Malay Online Virtual Integrated Corpus (MOVIC): A Systematic Review

Normi Sham Awang Abu Bakar, Hamwira Yaacob,
Department of Computer Science
International Islamic University Malaysia,
Kuala Lumpur, Malaysia
{nsham,hyaacob}@iium.edu.my

Dini Handayani,
School of Computing and IT
Taylor's University
Selangor, Malaysia
dinioktarina.dwihandayani@taylors.edu.my

Mustafa Ali Abuzaraida
Computer Science Department
Faculty of Information Technology
Misurata University, Libya
abuzaraida@it.misuratau.edu.ly

***Abstract*— The development of the various Malay corpora have given the opportunities to many researchers to explore their usage in diverse contexts. However, the corpora were distributed in various locations, and for the ease of access for users, a system called Malay Online Virtual Integrated Corpus (MOVIC) is proposed. This paper focuses on applying the systematic literature review (SLR) on the Malay corpus research to find out the recent development in the area. From the initial search, 3231 articles were extracted from five online databases, such as, IEEE Xplore, Scopus, ProQuest, Springer Link, and ACM. After several rounds of filtering, 11 papers were selected for review.**

Keywords*— *Development, Online Malay corpus, Integrated, Applications

I. INTRODUCTION

The recent outburst of opinions on social media, like Facebook and Twitter has motivated a large number of research on Sentiment Analysis or Opinion Mining. Sentiment analysis (SA) is a computational study of opinions, sentiments, emotions, and attitude expressed in texts towards an entity. It involves the tasks of detecting, extracting and classifying opinions, sentiments and attitudes concerning different topics, as expressed in textual input [1].

One of the sources of data for the SA research is using words from various text corpora. Text corpora are important because it can provide empirical data for researchers in various fields [2]. As such, the study of Malay corpora is essential to capture the trends or opinion among Malay speakers, locally or internationally.

The aim of this paper is to summarize and assess the current scientific literature on the Development of Online Malay Corpus through systematic literature review (SLR). SLRs are ways of aggregating knowledge about any topic or research question [3]. The SLR methodology should be as neutral as possible by being auditable and repeatable. SLRs are reported as secondary studies and the studies they analyse are known as primary studies [3].

This paper is arranged as follows: Section II presents the background of the study, while Section III elaborates the methodology used for the SLR. In addition, the findings are discussed in Section IV, and Section V highlights the conclusion and future work.

II. BACKGROUND OF STUDY

The study of Malay corpora has been investigated by a number of researchers, and for various purposes [2]. As a result, many Malay word corpora have been developed to cater for the various needs of the researchers. The examples of the existing Malay corpora are given in Table 1:

TABLE 1. MALAY TEXT CORPORA

Corpus	Context	Application	Provider
sealang.net	General	Collocation analysis Concordance Provide with example sentences from multiple sources	CRCL and University of Wisconsin-Madison Center for Southeast Asian Studies (CSEAS)
ms.oxforddictionaries.com	General/ education	Translating and define word from english to malay, malay to english	Oxford University Press
Mcp.anu.edu.au(Malay concordance project)	Education	Concordance	Australian National University
MyBaca.org	Education	To find words whether it is capable in starts,end or middle of the sentences	School of educational studies

978-1-5386-7525-0/18/$31.00 ©2018 IEEE
DOI 10.1109/ICT4M.2018.00052

Sbmb.dbp.gov.my (Korpus Dewan Bahasa dan Pustaka)	Education	Concordance Collocation analysis	Dewan Bahasa dan Pustaka
prpm.dbp.gov.my/ (Pusat Rujukan Persuratan Melayu)	Education	Define and provide detailed meaning of word.	Dewan Bahasa dan Pustaka

In general, all of the prior research only work on a single corpus and the multiple independent corpora are not integrated. Thus, there is a need to have an integrated corpus that will combine the important functions of the various corpora. The Malay Online Virtual Integrated Corpus (MOVIC) is proposed to address this gap and the main feature of this system is the integration part. In addition, this system will be developed as a web application, hence, the word "virtual" means available online.

The existing Malay corpora serve as the possible data source for our work, as the main objective is to develop an integrated Malay corpora that will ease the search of Malay words in one repository.

Essentially, once completed, MOVIC can be used in any domain that uses Malay terms because its applicability is not restricted to just Sentiment Analysis.

III. METHODOLOGY

Articles on the Development of Online Malay Corpus and its application are scattered across journals of various disciplines. Accordingly, the systematic review was performed using five online databases including IEEE Xplore, Scopus, ProQuest, Springer Link, and ACM.

In this paper, we use a systematic approach known as the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) [4], which includes the following steps:

1) Finding the number of record from several selected online databases;
2) Screening process, which involves the exclusion of the duplicate and unrelated articles;
3) Eligibility process, where the most relevant articles to the topic were selected.

The procedures used in the articles extraction are discussed in the following sub-sections. In addition, our article selection criteria and filtering process are also explained.

A. Articles Extraction Procedure

Our articles search was done using the basic search settings, where we input the search terms and phrases, such as: Development of Online Malay Corpus and its application.

The initial search resulted in 3231 articles. 898 articles were identified from IEEE Explore, while 2 articles were identified from scopus. On the other hand, 145 articles were identified from ProQuest, 158 were identified from SpringerLink, and 2028 articles were identified from ACM. our search process was further refined based on some pre-determined criteria. The following section illustrates our filtering/reviewing process.

Filtering/Reviewing Process

In order to perform this process, each article was manually reviewed in four stages, as shown in Fig. 1.

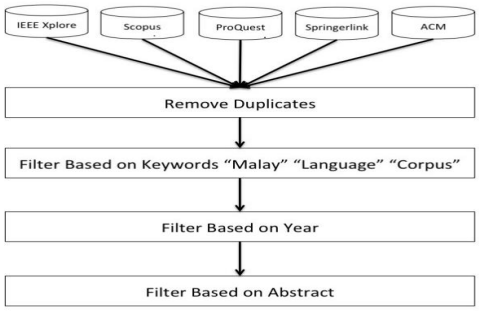

Fig. 1. Search strategy in Development of Malay Corpus and its Application according to the PRISMA statement (Source: [4])

1. During the first stage of review, originally, we managed to obtain 3231 articles. However, due to the duplication of articles, 106 articles were excluded. We were left with 3125 articles.

2. Our second stage of review includes doing the scanning of titles, manually. At this stage, 2756 articles do not have the keywords of "Malay", "Language" and "Corpus", therefore, excluded. At the end of this stage, we have 369 articles to be reviewed.

3. In the third stage, we chose articles that were published between 2008 and 2018. This 10-year period could be considered to correspond to the research period of

Development of Malay Corpus. At the end of this stage, we have 214 articles to be reviewed.

4. The final stage includes reading the abstract and analyzing each article according to our keywords. As a resut, 11 full text articles about Development of Malay Corpus were identified from the remaining articles.

The diagram in Fig. 1 shows the procedure used to filter and extract articles that meet our pre-determined criteria.

IV. FINDINGS

The main aims of the literature search were to extract information on several categories, such as the *development techniques* of the online Malay corpora, *features* that are included in the corpora, and the *domain* they were applied.

Generally, from the development techniques aspect, the researchers use different techniques to build the Malay corpus in relation to their needs. As such, in order to build their word corpus, Lee & Low [5] [6] used the Malay language textbooks used in the Malaysian primary schools as their primary source. They study the differences in the textbook designs across all the textbooks used.

In addition, Asyafi'ie et al [2] investigated the methods of extracting the corpus using the Friday sermon transcripts, obtained through the official government website. The program was developed using Personal Home Page (PHP) and the data were then stored into Sequential Query Language (MySQL).

A recent work by Hijazi et al [7] discuss about the method of selecting the corpora from 443 Facebook and Twitter posts. The corpus then were manually commented as positive, negative, or neutral by 3 annotators. Furthermore, Wang et al [8] proposed three approaches:

(1) word-level paraphrasing using confusion networks

(2) phrase-level paraphrasing using pivoted phrase tables.

(3) adaptation using a specialized text rewriting decoder

In their work, Saloot et al [9] attempted to create a corpus called the Malay Chat-style-text Corpus (MCC) that consists of 1 million Twitter messages, containing 14,484,384 word instances and 646,807 vocabularies.

In a similar line, Nicholson & Baldwin [10] applied a Malay–English translation dictionary, KAMI, with a limited very low frequency entries of 1.2M tokens of Malay text. In their work, they used a machine learning model and lexicon that do not depend on the programming languages. In order to build the Malay classifiers, they built a feature vector for each headword in the corpus.

A study by van Minde [11] discusses the usage of the Malay word ***yang*** in the linguistics context. The asserted that ***yang*** can be used in multiple contexts in the various areas. Similarly, Chung [12] uses a corpus-based approach in examining the prefix ***ter-*** in using a modern and a historical corpus. Amad Darwis et al [13] suggest an approach that can extract a valid root word by cutting the matched affixes. In addition, they also suggested the use of a Malay Word Register to handle the ambiguity problem of determining the correct root word. Lim et al [14] show how multi-lingual lexicons with under-resourced languages can be constructed using simple bilingual translation lists.

Next, we choose to analyse the features highlighted in the papers. The work of Lee & Low [5][6] highlighted the linguistic properties of the Malay language textbooks, such as: Frequency of occurrence in the sources, word length, phoneme length, number of syllable, type of inflection, word category and syllable structure. Asyafi'ie et al [2] focus on the phenome distribution of the fifty two Friday sermon while Hijazi et al [7] investigated the corpora from Facebook and Twitter posts in Sabah language.

A recent work of Wang et al [8] investigates the Indonesian/ Malay–English translation using the large adapted resource-rich bitext, and Saloot et al studied the method to study the normalization of the Malay Twitter messages based on corpus-driven analysis [9]. Similarly, Nicholson and Baldwin explained about the prediction of preferred count classifiers for nouns in Malay [10]. In general, these research papers mostly discuss the concordance in Malay words like ***ter*** [11] and ***yang*** [12].

Lastly, we look at the domains/genres where the research were applied and they are education [5][6], religion [2], social media [7][8][9], linguistics [10][11][12][13][14]. In particular, the researchers in [5][6] discuss about the Malay Language word corpus for primary school where the researchers in [2] investigate the Friday sermon transcript. The work in the social media domain include the sentiment-lexicon construction [7], building statistical machine translation systems for resource-poor languages [9] and to find the morphological features of the Malay words [10].

Most of the research work in the literature list belong to the linguistics domain. For example, [10] examine the capacity of Web and corpus frequency methods to predict preferred count

classifiers for nouns in Malay language. In addition, the work in [11][12] discuss the usage of the ***ter*** and ***yang*** in the Malay corpora while [13] look at the algorithm for word stemming.

From the literature, we found that most research in Malay language are in the area of linguistics and none of them work in the sentiment analysis domain. Thus, it is imperative that we choose the correct method to develop our system. All the prior work only work using one corpus and the challenge for us is to find the best way to combine the corpora.

In our project, most of the work would involve looking at the best methodology to combine several corpora and investigate the possibility to integrate several independent properties from the various corpora into one integrated system. Essentially, we plan to use our MOVIC project for the Sentiment Analysis domain and the development method would be using the Python language to crawl the selected corpora. Among the features that we plan to have are:

1) Malay words which will be integrated from multiple sources/corpora,
2) The combined information from the different corpora, i.e. the meaning, concordance, etc.

The summary of the literature is given in Table 2.

VI. CONCLUSION AND FUTURE WORK

The systematic literature review process has shed some lights on the direction of the research in the area of Malay corpus development. From the discussions in the reviewed papers, there are several development methodologies that we can try to implement in order to understand how corpus is developed, especially, the Malay corpus.

The future works of this research include the implementation of the most suitable development methods to develop the Malay Online Virtual Integrated Corpus (MOVIC) system. It is hoped that the completion of the system will ease the access to the wealth of Malay words.

ACKNOWLEDGMENT

This research is fully funded by the International Islamic University's RIGS grant, ID: RIGS-17-156-0731.

REFERENCES

[1] W. Medhat, A. Hassan and H. Korasky, "Sentiment analysis algorithms and applications: A survey". Ain Shams Engineering Journal, Vol. 5, pp 1093-1113, May 2014.

[2] M. A. Asyafi'ie,, M. Harun, M. I. Shapiai, and P. I. Khalid, "Identification of phoneme and its distribution of Malay language derived from Friday sermon transcripts," In *2014 IEEE Student Conference on Research and Development (SCOReD),* (pp. 1-6), 2014.

[3] B. Kitchenham, R. Pretorius, D. Budgen, O. P. Brereton, M. Turner, M. Niazi, S. Linkman, "Systematic literature reviews in software engineering – A tertiary study," Information and Software Technology, Vol. 52, Issue 8, pp 792-805, 2010.

[4] D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman, "Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement.," PLoS Med., vol. 6, no. 7, p. e1000097, July 2009.

[5] L. W. Lee and H. M. Low, "Developing an online Malay language word corpus for primary schools," *International Journal of Education and Development using Information and Communication Technology*, *7*(3), 96-101.

[6] L. W. Lee and H. M. Low, "The development and application of an online Malay language corpus-based lexical database," *Kajian Malaysia*, *32*(1), 151-166.

[7] M. H. A. Hijazi, L.Libin, R. Alfred, and F. Coenen, "Bias aware lexicon-based Sentiment Analysis of Malay dialect on social media data: A study on the Sabah Language," In IEEE *2nd International Conference on Science in Information Technology (ICSITech),* (pp. 356-361). *October 2016*.

[8] P. Wang, P. Nakov, and H. T.,Ng "Source language adaptation approaches for resource-poor machine translation," Computational Linguistics, 42(2), pp. 277-306, 2016.

[9] M. A. Saloot, N. Idris, and R. Mahmud, "An architecture for Malay Tweet normalization," Information Processing & Management, 50(5), pp.621-633, 2016.

[10] J. Nicholson and T. Baldwin, "Web and corpus methods for Malay count classifier prediction". In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, pp. 69-72, May 2009.

[11] D. van Minde, "The pragmatic function of Malay ***yang***," Journal of Pragmatics (JoP), 40(2008), pp1982-2001, 2008.

[12] S. F. Chung, "Uses of ter- in Malay: A corpus-based study," Journal of Pragmatics (JoP), 23, pp. 719-813, 2011.

[13] S. Amad Darwis, R. Abdullah, and N. Idris, "Exhaustive Affix Stripping And A Malay Word Register To Solve Stemming Errors And Ambiguity Problem In Malay Stemmers," Malaysian Journal of Computer Science. Vol. 25(4). pp 196-209, 2012.

[14] L. T.,Lim, L. K. Soon, Lim, T.Y. Tang, T. K., B. Ranaivo-Malancon, "Lexicon+TX: rapid construction of a multilingual lexicon with under-resourced languages," Lang Resources & Evaluation. Vol 48. pp 479-492, 2014.

TABLE 2. TAXONOMICAL TABLE ON MALAY CORPORA RESEARCH

No	Reference	Development Techniques	Features	Application/Domain
1	L. W. Lee and H. M. Low (2011)	The Malay language textbooks used in the Malaysian primary schools are the primary sources to build the word corpus.	Linguistic properties: • Frequency of occurrence in the sources • Word length • Phoneme length • Number of syllable • Type of inflection • Word category • Syllable structure	Malay Language Word Corpus for primary schools. (Education)
2	L. W. Lee and H. M. Low (2014)	Develop a lexical database of Malay words commonly encountered by elementary school children in Malaysia.	The database has an interactive interface that allows users to search in real-time primary linguistic features such as word frequency, word length, phoneme length, number and type of syllables, and word category.	Malay Language Word Corpus for primary schools. (Education)
3	M. A. Asyafi'ie,, M. Harun, M. I. Shapiai, and P. I. Khalid (2014)	The Friday sermon transcripts were obtained through the official government website and then standardized by removing images and foreign letters; expanding acronyms and short forms; converting numbers and symbols to appropriate Malay words. The program was written using Personal Home Page (PHP) and the data were then stored into MySQL (Sequential Query Language).	Phoneme distribution	Friday sermon transcripts (Religion)
4	M. H. A. Hijazi, L.Libin, R. Alfred, and F. Coenen. (2016)	Corpora selected from 443 Facebook and Tweet posts. Manually annotated as positive, negative or neutral by 3 annotators	Sabah language	Sentiment-Lexicon construction (Social media)
5	P. Wang, P. Nakov, and H.	Proposed three approaches:	Indonesian/ Malay–English translation using the large	A useful guideline for building statistical machine translation

	T.,Ng (2016)	(1) word-level paraphrasing using confusion networks. (2) phrase-level paraphrasing using pivoted phrase tables. (3) adaptation using a specialized text rewriting decoder	adapted resource-rich bitext	systems for resource-poor languages.(Social media)
6	M. A. Saloot, N. Idris, and R. Mahmud, (2014)	Gathered 1 million Twitter messages, consisting of 14,484,384 word instances and 646,807 vocabularies, and named it the Malay Chat-style-text Corpus (MCC).	Normalize the Malay Twitter messages based on corpus-driven analysis	To find the morphological features of the Malay word (Social media)
7	J. Nicholson and T. Baldwin (2009)	Use a machine learning model for Malay classifiers, designed to be language independent	Predict preferred count classifiers for nouns in Malay	Examine the capacity of Web and corpus frequency methods to predict preferred count classifiers for nouns in Malay (Linguistics)
8	D. van Minde (2008)	Explores the use of ***yang*** in various regional dialects and from various stages in the development of Malay words.	Corpus from the Malay Concordance Project (MCP, Australian National University, Canberra).	The usage of ***yang*** in older Malay and young Indonesian text (Linguistics)
9	S. F. Chung (2011)	Uses a corpus-based approach in examining the prefix ter- in using a modern and a historical corpus	All instances of ter- were compared in terms of their distribution in two corpora (one modern and the other historical).	Study of usage of ter- in corpora (Linguistics)
10	S. Amad Darwis, R. Abdullah, and N. Idris (2012)	Propose an approach that exhaustively strips all matched affixes to ensure that a valid root word will be extracted Propose the use of a Malay Word Register to address the ambiguity problem of determining the correct root word.	Stemmer uses a rule based approach to produce a linguistic root word.	Algorithm for word stemming (Linguistics)
11	L. T.,Lim, L. K. Soon, Lim, T.Y. Tang, T. K., B. Ranaivo-Malancon, (2014)	Show how multilingual lexicons with under-resourced languages can be constructed using simple bilingual translation lists	Six member languages: English, Malay, Chinese, French, Thai and Iban	Multilingual (Linguistics)