Jurnal Elektronik Ilmu Komputer Udayana
Volume 10 No. 1, August 2021

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Person Named Entity Recognition in Balinese

Kenny Kurniadi[a1], Ngurah Agus Sanjaya ER[a2]

[a]Informatics Departement, Faculty of Math and Science, Universitas Udayana
Bali, Indonesia
[1]kennikurniadi99@gmail.com
[2]agus_sanjaya@unud.ac.id

Abstract

Named Entity Recognition (NER) is part of information extraction whose task is to classify text which is categorized into several classes such as names of people (figures), organizations, and locations. In this study, the authors propose making a NER identify the names of characters in Balinese language documents. This study will use a rule-based method (rule-based). Rules are build based on the morphological structure and linguistic meaning of Balinese names. The research conducted, that the system has an accuracy of 67.41%, precision of 83.42%, recall of 77.83%, and F-Score of 80.53%.

Keywords: *Named Entitiy Recognition, Natural Language Processing, Balinese Language, Rule-Based Approach, Information Extraction.*

1. Introduction

Considering the condition of the Balinese language, the existence of the Balinese language needs to be preserved, both in education and through digital technology. Given the development of technology, making many types of Balinese text documents published digitally, this certainly plays an important role in the preservation of the Balinese language by providing Balinese reading sources that can be accessed easily [1]. The various text documents in the Balinese language can be summarized or extracted so that the younger generation who are less motivated to read and the younger generation who want to learn the Balinese language, can understand the essence and study the document without having to read the entire document.

One way to extract the essence of a document is by performing information extraction. To do information extraction, several components are required, such as syntactic parsing, entity extraction, relation extraction. Named Entity Recognition (NER) is part of information extraction whose task is to classify text which is categorized into several classes such as names of people (figures), organizations, and locations [2].

Making NER to identify the names of figures in Balinese language documents is entity extraction, therefore the NER created can be used as a feature in the extraction of document information in Balinese. The extraction of document information in Balinese will be used to extract the extract of the document. Besides, the NER created can be used as a feature of Balinese text-preprocessing which makes it easier to digitize Balinese language that can be accessed and will not be lost. There is also other research has been done in making Balinese text-preprocessing features such as lemmatization [3] and stemming [4].

NER development for Balinese documents has never been done before. However, there are similar studies, where these studies make NER for specific languages (such as local languages) [5]–[7]. Research conducted in [5]–[7] utilizes a rule-based method, this study uses a rule-based method. This approach is based on grammatical rules derived from linguistic knowledge, and a list of names for complex entities to control precisely [8].

Defining rules can be determined through direct observation of the existing corpus, such as seeing based on the similarities of an entity. Also, rules can be made by considering the structure of morphology, syntax, and semantics [9]. As in research [6] utilizes the morphological structure (detects prefix and suffix in the prefix of the name) of the language to get the entity.

The structure of the morphology of Balinese names has been studied before [10], [11]. The morphological structure obtained is such as the article of the article is an element used to limit or modify a noun. The word clothing in Balinese naming is an element that precedes the name and can distinguish the gender of the owner of the name. The articles referred to are I and Ni [10], examples of the use of the article in Balinese names such as "I Gusti Putu Ardana" and "Ni Komang Ayu". Another morphological structure is that there is a birth order in the name of the Balinese. This aspect becomes a marker for a person to be in what order [11]. Examples of birth order words such as "Wayan", "Komang", "Made", and others. Examples of using birth order in a Balinese name are like "I Wayan Indrayasa".

In addition to the morphological structure, Balinese names also have a linguistic meaning, namely the meaning of names in Balinese society based on their structure or shape. The contextual meaning referred to in this paper is the meaning behind the Balinese name. Like the meaning of hope which means the meaning that contains the hope that the owner of the name is like the meaning of his name [10]. Examples of such names are "Raditya Putra" which means son of the sun, "Susila" which means good deeds.

This linguistic meaningful name is usually taken from the Sanskrit language, therefore this can also be used as a feature of detecting the names of Balinese figures. From this research, we can use the morphological structure and linguistic meaning to serve as rules in the rules-based approach method to detect the names of characters in Balinese documents.

In this study, the authors propose making a NER identify the names of characters in Balinese language documents. This study will use a rule-based method (rule-based). The purpose of this research is to create a NER system that can identify entities that focus on character names with sufficient accuracy. This research is expected to be used for the development of document information extraction in Balinese or features for digitizing Balinese.

2. Reseach Methods

This research consists of several stages, including text-preprocessing wherein the text-preprocessing process there are several processes, that is removing punctuation, and tokenizing text, the next stage is making rules for rule-based methods, system design, and results and discussion.

2.1. Text-Preprocessing

Text-preprocessing is the first stage for preparing the text in documents into data that can be processed into the next process. There are stages of text-preprocessing, generally case folding, tokenizing, and filtering. In this study, the case folding and filtering stages were not used because some features in the system were still needed, such as capital letters and filtering, which would eliminate some characters in the name. The Text-Preprocessing stage starts from removing punctuation, which is the process for removing punctuation marks such as "!#$%&'()*+-/:;" except comma punctuation, period punctuation eventually will be deleted after text divided into sentences because it is needed as a separator between the words concerned. Next is tokenizing where the text will be separated into a token, where the token is the words in the text.

2.2. Rule-Based

The resulting token from text-preprocessing will enter the stage of character name classification with NER using a rule-based method. The rules are formed based on the morphological structure and linguistic meaning of Balinese names. In Balinese, the morphological structure of Balinese names is divided into several signifying aspects, including gender, birth order, caste system and name abbreviations which are described in table 1. In addition to the morological structure there are linguistic meanings such as the meaning of hope (Dharma, Candra Dewi, Hapsari) towards that person, these meaningful names of hope come from the Sanskrit vocabulary which will be used as a features.

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Table 1. Features Based on Morphological Structure

Features	Explanation	Example
Gender	Gender-distinguishing word	I, Ni
Birth Order	A word that distinguishes birth order	Putu, Made, Wayan, Ketut
Caste System	The word that tells the caste of a person	Gusti, Cokorda, Dewa, Desak, Anak Agung
Abbreviations	Abbreviation for other feature	Md, I Gst., A.A.,

The following is an example of the rules formed from the aspects of the name markers of Balinese figures.

IF [word[index] in item for item in aspek_penanda] and word[index][0].isupper()

THEN names.append(word[index])

WHILE word[index+1].isupper()
THEN names.append(word[index+1])

The example of the sentence is “Mangku Gede Pura Goa Raja Taksaka inggih punika Mangku Nyoman Rawet, nguningan, patah sekadi pujawali-pujawali sane sampun-sampun ring pujawali…” then NER in accordance with the above rules will produce the output "Nyoman Rawet".

2.3. Design System

In this research, the system is built based on the rules that have been made to identify the names of characters from Balinese language documents. The following is the flow of the system built process.

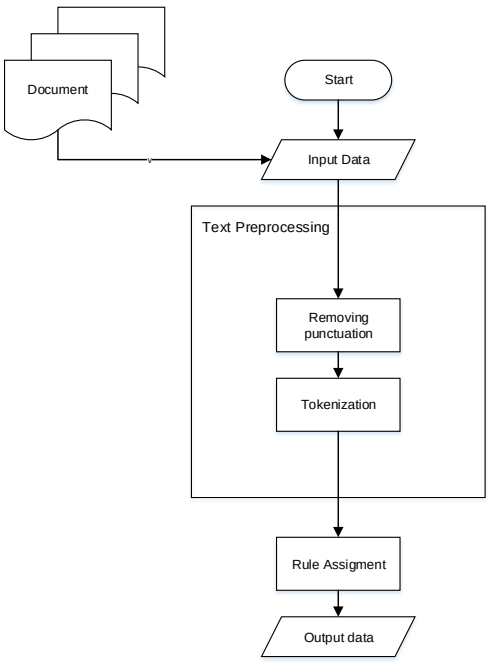

Figure 1. System Flow Chart

2.4. System Evaluation

After the NER has succeeded in identifying the name of the character in the Balinese language document, it is necessary to conduct an evaluation. The methods often used for evaluation are Precision, Recall, and F-score. The precision is calculated based on the number of correct classifications divided by the total classified by the system. Recall is calculated from the number of correct classifications divided by the total number of correct data. The value of the F-Score (or F-Measure) calculation is used to make comparisons between other classification models, the calculation of the F-Score using calculated precision and recall. Here are the equations of precision (1), recall (2), and F-Score (3).

$$\text{Precision} = \frac{\text{number of correct answer from system}}{number\ of\ system\ output} \quad (1)$$

$$\text{Recall} = \frac{\text{number of correct answer from system}}{\text{number of correct answer}} \quad (2)$$

$$\text{F} - \text{measure} = \frac{\text{Precision} * \text{Recall}}{0{,}5 * (\text{Precision} + \text{Recall})} \quad (3)$$

3. Result and Discussion

To prove that the NER work optimally, the authors will carry out the testing and evaluation phase. The evaluation will be carried out by calculating the Precision, Recall, and F-score based on the data obtained. The researcher used 50 Balinese language documents as test data consisting of various kinds of stories and news in Balinese which were saved in the .txt file format. The following are the test results which can be seen in table 3 and the evaluation results in table 4.

Tabel 3. Example of Testing Data

No	Input Sentence	System Extraction
1	Tiosan raris Ketua Gabungan Industri Pariwisata Bali Ida Bagus Agung Partha Adnyana maosan sampun masadu ajeng sareng Konsul Jenderal Tiongkok ring Denpasar	Ida Bagus Agung Partha
2	Asapunika kabaos olih Kepala Dinas Pariwisata Provinsi Bali Putu Astawa majeng juru warta ring kantor Dinas Pariwisata Provinsi Bali, Denpasar,	Putu Astawa
3	Sapunika kasobyang oleh Gubernur Bali Wayan Koster rikala pamungkah Bulan Bahasa Bali ring Taman Budaya Denpasar	Wayan Koster
4	Baan belogne ia adanina I Belog	I Belog
5	Ada tuturan satua I Lutung teken I Kekua	I Lutung, I Kekua
6	Disubané makejang kategul, Luh Ayu Manik Mas buin nylibsib ngalih tongos bet masalin raga dadi Luh Ayu Manik	Luh Ayu Manik Mas, Luh Ayu Manik
7	Ring salantang jalan, Luh Putu Suastini setata ijeg motrékin jagaté	Luh Putu Suastini
8	Utamane katuju tersebut majeng desainer Dayu Karang sane ngadungan karya busana antuk pepayasan jinah bolong	Dayu Karang

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Tabel 3. System Evaluation

Accuracy	Precision	Recall	F-measure
67.41%	83.42%	77.83%	80.53%

4. Conclusion

From the research conducted, NER to identify the names of characters in Balinese language documents using a rule-based method has an accuracy of 67.41%, precision of 83.42%, recall of 77.83%, and F-Score of 80.53%. The performance of NER, which is carried out based on the rules of the morphological structure and linguistic meaning is not too good because not all the names of the characters in the Balinese language document have the same morphological structure, and some of them doesn't have linguistic meaning. The lack of accuracy is also caused by the presence of typos and incorrect placement of punctuation marks in the document.

The suggestion in the future for this research is the need to add rules based on the features of other character feature such as POS Tagger, and Chunking to grouping the word.

References

[1] I. B. G. W. PUTRA, M. SUDARMA, and I. N. S. KUMARA, "Klasifikasi Teks Bahasa Bali dengan Metode Supervised Learning Naive Bayes Classifier," *Teknol. Elektro*, vol. 15, no. 2, pp. 81–86, 2016, [Online]. Available: https://ojs.unud.ac.id/index.php/JTE/article/view/ID21577.

[2] M. Y. S. Dirgantara, M. A. Fauzi, and R. S. Perdana, "Penerapan Named Entity Recognition Untuk Mengenali Fitur Produk Pada E-Commerce Menggunakan Rule Template Dan Hidden Markov Model," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 10, pp. 3912–3920, 2018.

[3] I. G. A. P. Arimbawa and N. A. S. ER, "Lemmatization in Balinese Language," *JELIKU - J. Elektron. Ilmu Komput. Udayana*, vol. 8, no. 3, pp. 235–242, 2020, [Online]. Available: https://ojs.unud.ac.id/index.php/JLK/article/view/51892.

[4] I. M. Wahyu and G. Negara, "Basic Word Extraction Algorithm Based on Morphological Rules for Balinese Basic Word Extraction Algorithm Based on Morphological Rules for Balinese Texts," no. May, 2020.

[5] Y. Kaur and E. R. Kaur, "Named Entity Recognition (NER) System for Hindi Language Using Combination of Rule Based Approach and List Look Up Approach," *ternational J. Sci. Res. Manag.*, vol. 3, no. 3, pp. 2300–2307, 2015.

[6] M. Hjouj, A. Alarabeyyat, and I. Olab, "Rule Based Approach for Arabic Part of Speech Tagging and Name Entity Recognition," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 6, pp. 331–335, 2016, doi: 10.14569/ijacsa.2016.070642.

[7] M. D. Drovo, M. Chowdhury, S. I. Uday, and A. K. Das, "Named Entity Recognition in Bengali Text Using Merged Hidden Markov Model and Rule Base Approach," *2019 7th Int. Conf. Smart Comput. Commun. ICSCC 2019*, no. September, pp. 1–5, 2019, doi: 10.1109/ICSCC.2019.8843661.

[8] R. E. Salah and L. Q. B. Zakaria, "Arabic rule-based named entity recognition systems: Progress and challenges," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 7, no. 3, pp. 815–821, 2017, doi: 10.18517/ijaseit.7.3.1811.

[9] M. Sailler and S. Markantonatou, *Multiword expressions*. 2018.

[10] I. G. W. S. Bandana, "Sistem Nama Orang Bali: Kajian Struktur dan Makna," *Aksara*, vol. 27, no. 1, pp. 1–11, 2015, doi: 10.29255/aksara.v27i1.166.1-11.

[11] I. G. B. W. B. Temaja, "Sistem Penamaan Orang Bali," *Humanika*, vol. 24, no. 2, pp. 60–72, 2018, doi: 10.14710/humanika.v24i2.17284.

This page is intentionally left blank