Multidisciplinary Applied Research and Innovation Vol. 2 No. 3 (2021) 001-004
© Universiti Tun Hussein Onn Malaysia Publisher's Office

MARI

Homepage: http://publisher.uthm.edu.my/periodicals/index.php/mari
e-ISSN :2773-4773

Malay Roman Corpus Annotation System

Safwan Sufian Chang*, Juhaida Abu Bakar*, Norliza Katuk

School of Computing, UUM College of Arts and Sciences,
Universiti Utara Malaysia, Sintok, Kedah, 06010, MALAYSIA

*Corresponding Author Designation

DOI: https://doi.org/10.30880/mari.2021.02.03.001
Received 05 September 2021; Accepted 05 October 2021; Available online 15 December 2021

Abstract : The Malay Roman Corpus Annotation is a web-based Natural Language Processing system. The system was developed using Flask and is powered by Polyglot, a natural language pipeline. Polyglot supports multilingual applications and the Malay language is one of the supported languages in the library. There are many unstructured texts in the WWW resources and those texts are incomprehensible to computers. Then, text analysis takes longer time and is inefficient. Furthermore, these unstructured texts contain an excessive amount of information, such as people's names, places, and locations, which will almost always result in incorrect information being evaluated. Hence, this work is to define and extract Malay Roman Named Entity Recognition characteristics from an unstructured document. Besides, this work was also created to develop a system that is able to annotate Malay Roman by using a suitable approach. The system built able to help users extract information correctly. The method used to develop this work consists of 5 phases, which are sentence segmentation, tokenization, part of speech tagging, entity recognition, and relationship recognition. Manage users, manage text input, manage clear text, view entity labels, manage analyse text and manage results are the functionalities developed. This innovation can help news providers by automatically going through the entire articles and identifying the entities, which helps in categorizing articles and saves students time by helping them summarize the documents.

Keywords: Named Entity Recognition, Polyglot, language model

1. Introduction

Natural Language Processing (NLP) is a theoretically inspired set of computational techniques for the study and representation of naturally occurring texts at one or more stages of linguistic analysis. This technique is to achieve human-like language processing for a number of tasks or applications [1]. Named Entity Recognition (NER) is a part of the sub task of Information Extraction (IE), which is one of the fields that applies NLP technologies. NER is used to produce a more meaningful corpus by identifying proper names in the corpus and categorizing them into groups [2]. Information consists of various types, such as text, images, audio, and so one keeps increasing on the internet, which is largely unstructured. Hence, effective management and organization of information is the key strategy for

**Corresponding author: safwan_sufian_cha@soc.uum.edu.my, juhaida.ab@uum.edu.my* addressing the problem of finding useful information [3]. Therefore, this paper is being proposed. The objective of this study is to define and extract Malay Roman NER characteristics from an unstructured document, develop a system that is able to annotate Malay by using a suitable approach, and test the usability of the system.

2. Materials and Methods

This section consists of the materials and methods section, which describes all the necessary information that is required to obtain the results of the study.

2.1 Materials

The Malay Roman Corpus Annotation was developed based on the requirements gathered. The materials involved in this development are being developed using Flask, Python, HTML, SQL, PyCharm and Google Form.

2.2 Methods

The method used in this work consists of 5 phases (refer **Figure 1**), which are sentence segmentation, tokenization, part of speech (POS) tagging, entity recognition, and relationship recognition [4].

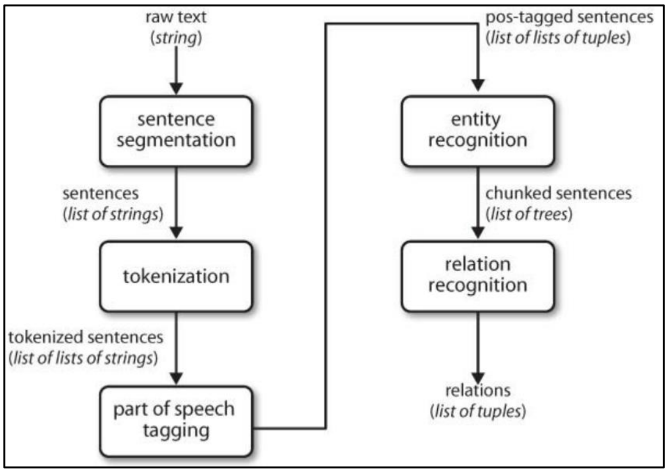

Figure 1. Pipeline architecture for information extraction system (*Reference*: [4])

The process begins by extracting raw text from a document which is later split into sentences using sentence segmentation. Next, the sentence or the list of strings should be further subdivided into words using a tokenizer. Tokenization is a common task in NLP. Tokenization is a process of separating a piece of information, which in this work shall be a list of strings or sentences, into smaller units called tokens. After that, each of the sentences was tagged with POS tags. The tags in POS will be very helpful in the next phase, which is NER. In the NER phase, we searched for entities in each sentence. Lastly, we searched for likely relations between those entities in the last phase, which is relation recognition.

3. Results and Discussion

3.1 Results

The Malay Roman Corpus Annotation (refer **Figure 2**) was developed so that the user can analyse their unstructured text or to help the user to categorize their article. The Malay Roman Corpus Annotation is a web-based NLP system. The web-based system was developed using Flask-Python and is powered by Polyglot. Polyglot is a natural language pipeline that was developed by Rami Al-Rfou [5]. It supports multilingual applications and the Malay language is one of the supported languages which has been used in the work. The system is able to manage users, manage text input, manage clear text, view entity labels, manage analyse text and manage results. The categories of NER that can be learned using Malay Roman Corpus Annotation are shown in **Table 1**. Person (PER) names, Organization (ORG) names, and Location (LOC) names are all addressed.

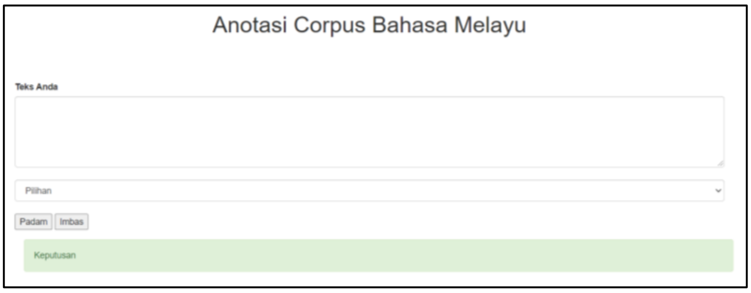

Figure 2. Main Page of Malay Roman Corpus Annotation

Table 1. Example of NER

NER Category	Example
Person (PER)	Safwan, Abu, Ali, Dr. Mahatir, Lorenzo
Organizations (ORG)	PNB, SPRM
Location (LOC)	Itali, Switzerland, Paris

A usability test was also conducted to identify the usefulness, ease of use, and satisfaction of the Malay Roman Corpus Annotation. The results are shown in the **Table 2** below.

Table 2. Usability Testing Result

Questions	Mode (Most Agree)	Percentage (%)
Usefulness of Malay Roman Corpus Annotation		
Malay Roman Corpus Annotation increases my productivity	25	78.125
Malay Roman Corpus Annotation enables me to categorize article more quickly.	27	84.375
Malay Roman Corpus Annotation saves me time when I use it.	31	96.875
Malay Roman Corpus Annotation does everything I would expect it to do.	25	78.125
Malay Roman Corpus is useful in overall.	26	81.25
Average Usefulness of Malay Roman Corpus Annotation	83.75	
Ease of Use of Malay Roman Corpus Annotation		
Malay Roman Corpus Annotation is easy to use.	28	87.5
Malay Roman Corpus Annotation is user friendly	25	78.125
Malay Roman Corpus Annotation is easy to learn how to use it.	26	81.25
I can use Malay Roman Corpus Annotation without written instructions.	24	75
I can use Malay Roman Corpus Annotation successfully every time.	26	81.25
Average Ease of Use of Malay Roman Corpus Annotation	80.625	
Satisfaction of Malay Roman Corpus Annotation		
I am satisfied with Malay Roman Corpus Annotation.	25	78.125
I would recommend Malay Roman Corpus Annotation to my friend.	24	75
Malay Roman Corpus Annotation works the way I want it to work.	24	75
I feel I need to have Malay Roman Corpus Annotation.	24	75
Malay Roman Corpus Annotation is wonderful and pleasant to use.	21	65.625
Average Satisfaction of Malay Roman Corpus Annotation	73.75	

3.2 Discussions

The usability test was conducted through Google Form. The participants for this test consisted of 30 students from across Malaysia. Based on the results, most of the respondents agree with those three aspects, which are as shown in **Table 2**. Three aspects are based on the Technology Acceptance Model, which are usefulness, ease of use, and satisfaction of Malay Roman Corpus Annotation. Based on the findings, it shows that usefulness gets a higher percentage compared to ease of use and satisfaction. Respondents believe that the functionalities of the system are satisfactory, but that the ease of use could be improved.

4. Conclusion

NER, a part of the subtask of IE, which is a field that applies NLP technologies, is used to identify named entities in open documents. To find appropriate techniques and methods to process and extract the essential knowledge contained in this information, Malay Roman Corpus Annotation is proposed. The objective of this paper is to define and extract Malay Roman NER characteristics from an unstructured document, develop a system that is able to annotate Malay by using a suitable approach, and test the usability of the system. To develop the system, a method consisting of 5 phases was used. The Malay Roman Corpus Annotation was successfully developed based on the functional requirements that were gathered. The system was developed using Flask Python and was powered by Polyglot, a natural language pipeline that supports multilingualism. The result of the usability test shows most of the respondents are satisfied. Hence, we can conclude the system is ready to be implemented. In the future, more entities will be added with the increase of Polyglot functions to provide more than three entities; person, location, and organization.

Acknowledgement

This research was supported by Ministry of Higher Education (MoHE) of Malaysia through Fundamental Research Grant Scheme for Research Acculturation of Early Career Researchers (RACER/1/2019/ICT02/UUM/1).

References

[1] E. Liddy, Natural Language Processing. In Encyclopedia of Library and Information Science, 2nd Ed. New York: Marcel Decker, Inc, 2001.

[2] R. L. Alfred, "Malay Named Entity Recognition Based on Rule-Based," International Journal of Machine Learning and Computing vol. 4, no. 3, 2014.

[3] S. A. Asmai, M. S. Salleh, H. Basiron, S. Ahmad, "An Enhanced Malay Named Entity Recognition using Combination Approach for Crime Textual Data Analysis," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 9, no. 9, pp. 474-483, 2018.

[4] S. Bird, E. Klein, and E. Loper, Natural language processing with Python: analyzing text with the natural language toolkit, O'Reilly Media, Inc., 2009.

[5] R. Al-Rfou, B. Perozzi, and S. Skiena, Polyglot: Distributed word representations for multilingual nlp. arXiv preprint arXiv:1307.1662, 2013.