase is decided conceptually based on the literature review and comments by other researchers.

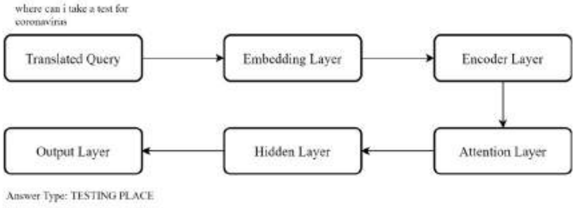


Fig. 7. Question Classification by Convolutional Neural Network (CNN) in Phase 2.

The detailed process of classification is described once the deeper exploration and initial experiment is done, to compare each of the methods that are mentioned by other researchers to investigate the novelty part to be added, and to optimize the most suitable parameter for the chosen approach. To evaluate the performance of the proposed classification model, we apply the evaluation metrics which are usually utilized in classification tasks: Accuracy (Acc), Precision, Recall and F-measure, which will be discussed in Section E.

D. Phase 3: Answer Extraction

Phase 3 is the last phase of the research experiment. In this phase, the best answers are generated to answer the mixed-language question. The major steps for designing and implementing the retrieval model are:

1) Named Entity Recognition (NER): The collection of answers from the DoctorOnCall website and the journals from PubMed are passed into a Named Entity Recognition (NER) method for identifying the entity type of each word as well as each answer.

2) Check and Retrive: Check and retrieve the NER data (answers collection) that match with the answer type of the mixed-language question as a list of candidate answers.

3) Answer Ranking: The candidate answers are ranked by using the ranking model such as Vector Space Model, Boolean Model or Machine Learning Model, to obtain the best answer.

Similar to the previous phase, a deeper exploration and decision on what method to be used in the retrieval model is done. The evaluation of performance of the retrieval model uses the sane metrics as the previous phase which are accuracy, precision, recall and F-measure. Another evaluation carried out is an expert evaluation, where general practitioners in medical/medicine field are asked to assess the automated answers created by the system.

E. *Performance Evaluation*

In this section, the evaluation that to be conducted on our proposed framework is described. From the previous section, it can be observed that the evaluation is involved in Phase 2 and Phase 3. Phase 2 cover the experiment concerning the classification of Malay-English mixed-language questions by using the Convolutional Neural Network (CNN), whereas Phase 3 is the retrieval model that generate the best answer for the Malay-English mixed-language question. The evaluation metrics often used in classification tasks, which are Precision, Recall, F-measure, and Accuracy are utilized in this research. Besides this, there is additional evaluation for Phase 3 which is an expert evaluation as described below:

1) Expert Evaluation: The final output of the research will be the best answer for the Malay-English mixed-lagnuage questions. These best answers will be evaluated by general practitioners in the medical/medicine field to validate and judge the system generated answer for the particular mixed-language question.

2) Automatic Evaluation: The overall performance of the proposed classification model and retrieval model for Malay-English mixed-language questions will be calculated by accuracy (Acc), while the class specific performances will be measured using standard evaluation measures Precision, Recall and F-measure.

$$Acc = \frac{no.\ of\ question\ that\ classified\ correctly}{Total\ no.\ of\ questions}\ x\ 100\% \quad (1)$$

$$Precision = \frac{TP}{TP+F} \quad (2)$$

$$Recall = \frac{TP}{TP+F} \quad (3)$$

$$Fmeasure = \frac{2*precision*recall}{precision+rec} \quad (4)$$

IV. Conclusion

Mixed-language is now commonly used in social media in multilingual community countries such as Malaysia. However, there is little discussion about Malay-English mixed-language in computational study, especially the use of both languages in Question Answering (QA) systems. The vast majority of automated QA systems that exist today only support single language solutions, and this has posed some inconveniences towards bilingual users.

Therefore, in this paper, we presented a conceptual framework for a Malay-English mixed-language question answering system. The proposed framework is mainly aimed to solve the problem of Malay-English mixed-language questions by the users in an online forum. And through the combination of machine translation model and a deep learning question classification model (CNN), we believe that this would give a diverse selection of answers for users. We also believed that this combination model can reduce the data complexity and improve the performance of the mixed-language question answering system. It is hoped that this research will provide a significant contribution for multilingual forum platforms and also to intelligent QA systems (chatbots).

This research, however, is subject to several limitations. The lack of available Malay-English mixed-language data for a question answering system is one such limitation. As a result, we had to confine our analysis to the health domain as well as the Covid-19 domain. Moreover, the language identification stage may rely on the existing Malay corpus, which may lead to identification errors. Furthermore, the framework may be limited to the Malay-English language pair, as the rules produced for text normalization are applicable to Malay or English terms. A further study could access the data construction for the Malay-English mixed-language question answering process in open domains, and an investigation can be done on the integration of language identification with text normalization to overcome the identification errors.

Acknowledgment

The authors would like to express the deepest appreciation to the Ministry of Education Malaysia for financially supporting this research. Additional supports in terms of facilities were also provided by School of Computing (SC), Faculty of Engineering, Universiti Teknologi Malaysia (UTM).

References

[1] N. Bukhari, A. F. Anuar, K. M. Khazin, and T. Abdul, "English-Malay Code-Mixing Innovation in Facebook among Malaysian University Students," *Artic. Res. World-Journal Arts Sci. Commer.*, vol. 6, no. 4, pp. 1–10, 2015.

[2] E. Kasmuri and H. Basiron, "Segregation of Code-Switching Sentences using Rule-Based Technique," *Int. J. Adv. Soft Compu. Appl*, vol. 12, no. 1, 2020.

[3] Nurhayati-Wolff, P., &amp; 28, J. (2020, July 28). Malaysia number of internet users. Retrieved March 29, 2021, from https://www.statista.com/statistics/553752/number-of-internet-users-in-malaysia/

[4] N. I. Zabha, Z. Ayop, S. Anawar, E. Hamid, and Z. Z. Abidin, "Developing Cross-lingual Sentiment Analysis of Malay Twitter Data Using Lexicon-based Approach," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 1, pp. 346–351, 2019.

[5] K. C. Raghavi, M. K. Chinnakotla, and M. Shrivastava, "'Answer ka type kya he?' Learning to Classify Questions in Code-Mixed Language," in *Proceedings of the 24th International Conference on World Wide Web*, 2015, pp. 853–858.

[6] B. R. Chakravarthi, N. Jose, S. Suryawanshi, E. Sherly, and J. P. McCrae, “A Sentiment Analysis Dataset for Code-Mixed Malayalam-English,” *arXiv Prepr. arXiv2006.00210*, 2020.

[7] N. H. Mahadzir, M. F. Omar, and M. N. M. Nawi, “Semantic similarity measures for Malay-English ambiguous words,” *J. Telecommun. Electron. Comput. Eng.*, vol. 10, no. 1–11, pp. 109–112, 2018.

[8] K. S. N. Tan, T. M. Lim, and Y. M. Lim, “EMOTION ANALYSIS USING SELF-TRAINING ON MALAYSIAN CODE-MIXED TWITTER DATA,” in *International Conferences ICT, Society, and Human Beings 2020; Connected Smart Cities 2020; and Web Based Communities and Social Media 2020*, 2020, pp. 181–188.

[9] D. Gupta, S. Tripathi, and P. Bhattacharyya, “A Hybrid Approach for Entity Extraction in Code-Mixed Social Media Data,” *MONEY*, vol. 25, no. 66, 2016.

[10] K. Soman, “Amrita-CEN@MSIR-FIRE2016: Code-mixed question classification using BoWs and RNN Embeddings,” in *FIRE (Working notes)*, 2016, pp. 122–125.

[11] Siti Noor Allia Noor Ariffin and Sabrina Tiun, “Rule-based text normalization for malay social media texts,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 10, pp. 156–162, 2020.

[12] S. Thara and P. Poornachandran, “Code-Mixing: A Brief Survey,” in *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2018, pp. 2382–2388.

[13] E. G. Bokamba, “Are there syntactic constraints on code-mixing?,” *World Englishes*, vol. 8, no. 3, pp. 277–292, Nov. 1989.

[14] H. Abu-Krooz, Q. O. Al-Azzawi, and M. M. Saadoon, “Code switching and code mixing: A sociolinguistic study of Senegalese international students in Iraqi colleges,” *J. Univ. Babylon Humanit.*, vol. 26, no. 3, pp. 112–123, 2018.

[15] P. Muysken and P. C. Muysken, *Bilingual speech: A typology of code-mixing*. Cambridge University Press (CUP), 2000.

[16] G. Sreeram and R. Sinha, “Exploration of End-to-End Framework for Code-Switching Speech Recognition Task: Challenges and Enhancements,” *IEEE Access*, vol. 8, pp. 68146–68157, 2020.

[17] A. Bohra, D. Vijay, V. Singh, S. S. Akhtar, and M. Shrivastava, “A Dataset of Hindi-English Code-Mixed Social Media Text for Hate Speech Detection,” in *Proceedings of the second workshop on computational modeling of people’s opinions, personality, and emotions in social media*, 2018, pp. 36–41.

[18] S. Thara, E. Sampath, and P. Reddy, “Code mixed question answering Challenge using deep learning methods,” in *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, 2020, pp. 1331–1337.

[19] G. Sreeram and R. Sinha, “Language Modeling for Code-Switched Data: Challenges and Approaches,” *arXiv Prepr. arXiv1711.03541*, 2017.

[20] C. W. Pfaff, “Constraints on Language Mixing: Intrasentential Code-Switching and Borrowing in Spanish/English,” *Language (Baltim).*, pp. 291–318, 1979.

[21] C. Tsoukala, S. L. Frank, A. Van Den Bosch, J. V. Kroff, and M. Broersma, “Modeling the auxiliary phrase asymmetry in code-switched Spanish-English,” *Biling. Lang. Cogn.*, pp. 1–10, 2020.

[22] T. Solorio and Y. Liu, “Part-of-Speech Tagging for English-Spanish Code-Switched Text,” in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 2008, pp. 1051–1060.

[23] J. Y. C. Chan, P. C. Ching, and T. Lee, “Development of a Cantonese-English Code-mixing Speech Corpus,” in *Ninth European Conference on Speech Communication and Technology*, 2005.

[24] M. Choudhury, A. Srinivasan, and S. Dandapat, “Processing and Understanding Mixed Language Data,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): Tutorial Abstracts*, 2019.

[25] S. Banerjee, S. Naskar, P. Rosso, and S. Bandyopadhyay, “Code mixed cross script factoid question classification-A deep learning approach,” *J. Intell. Fuzzy Syst.*, vol. 34, no. 5, pp. 2959–2969, 2018.

[26] S. Mandal and D. Das, “Analyzing Roles of Classifiers and Code-Mixed factors for Sentiment Identification,” *arXiv Prepr. arXiv1801.02581*, 2018.