Jurnal Elektronik Ilmu Komputer Udayana
Volume 9 No. 4, May 2021

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

Rule-based Named Entity Recognition (NER) to Determine Time Expression for Balinese Text Document

Ni Made Sinta Wahyuni[a1], Ngurah Agus Sanjaya ER[a2]

[a]Informatic Departement, Udayana University
Bali, Indonesia
[1]madesintawahyuni@gmail.com
[2]agus_sanjaya@unud.ac.id

Abstract

Named Entity Recognition (NER) is a process to identify words or phrases as a named entity, such as a person, location, time expression, or organization. In this research, we are interested in developing a NER which able to identify the time expression entity in Balinese text documents. The time expression entity becomes an important component in the text because it is usually followed by important facts and information. NER was built using a rules-based approach. The rules are built based on direct observation of documents and pay attention to the morphological and contextual structures. Based on the experiments conducted, the average results of the precision, recall, and f-measure values were 0.85, 0.87, and 0.85.

Keywords: *NER, Balinese text documents, time expression, rule-based, rules*

1. Introduction

Indonesia is a rich country with cultural diversity. Each region in Indonesia has its cultural characteristics. One form of Indonesian cultural diversity is regional languages. In 2017, Indonesia has 652 regional languages [1]. Balinese is one of the regional languages of Indonesia. The use of Balinese is still maintained and actively used in various aspects of life such as communication, customs, religion, education, and even the media. Balinese is categorized as a major regional language because it has a large number of speakers, a written system, and a literary tradition [2].

The rapid development of technology and information is in line with the amount of availability of Balinese text documents in digital format. Besides, the need for fast and accurate information also increases. However, a lot of important information is still scattered in narrative documents, so it can take a long time to find it. One way to find this important information is by using NER.

Named Entity Recognition (NER) is a derivative of information extraction (IE). Information extraction (IE) is the process of finding information from a document or natural language with the results in the form of structured information in a certain format. The purpose of NER is to recognize and identify named entities and classify them into predetermined categories [3]. The named entity can be a person, location, organization, time expression, and others. NER also has an important role in various natural language processing (NLP) applications, such as text understanding, information retrieval, automatic text summarization, question answering, machine translation, and knowledge base construction [4].

NER has three general approaches to use, there are the rule-based method, the learning-based method, or the hybrid approach [5]. The rule-based approach relies on the rules and patterns of named entities contained in sentences and is defined manually using regular expressions based on linguistic knowledge and entity characteristics [6], [7], [8]. Linguistic knowledge can include grammar, contextual, lexicon, and algorithms to determine each of the operations involved [6].

Several studies related to NER have been carried out using various methods in various languages. However, research related to NER for Balinese has never been done before. In

research [9], the authors constructed NER by using a set of rules by combining contextual, morphological, and part-of-speech knowledge. NER is built for Indonesian. The evaluation is done by calculating the precision, recall, and f-measure values with the results of 63.63% recall and 71.84% precision. These results outperform those obtained using the maximum entropy method and associated rule-based method. Research [6], built a NER that able to identify entities in the biomedical domain using rule-based and Naïve Bayes classifiers. 18 rules are built based on observations on training data. The results obtained in this study are the highest average value of precision, recall, and f-measure is 0.85 with a micro average.

In a document, the existence of a time expression is usually followed by important facts or information in the vicinity [10]. Expressions of time have several types, there are range, sequence, duration, function, other types, and omitted phenomena [11]. A range is a period between two times or dates. The sequence is a sequence of several time ranges. Duration is a period. Functions include implied semantic entities. Another type is complementary in representing time expressions such as numbers which indicate numbers without any inherent temporal meaning. Omitted phenomena focus on events or states relative to expressions of time.

In this research, a NER will be developed to identify and recognized time expression entities for Balinese text documents using a rule-based approach. The rules are made based on direct observation of documents that have a certain pattern by observing the morphological and contextual structures. Evaluation is done by calculating the average of the precision, recall, and f-measure values of the resulting NER.

2. Research Methods

The process stages of developing a rule-based NER for Balinese text include data preprocessing, identifying the time expression entity with the NER, then generate output as time expression entity according to the established rules. The preprocessing stage includes case folding, data cleaning, and tokenization. The results from the preprocessing stage will be used as input for the identification process of the time expression entity with a rule-based NER for the Balinese text document. The following is a flow chart of the NER development process which can be seen in figure 1.

2.1. Types of Time Expression

In this research, NER will be built to identify and recognized time expression entities. Time expressions have some type, include range, sequence, duration, function, other types, and omitted phenomena [11].

a. Range
   The range is the period between two times or dates that represent or are called intervals. An instance of a range is 17 Mei - 20Agustus 2020.
b. Sequence
   The sequence is a sequence of several time ranges. The instance is rahina Redite mangkin, rahina Redite sane jagi rauh.
c. Duration
   Duration is a period of time. The instance is tigang rahina, telung tiban, aminggu.
d. Function
   This type includes implied semantic entities. An instance of the function is 2 dina sane lintang
e. Other types
   This type is complementary in representing time expressions such as numbers without any inherent temporal meaning. An instance is 2012.
f. Omitted phenomena
   This type focuses on events or states relative to a timestamp expression. An instance is warsa 90-an.

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

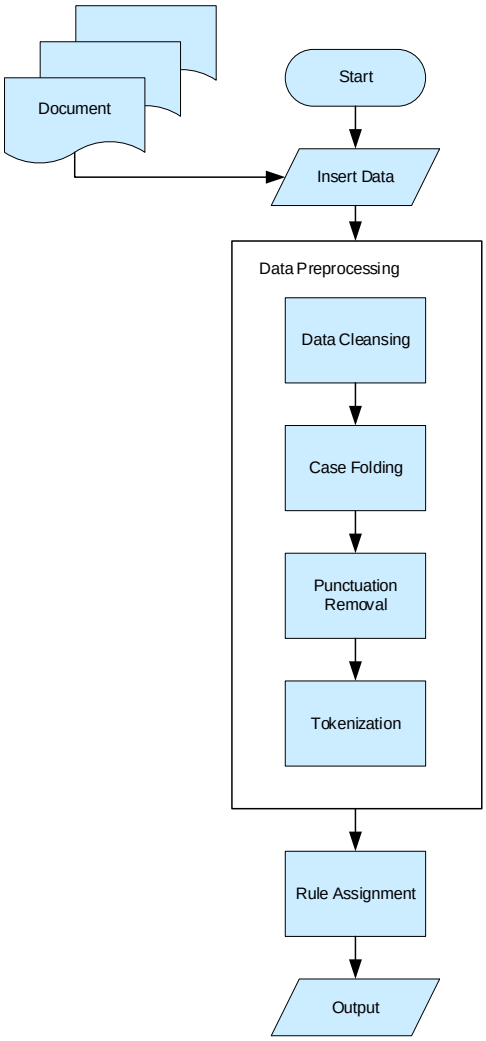

Figure 1. System Flow Chart

2.2. Data Preprocessing

Data preprocessing is a stage for preparing text into data that is ready to be processed for the next stage. In this study, there are four preprocessing stages, namely data cleansing, case folding, punctuation removal, and tokenization. Data cleansing is the stage of detecting, eliminating errors, and inconsistencies in data to improve data quality, such as changing the character é to e. Case folding is the stage of changing all letters in the text document to lowercase. Punctuation removal is the stage of removing punctuation marks for the example ',! \ "# $% & () * +.:; <=>? @ [\] ^ _` {|} ~ \ that is contained in the text. Tokenization is the breakdown step string into the smallest unit called a token. In this research, what is meant by the token is the word.

2.3. *Rule-based*

The results obtained from preprocessing data, then used for the next stage, namely NER by using a rule-based approach. At this stage, rule-based checks the suitability of each token toward each rule that has been made. If any of the rules matches, the token is identified as a named entity i.e. a time expression. However, if no rules match, it will proceed to the next token.

The rules for NER are made based on direct observation of documents that have a certain pattern. Besides, rules are made with due regard to the morphological and contextual structures. The morphological structure is based on the structure of the word, as in table 1. The contextual structure observing to the word to be used as initially annotated text, as in table 2. For example,

when you find the word "pinanggal" it is usually followed by a time expression such as 17/8 so that the next word is a time expression.

Tabel 1. List of Morphological Features

Feature	Explanation	Example
numStr	number in word	kalih, telu, dasa, limolas
digitSlash	number with slash	12/09/2019
digitHyphen	number with hyphen	13-8-2020
digit	all number	2020

Tabel 2. List of Contextual Features

Feature	Explanation	Example
day	list of days	redite, soma, anggara
month	list of months	januari, februari, maret
time_prefix	time prefix	pinanggal, duk, periode
time_sufix	time suffix	lintang, kaon, rauh
time_range	range of time	nyantos
duration	duration	detik, jam, menit
durationStr	duration in word	abulan, atiban, awai
notDatePrefix	not in list of date prefix	no, nomer, pergub, perda

The following is an example of a rule built to identify time marking entities in Balinese text documents.

IF token[index] in time_prefix and (token[index+1].isnumeric() == True or token[index+1] in numStr) and token[index+2] in duration
THEN time_entity = token[index+1] + token[index+2]

If given the input sentence " Parikrama sane kamargiang antuk akeh dudonan puniki kakawitin saking workshop nyantos kompetisi pantaraning mahasisya suennyane 3 dina ring Bali puniki.". Based on the rule above, the NER will identify the “3 dina” as a time expression entity.

2.4. *System Evaluation*

In this research, to determine the performance produced by NER in identifying and recognized time expression entities, an evaluation measurement technique be required. Evaluation is carried out to obtain the average precision, recall, and f-measure values. Precision is a comparison of the amount of relevant information obtained by the system with the total amount of information retrieved by the system, whether relevant or not. Recall is a comparison of the amount of relevant information obtained by the system with the amount of all relevant information contained in the collection of information either that has been or has not been retrieved by the system. F-measure is the relationship between precision and recall that provides system accuracy. The following is an equation for calculating precision (1), recall (2), and f-measure (3).

$$\text{Precision} = \frac{\text{number of correct responses}}{\text{number of responses}} \quad (1)$$

$$\text{Recall} = \frac{\text{number of correct responses}}{\text{number correct in key}} \quad (2)$$

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

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

3. Result and Discussion

To determine the performance of the NER system, testing or evaluation is carried out to obtain the average values of precision, recall, and f-measure. The type of data used in this research is secondary data. Secondary data is data that is already available before the researcher starts the research and that is related to the research to be carried out. Testing was carried out using 50 Balinese text documents consisting of news and short stories with file format * .txt. All documents are used as test data. The following are the example of the result obtained by the system that is shown in table 3 and the precision, recall, and f-measure result of some document can be seen in table 4.

Tabel 3. Example of Testing Data

No	Input Sentence	Entity Result
1	Mirib sing sida naanang sakit ati, atiban ané liwat, méménné nuturin bapanné apang suud ja ngantén-palas, mamitra miwah mamotoh.	atiban liwat
2	Ring Sensus Penduduk (SP) Online sane memargi saking pinanggal 15 Februari kantos 29 Mei 2020, kasurat 1.571.119 krama utawi 35,59% saking 4.414.431 krama Provinsi Bali sampun ngamiletin SP Online.	15 Februari-29 Mei 2020
3	Ring rahina Wrespati (9/7/2020), Perhimpunan Penggemar Mobil Kuno Indonesia (PPMKI) Bali pacang ngelaksanayang tour sane kakawitin ring ajeng Kantor Gubernur Bali.	Wrespati 9/7/2020
4	Ring program sane megalah 60 menit punika taler ngerauhang juru raos praktisi tanaman toga lan pangan Yuliani Djajanegara, miwah kadagingin baga metaken nyawis interaktif indik orti ketahanan pangan kulawarga.	60 menit
5	Pikobet kulawarga puniki kantun memargi santukan Sutarini sane kaandelan kulawarga keni stroke nem sasih lintang.	nem sasih lintang
6	Telung tiban Madé Loka ngarap gumi, tusing taén mupu.	telung tiban
7	Gubernur Bali pinanggal 20 September 2020 Wayan Koster nyihnayang pamujinnyane duaning ring sajeroning acara pamungkah seminar nasional puniki para pamilet nganggen destar khas Bali.	20 September 2020
8	Gubernur Bali Wayan Koster ngamedalan Peraturan Gubernur nomor 26 warsa 2020 indik Sistem Pengamanan Lingkungan Terpadu Berbasis Desa Adat (SIPANDU BERADAT), Sukra (10 Juli 2020).	Sukra 10 Juli 2020
9	Parikrama sane kamargiang antuk akeh dudonan puniki kakawitin saking workshop nyantos kompetisi pantaraning mahasisya suennyane 3 dina ring Bali puniki, baos Prof Santoso, kamargiang olih pihaknyane kasarengin olih Fakultas Teknologi Pertanian Universitas Udayana (Unud).	3 dina
10	Manut dane kuartal kaping tiga warsa 2020 puniki dados kunci ngewalian ekonomi Indonesia kantos nenten tedun me jurang resesi.	2020

Tabel 4. Evaluation Result of Some Testing Documents

Document	Precision	Recall	F-measure
Document 1.txt	0.5	0.33	0.4
Document 2.txt	1.0	1.0	1.0
Document 3.txt	0.5	0.5	0.5
Document 4.txt	1.0	1.0	1.0
Document 5.txt	1.0	1.0	1.0
Document 6.txt	1.0	1.0	1.0
Document 7.txt	1.0	1.0	1.0
Document 8.txt	0.66	0.75	0.70
Document 9.txt	0.71	0.62	0.66
Document 10.txt	1.0	1.0	1.0
Document 40.txt	1.0	1.0	1.0
Document 41.txt	1.0	1.0	1.0
Document 42.txt	1.0	1.0	1.0
Document 43.txt	1.0	1.0	1.0
Document 44.txt	1.0	1.0	1.0
Document 45.txt	1.0	1.0	1.0
Document 46.txt	1.0	1.0	1.0
Document 47.txt	1.0	1.0	1.0
Document 48.txt	1.0	1.0	1.0
Document 49.txt	1.0	1.0	1.0
Document 50.txt	1.0	1.0	1.0
Average	0.85	0.87	0.85

Based on the tests, we get the average value of precision is 0.85, recall is 0.87, and f-measure is 0.85. In the table, it is shown that several documents have less than optimal values of precision, recall, and f-measure. This is affected by several factors, such as errors in writing time expressions such as days, the existence of an entity that has not been recognized by NER as a time expression entity because the rules have not been defined and some words or phrases which are not classified as time expression entities but are recognized by the system.

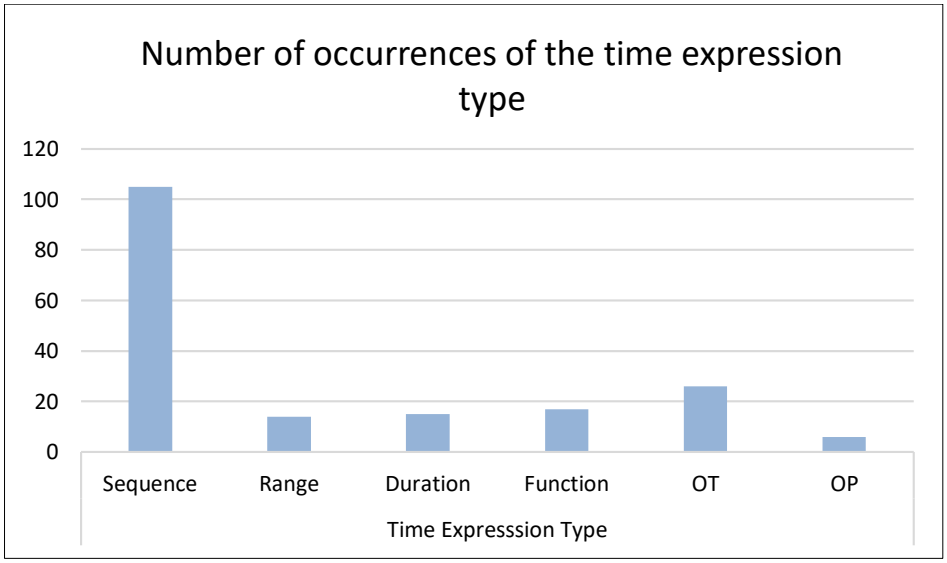

Figure 2. Number of Occurrences of Time Expression Type

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

Besides that, we can see the number of occurrences of the most recognizable time expression type in Figure 2. A sequence is a type of time expression that is most easily recognized and retrieved by the system. This is because most of the testing data are news articles, which contain time expressions such as sequences to explain the sequence of events. In this system, the types of time expressions that can decrease accuracy are duration and function. This is because there are various styles of writing, such as 20-30 January 2020, on the other hand, there are those who write 20 January-30 January 2020 or 20 January nyantos 30 January 2020.

4. Conclusion

We have built Named Entity Recognition (NER) using a rule-based approach. NER recognizes time expression entities in Balinese text documents. Rules are built based on direct observation of documents and use the morphological and contextual structures. Based on the research conducted, it is concluded that the rule-based approach can be used in NER on Balinese text documents and able to provide results based on the compatibility with the rules. The results showed that the average precision, recall, and f-measure values are 0.85, 0.87, and 0.85. A sequence is the most recognizable time expression type because most of the testing data are news articles, which contain time expressions to explain the sequence of events. Hopefully in the future, the development of NER to identify time expression entities in Balinese documents will continue to be carried out with other approaches so that it is expected to be able to recognize all types of time expressions.

References

[1] "Kementerian Pendidikan dan Kebudayaan » Republik Indonesia." https://www.kemdikbud.go.id/main/blog/2018/07/badan-bahasa-petakan-652-bahasa-daerah-di-indonesia (accessed Sep. 21, 2020).

[2] W. A. S. Gitananda, "DALAM BAHASA BALI Analisis Morfofonemik dan Fungsi Sintaksis," pp. 1–7, 2002.

[3] Y. Kurniawati and P. P. Adikara, "Implementasi Named Entity Recognition Pada Factoid Question Answering System Untuk Cerita Rakyat Indonesia," *J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya*, vol. 2, no. 9, pp. 3142–3149, 2018, [Online]. Available: http://puslit2.petra.ac.id/ejournal/index.php/inf/article/view/1647.

[4] J. Li, A. Sun, J. Han, and C. Li, "A Survey on Deep Learning for Named Entity Recognition," *IEEE Trans. Knowl. Data Eng.*, vol. XX, no. XX, pp. 1–1, 2020, doi: 10.1109/tkde.2020.2981314.

[5] K. Adnan and R. Akbar, *An analytical study of information extraction from unstructured and multidimensional big data*, vol. 6, no. 1. Springer International Publishing, 2019.

[6] D. W. Wulandari, P. P. Adikara, and S. Adinugroho, "Named Entity Recognition (NER) Pada Dokumen Biologi Menggunakan Rule Based dan Naïve Bayes Classifier," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 11, pp. 4555–4563, 2018, [Online]. Available: http://j-ptiik.ub.ac.id.

[7] H. J. Song, B. C. Jo, C. Y. Park, J. D. Kim, and Y. S. Kim, "Comparison of named entity recognition methodologies in biomedical documents," *Biomed. Eng. Online*, vol. 17, no. s2, pp. 1–14, 2018, doi: 10.1186/s12938-018-0573-6.

[8] T. Eftimov, B. K. Seljak, and P. Korošec, *A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations*, vol. 12, no. 6. 2017.

[9] I. Budi, S. Bressan, G. Wahyudi, Z. A. Hasibuan, and B. A. A. Nazief, "Named Entity Recognition for the Indonesian language: Combining contextual, morphological and part-

of-speech features into a knowledge engineering approach," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 3735 LNAI, pp. 57–69, 2005, doi: 10.1007/11563983_7.

[10] J. M. Lim, I. S. Kang, J. H. J. Bae, and J. H. Lee, "Sentence extraction using time features in multi-document summarization," *Lect. Notes Comput. Sci.*, vol. 3411, pp. 82–93, 2005, doi: 10.1007/978-3-540-31871-2_8.

[11] K. Lee, Y. Artzi, J. Dodge, and L. Zettlemoyer, "Context-dependent Semantic Parsing for Time Expressions," pp. 1437–1447, 2014.