Part of Speech Tagging for Javanese Ngoko Language with Hidden Markov Model

Ryan Armiditya Pratama, Arie Ardiyanti Suryani, Warih Maharani
School of Computing, Telkom University
Jl. Telekomunikasi Terusan Buah Batu, Bandung
Email: ryanarmiditya@student.telkomuniversity.ac.id, ardiyanti@telkomuniversity.ac.id, wmaharani@telkomuniversity.ac.id

***Abstract*—Indonesia has many tribes; one of the largest number of tribes is Javanese with the Javanese languages spoken in Central Java, Jogjakarta, and East Java. Its grammatical structure is similar to the Indonesian language. Part of Speech Tagging labels words in a sentence with their word class based on the word function in the sentence. Some POS tagging research have been conducted for Indonesian, but its rare for Javanese. This study aims to automatically PoS tag words in a sentence written in Javanese Ngoko language, using a dataset crawling from online news. This study used Hidden Markov Model (HMM) and got an accuracy of 92.6 %.**

Key Words: POS Tagging, Javanese Ngoko, Labeling, Hidden Markov Model.

I. Introduction

POS tagger is a system that sets word class labels for each word in a sentence automatically [1]. POS tagger is part of Natural Language Processing (NLP) and benefits to Multi-Word Expression (MWE), Word Sense Disambiguation (WSD) and Statistical Machine Translation.

Javanese is a spoken and writen language largely used by people living in the Java island [2]. It has levels of language based on the politeness that is Ngoko for daily conversation and Krama is used to communicate with older or higher levels people [3]. They are still be used currently.

There are various approaches of labeling words with probability based and ruled based. Probability based is a bottom-up approach that uses corpus as training data, and then the system will set the word with a tag whose probability is the highest. In this approach, the training corpus should be labeled first. Rule based is a top-down method that consults linguists regarding the rules [4].

The study of POS Tag for the Indonesian language has been carried out with variety methods and obtained good accuracy, such as HMM Based Part of Speech Tagging for Indonesian with the best accuracy of 96.5% [5] and POS Tagging Indonesian with HMM and Rule Based with the best accuracy of 92.2% [6].

Research for Javanese language PoS tags is very rare despite Javanese Ngoko is widely used in writing news or articles. One study done by [2] was POS Tag for the Krama Javanese Language with Rule Based and Maximum Entropy method and got 97.67% accuracy. Different from those previous research, in this study used Ngoko Javanese language dataset with Hidden Markov Model (HMM) method. To the best of our knowledge, this study is the first statistically POS tagger with HMM on the Ngoko Javanese dataset. POS tagging is a sequential task; the tag in the previous word affects the word that to be tagged. Therefore, this research used HMM that is a good probabilistic method and is suitable to be used for POS tag. It uses temporal sequential data logic circuit whose output depends on the previous results [5]. The work on Ngoko Javanese because it is used in daily communication among Javanese people. Moreover, the development of the Javanese language in the NLP technology is rapid as it available in Google Translate.

II. Related Works

The research about POS Tag for the Indonesian language with the Viterbi algorithm resulted good accuracy [1]. The size of used dataset was 16.291 words. Testing is done with two scenarios, the first was to use 16.290 words in the dataset and the second was by adding a word ‘zz’ to the dictionary for words that are not in the dataset. Every method was evaluated with precision and recall. The highest precision obtained was to 93.6 % while recall was 94.58 %.

The research about POS Tag for the Javanese language has been done by [2] by combining the Maximum Entropy and Rule Based method. The size of the dataset was 2,380 words as the first training data and 8488 words as the second training data. The highest accuracy obtained was 97.67 % using in training data and 95.75 % accuracy using different test data. In this study Rule Based helped the resulted accuracy because it handled testing data that is not known in the training data.

On the other hand, in the research [6] labeling word classes for Indonesian language texts using the Hidden Markov Model and Rule-Based methods has high accuracy results, with the highest 100 % for the text that is in the corpus. When compared with POS tagging that uses only HMM, the merging of 2 methods in this study gives better results, the highest accuracy obtained is 100 % for the same text as the corpus while POS only with HMM has the highest accuracy of 99.29 %. The system in this study is able to process word input in the correct writing sequence and this is the difference or improvement of similar research that has been done before. This research

requires a large corpus in order to provide more precise labeling.

The research by [7] with a combination of Unigram, HMM, and Brill Tagger methods for POS Tag Indonesian Language. Used data of 30,000 sentences with 700,000 words. The best accuracy obtained using the Unigram, HMM, and Brill tagger methods was 88.37 %, 62.69 %, and 76.78 % respectively. After get these results it can be concluded that the unigram method produces very high accuracy compared to the HMM and brill tagger methods. HMM method is very dependent on the large number of words in the training data to produce high accuracy.

Research [8] conducted POS tag for Indonesian language by combining six methods, namely Unigram, HMM, TnT, Brill Tagger, Naive Bayes and Maximum Entropy. In this study the best method was Maximum Entropy with the best accuracy of 90.6 % then. HMM with the best accuracy of 90.39 % and Unigram 88.97 %.

III. METHODS

The system built in this research is POS tagging for the Javanese language and produces tagging for each input word by using the Hidden Markov Model method and applying the Viterbi algorithm for the testing process. An overview of the process PoS tag in the system can be seen in Figure 1

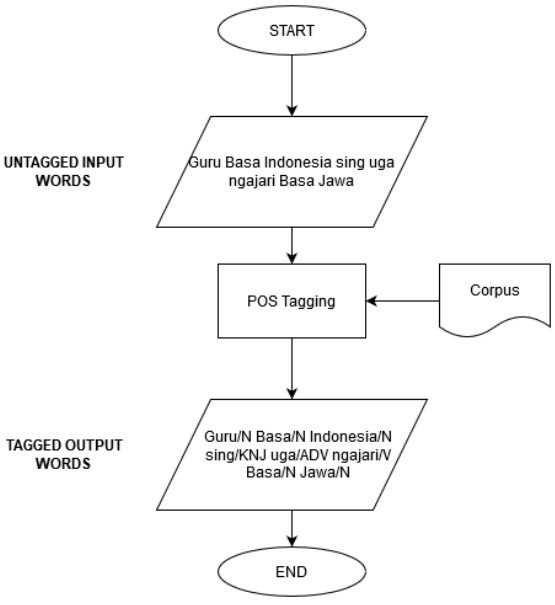

Fig 1. The process of POS Tagging

The untagged input word is an input word that does not have a label and will be labeled by the system automatically. Then input word-processed by the system to be labeled based on the corpus file that has been trained to the system before. Corpus file contains a collection of sentences in Javanese that have been collected and labeled the word classes for training the system for tagging.

The following is an example of part of the data that used in the corpus.

1) Loro-lorone/PR padudon/V nganti/ADV tiwas/ADJ ,/, sampyuh/V ,/, gara-gara/N salah/ADJ pangerten/N marang/KNJ dhawuhe/N Ajisaka/N ./. Persentase/N nom-nomane/N sing/KNJ gelem/V ndeleng/V kethoprak/N antarane/N mung/ADJ 30/NUM %/SYM ./. Dene/KNJ ,/, yen/KNJ babagan/KNJ paraga/N kethoprak/N ,/, kandhane/V Ali/N ,/, akeh/ADJ nom-noman/N sing/KNJ gelem/V melu/V pentas/N ./.

2) Ing/PRP masyarakat/N Jawa/N ,/, dongeng/N sing/KNJ kerep/ADJ dirungokake/V marang/KNJ bocah-bocah/N kayata/KNJ Timun/N Mas/N ,/, Kancil/N Nyolong/V Timun/N ,/, lan/KNJ Malin/N Kundang/N ./. Nanging/KNJ ,/, kerep/ADJ diewangi/V mahasiswa/N saka/PRP Institut/N Seni/N Indonesia/N (/kurung buka ISI/N)/kurung tutup Solo/N lan/KNJ liyane/ADJ nalika/KNJ pentas/N ./.

3) Dosen/N Program/N Studi/N Basa/N Jawa/N Fakultas/N Keguruan/N dan/KNJ Ilmu/N Pendidikan/N (/kurung buka FKIP/N)/kurung tutup Universitas/N Sebelas/N Maret/N (/kurung buka UNS/N)/kurung tutup Solo/N ,/, Rahmat/N ,/, Selasa/N ,/, ngandharake/V menawa/KNJ nom-noman/N sing/KNJ seneng/ADJ sinau/V aksara/N Jawa/N ora/ADV pati/N akeh/ADJ ./.

4) Kamangka/ADV yen/KNJ mulang/V ing/PRP sekolah/N ,/, paling/KNJ ya/EM mung/ADJ maca/V teks/N aksara/N Jawa/N ./. Kudu/ADV ngupaya/N aja/ADV nganti/ADV ana/V basa/N campuran/N ./. Ya/EM minangka/PRP kupiya/N nguri-uri/V basa/N Jawa/N ,/, tuture/V Ali/N ./. Saben/KNJ pepanthan/N pancen/ADV duwe/V cara/N sing/KNJ beda-beda/ADJ ngleluri/V kethoprak/N ./. Kethoprak/N Ngampung/N kerep/ADJ ngadani/V pentas/N ing/PRP desa-desa/N sing/KNJ adoh/ADJ saka/PRP kutha/N Solo/N kayata/KNJ Wonogiri/N ,/, Klaten/N ,/, lan/KNJ Sragen/N ./.

5) Kanggo/PRP sangu/N sapa/PR ngerti/V mengko/N yen/KNJ wis/ADV lulus/V ora/ADV mung/ADJ dadi/V guru/N ,/, nanging/KNJ kerja/V ing/PRP kantor/N arsip/N naskah/N ,/, perpustakaan/N ,/, lan/KNJ sakpiturute/PR ,/, kandhane/V ./. Nganti/ADV lulus/V mung/ADJ mbabar/V enem/NUM jinis/N tembang/N kamangka/ADV Macapat/N ana/V 11/NUM ./.

Figure 2 is an overview of the process POS Tagging in the system with the Hidden Markov Model method. The corpus file going through the tokenization process to separate the sentence into the units of words so that it will be easier in the POS Tag. The tokenization process detects a space separator to be separated into words. The example of result tokenization process can be seen in Table I

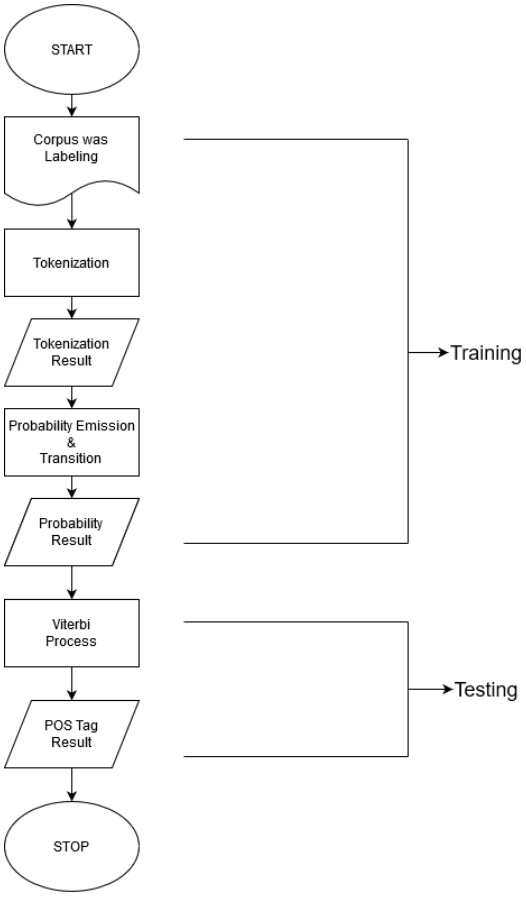

Fig 2. The process of POS Tagging with HMM

Table I. RESULT OF TOKENIZATION PROCESS

Sentence before Tokenization	Sentence after Tokenization
Saben pepanthan pancen duwe cara sing beda-beda ngleluri kethoprak .	'Saben', 'pepanthan', 'pancen', 'duwe', 'cara', 'sing', 'beda-beda', 'ngleluri', 'kethoprak', '.'
Dongeng sing diwaca marang bocah-bocah iku lumrahe ora mung kanggo rungon-rungon sakdurunge turu,nanging ya ngandhut piwulang sing wigati bab kauripan.	'Dongeng', 'sing', 'diwaca', 'marang','bocah-bocah', 'iku', 'lumrahe', 'ora', 'mung', 'kanggo', ',', 'rungon-rungon', 'sakdurunge', 'turu', 'nanging', 'ya', 'ngandhut', 'piwulang', 'sing', 'wigati', 'bab', 'kauripan', '.'
Ketua sing uga ngedekake Istana Dongeng Nusantara, Rebo (20/2/2019), ngandharake dongeng iku kudune ora mung diwacakke sakdurunge turu.	'Ketua', 'sing', 'uga', 'ngedekake', 'Istana', 'Dongeng', 'Nusantara', ',', 'Rebo', '(', '20', '/'. '2', '/', '2019', ')', ',' , 'ngandharake', 'dongeng', 'iku', 'kudune', 'ora', 'mung', 'diwacakke', 'sakdurunge', 'turu', '.'
Nasyir ngandharake dongeng uga kerep didadekake sarana kanggo ngilangi trauma wong-wong sing dadi korban banjir, longsor, lan liyane.	'Nasyir', 'ngandharake', 'dongeng', 'uga', 'kerep', 'didadekake', 'sarana', 'kanggo', 'ngilangi', 'trauma', 'wong-wong', 'sing', 'dadi', 'korban', 'banjir', ',', 'longsor', ',', 'lan', 'liyane', '.'
Sing luwih penting,dongeng iku ora mung kanggo bocah cilik, nanging uga bisa ditrepake marang mudha-mudhi remaja.	'Sing', 'luwih', 'penting', ',', 'dongeng', 'iku', 'ora', 'mung', 'kanggo', 'bocah', 'cilik', ',', 'nanging', 'uga', 'bisa', 'ditrepake', 'marang', 'mudha-mudhi', 'remaja', '.'

A. Hidden Markov Model (HMM)

Hidden Markov Model is a statistical model that uses probabilities to get the best tag for each word. Hidden Markov Model has two important entities, namely observed states (observable) and hidden states (not observable) In this case, the hidden state is the order of the word class and the word order as an observed state [6]. The HMM model have three processes, namely initialization, transition, and emissions.

1) Initialization is the process of getting the number of labels from each word contained in the training data. In initialization, the input is the labeled words and their label. Labeled words are collections of sentences that have been labeled and the labels are the types of labels contained in the data set. The results of the initialization are given the symbol π.

2) Transition is a process of looking for a word label after the current word label. Transitions are obtained by entering labeled words and word labels into the system. The results of the transition are symbolized as a.

3) Emission is the process of finding the number of words from each label contained in the training data. In obtaining emissions, it is necessary to include labeled words and word labels. The word dictionary is a collection of words contained in the data set. The results from emissions are symbolized as b.

B. Probability Emission and Transition

Probability emission is the chance of a word appears if it is given a specific word class. An example of emissions probability P (gawe | V) is the chance of the word "gawe" appearing if it has the Verb word class. Eq. 1 the calculation of probability emission [9] :

$$P(w_i|t_i) = \frac{C(t_i, w_i)}{C(t_i)} \quad (1)$$

Probability transition is the chance of the emergence of a class of words where a particular word class has previously appeared. Example of transition probability P (KNJ | V) which shows the probability of the appearance of the word class Conjunctions after Verb. Eq. 2 for the calculation of probability transition [9] :

$$P(t_i|t_i - 1) = \frac{C(t_i - 1, t_i)}{C(t_i - 1)} \quad (2)$$

C. Viterbi Algorithm

Viterbi algorithm is a dynamic programming algorithm that has been widely used in the application of Natural Language Processing [10]. Viterbi runs on the testing process by using the calculation results from emission and transition to get the best path order called Viterbi path.

In determining the best path, Viterbi algorithm uses 2, forward and backward steps. The forward step process calculate the word class of an input word by using the lowest negative log probability value [11]. The forward step process can be seen in Figure 3 [11].

E-ISSN:2541-0806
P-ISSN:2540-8895

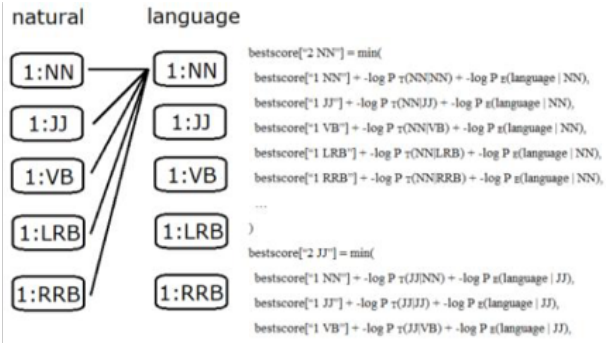

Fig 3. The process of calculating the lowest negative log probability

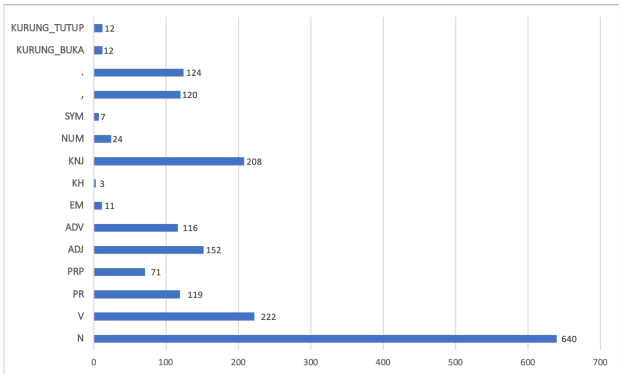

Fig 4. Number of Tagset

The forward step process continues for each word until the end of the sentence and the lowest value used as a word class. The backward step is running if the forward step has been completed. The lowest negative log probability results are then saved in a list consists of an ordered word class to label the input word [11].

Forward and backward algorithms are derivatives of forward-backward algorithms that can be used to handle OOV problems. Both of these algorithms have the same characteristics and are very possible to be used in OOV handling in POS Tagging HMM.

D. Out Of Vocabulary (OOV)

The problem with POS tag is the existence of OOV due to the limited size of corpus in the train data. It affects the value of the transition probability and the word will not get the appropriate label. This research solve the OOV by using Trigram state. OOV causes the transition probability value to be small because the POS label is not present on the training data. To solve this problem, the transition probability recalculation is performed using the formula in Eq.3 [12]:

$$P(t3|t2,t1) = \frac{Count(t1,t2,t3)}{Count(t1,t2)} \quad (3)$$

IV. Experiment Setup

The dataset in this study collected from online news solopos.com/jagad-jawa containing news in the Javanese Ngoko language. After the data is collected, we create a label with the tool from datasaur.ai/beta and labeled manually. Datasaur.ai/beta is an intelligent tool to help users label data so they can work more productively and efficiently. The system works using AI-based modeling and is supported by Natural Language Processing (NLP), which proactively suggests labels. The data contains 1,770 words with 126 sentences. Figure 4 shows the total tagset used in this research.

In general, there are five tagset used namely Verb, Noun, Adjective, Adverb, and Conjunction. Based on the division of these 5 categories, it still has subcategories of each word class.

The verb is a class of words that indicate an action, event, or action. Verbs can be divided into a transitive verb and intransitive verb. Transitive Verbs require additional objects so that sentences become whole and meaningful. In other words, the object that follows the verbs gets the action from the verbs. Intransitive is a verb that does not need a direct object so that the sentence becomes whole and has meaning. Unlike other verbs, there is only one form for this verb. Modal Auxiliary Verbs are words placed before the main Verb to modify the meaning of the main verb.Its function is to express willingness or ability, necessity, and possibility (possibility).

A noun is a category that used to name people, objects, animals, places, and can be divided into a countable noun, uncountable noun, genitive common noun, and proper noun. These subcategories have different examples, such as countable nouns whose numbers can be counted, while uncountable are nouns whose numbers cannot be counted. Adjectives are words used to describe nouns or pronouns that can form people, places, animals, things, or objects.A proper noun is a specific classification of nouns such as names of places and people that use capital letters at the beginning of the letters. A Cardinal Number states an amount that can be calculated as one, two, three, million. Prepositions, conjunctions, and interjections are included in the subcategory of the function word.

Table II shows the examples of tagsets for Indonesian which contains 37 tags[13].

Table II. Tagset Indonesian [13]

No.	Tag	Description	Example
1.	(	Opening Parenthesis	({ [
2.	)	Closing Parenthesis	) }]

The tagset in this study follows references from the Buku Kosa Kata Bahasa Jawa that have 10 categories tagset for the Javanese language [14] with modifications or additions made such as opening parenthesis, closing parenthesis, and sentence terminator that previously not

3.	,	Comma	,
4.	.	Sentence Terminator	. ? !
5.	:	Colon or Ellipsis	: ;
6.	–	Dash	–
7.	“	Opening Quotation Mark	‘ “
8.	”	Closing Quotation Mark	’ ”
9.	$	Dollar	$
10.	Rp	Rupiah	Rp
11.	SYM	Symbols	$ % & @
12.	NNC	Countable Common Nouns	Buku, Rumah, Karyawan
13.	NNU	Uncountable Common Nouns	Air, Gula, nasi, Hujan
14.	NNG	Genitive Common Nouns	Idealnya,komposisinya, fungsinya,jayanya
15.	NNP	Proper Nouns	Jakarta, Australia, Soekarno-Hatta
16.	PRP	Personal Pronouns	Saya, aku, dia,kami
17.	PRN	Number Pronouns	Kedua-duanya, Ketiga-tiganya
18.	PRL	Locative Pronouns	Sini, Situ, Sana
19.	WP	WH-Pronouns	Apa, Siapa, Mengapa, Bagaimana
20.	VBT	Transitive Verbs	Berbicara, Mengangkat, Menyanyi
21.	VBI	Intransitive Verbs	Bermain, terdiam, berputar-putar
22.	MD	Modal or Auxilaries Verb	Sudah, boleh, harus, mesti
23.	JJ	Adjectives	Mahal, kaya, besar, malas
24.	CDP	Primary Cardinal Numerals	Satu, juta,milyar
25.	CDO	Ordinal Cardinal Numerals	Pertama, kedua,ketiga
26.	CDI	Irregular Cardinal Numerals	Beberapa,segala, semua
27.	CDC	Collective Cardinal Numerals	Bertiga, bertujuh, berempat
28.	NEG	Negations	Bukan, tidak,belum, jangan
29.	IN	Prepositions	Di, ke, dari,pada, dengan
30.	CC	Coordinate Conjunction	Dan, atau,tetapi
31.	SC	Subordinate Conjunction	Yang, Ketika, Setelah
32.	RB	Adverbs	Sekarang, nanti, sementara, sebab, sehingga
33.	UH	Interjections	Wah, aduh, astaga, oh
34.	DT	Determiners	Para, ini, masing masing, itu
35.	WDT	WH-Determiners	Apa, siapa, barangsiapa
36.	RP	Particles	Kan, kah, lah, pun
37.	FW	Foreign Word	All word except Bahasa

available in the reference. The addition of the opening parenthesis tagset is based on table II, so it can accommodate the presence of symbols and punctuations. Table III lists 19 Javanese language tagsets.

Table III. THE TAGSET OF JAVANESE.

No	Tag Javanese Language	Description	Example
1	N	Nouns	Bayu, Wedhus
2	V	Verb	Tindak,Tumbas
3	ADJ	Adjective	Apik,Ayu,Bagus
4	ADV	Adverb	Mangkih,kadung,bablas
5	KNJ	Conjunction	Lamun, wiwit
6	PRP	Preposition	Marang
7	KH	Pengkhusus	Banget
8	SO	Subordinator	Nalika
9	EM	Emotif	Eh, Aduh
10	PR	Pronouns	Niki,sampean,kuwi,kulo
11	SYM	Symbol	Rp $ % & @
12	({ [	Opening Parenthesis	({ [
13	) }]	Closing Parenthesis	) }]
14	,	Comma	,
15	‘ “	Quotation Mark	’ ”
16	. ? !	Sentence Terminator	. ? !
17	– -	Dash	– -
18	:	Colon	:
19	;	Semicolon	;

V. RESULTS AND DISCUSSION

The results of this study are a collection of words in Javanese Ngoko that were POS tagged automatically based on the calculation of emission and transition probabilities followed Eq. 4.

$$Accuracy = \frac{Correctly\ Tagged\ Words}{Total\ Words} * 100 \quad (4)$$

Table IV shows examples output of the scheme I. It can be seen that not all words in the input sentences could be recognized correctly because the system never saw words that we called as ‘unknown words’. The system tagged the unknown words with “,” label. Table V lists a collection of unknown words from each input sentence. Unfortunately, given the label “,” did not affect the accuracy of the system because the label only indicated that the system failed to label the words.

The unknown word does not have a POS tag label, so its accuracy is lower. Unknown words occur because of OOV and the system cannot provide matching labels. Unknown words appear because OOV and the system cannot provide matching labels. Another thing that affects this problem is due to the lack of sentences in the dataset, so the system has never been trained to handle the appearance of the word. To solve these problems, so the system recalculates the transition probability with the trigram state and then produces a label for the unknown word. In the table VI, you can see the labeling results for words that previously did not have word class labels.

Table IV. SYSTEM OUTPUTS

Input Words	Output POS Tag
Kitab kuna sing wujude wis rusak banget biyasane dislametake kanthi cara digitalisasi naskah. Carane yaiku naskah difoto siji-siji banjur dokumene disimpen.	[('Kitab', 'N'), ('kuna', 'ADJ'), ('sing', 'KNJ'), ('wujude', 'N'), ('wis', 'ADV'), ('rusak', 'V'), ('banget', 'N'), ('biyasane', 'N'), ('dislametake', 'N'), ('kanthi', 'KNJ'), ('cara', 'N'), ('digitalisasi', 'N'), ('naskah', 'N'), ('.', '.'), ('Carane', 'N'), ('yaiku', 'KNJ'), ('naskah', 'N'), ('difoto', 'N'), ('siji-siji', 'N'), ('banjur', 'ADV'), ('dokumene', 'V'), ('disimpen', 'N'), ('.', '.')]
Saben pepanthan pancen duwe cara sing beda-beda ngleluri kethoprak .	[('Saben', 'KNJ'), ('pepanthan', 'N'), ('pancen', 'ADV'), ('duwe', 'V'), ('cara', 'N'), ('sing', 'KNJ'), ('beda-beda', 'ADJ'), ('ngleluri', 'V'), ('kethoprak', 'N'), ('.', '.')]
Dhagan ngandharake dene ngajari macapat marang bocah-bocah pancen rada angel.	[('Dhagan', 'N'), ('ngandharake', 'V'), ('dene', 'KNJ'), ('ngajari', 'V'), ('macapat', 'N'), ('marang', 'KNJ'), ('bocah-bocah', 'N'), ('pancen', 'ADV'), ('rada', 'ADV'), ('angel', 'ADJ'), ('.', '.')]

Table V. UNKNOWN WORDS NGOKO JAVANESE LANGUAGE

Input Words	Unknown Words
Saben ana sambatan, utawa adicara kabudayan mesti disengkuyung bareng. Ora mbedak-mbedakake endi Jawa endi Thionghoa.	[('sambatan', ','), ('adicara', ','), ('disengkuyung', ','), ('bareng', ','), ('mbedak-mbedakake', ','), ('endi', ','), ('endi', ','), ('Thionghoa', ',')]
Mula saka kuwi dhalang sing bener ya kudu bisa mbabar telung prekara kasebut. Ora kena nglantur, kudu kebak pitutur.	[('dhalang', ','),('prekara', ','),('kasebut', ','),('kena', ','),('nglantur', ',')]
Ing jaman biyen Imam uga kerep nglakoni pasa telung dina sakdurunge mbabar wayang.	[('Imam', ','), ('nglakoni', ','), ('pasa', ',')]

Table VI. UNKNOWN WORDS TAGGED

Input Words	Unknown Words Tagged
Saben ana sambatan, utawa adicara kabudayan mesti disengkuyung bareng. Ora mbedak-mbedakake endi Jawa endi Thionghoa.	[('sambatan', 'N'), ('adicara', 'N'), ('disengkuyung', '.'), ('bareng', 'N'), ('mbedak-mbedakake', 'V'), ('endi', 'N'), ('endi', 'N'), ('Thionghoa', 'N')]
Mula saka kuwi dhalang sing bener ya kudu bisa mbabar telung prekara kasebut. Ora kena nglantur, kudu kebak pitutur.	[('dhalang', 'ADV'), ('prekara', 'N'), ('kasebut', 'ADJ'), ('kena', 'V'), ('nglantur', 'N')]
Ing jaman biyen Imam uga kerep nglakoni pasa telung dina sakdurunge mbabar wayang.	[('Imam', 'N'), ('nglakoni', 'V'), ('pasa', 'N')]

Table VII contains the accuracy obtained from the results of the experiment, where the input words were out of the words in the training data.

Table VII is the results of the testing scheme I, carried out by inserting a different sentence from the sentence in the training data. The experiment was carried out ten times, where each experiment used a different sentences. In the Table VII, it can be seen that the highest accuracy obtained is 92.6 % in the first trial. Because the input word for this experiment is different from the data train, so there are a lot of unknown word, and the accuracy in this trial has been solved the OOV problems with trigram state. Input words in this experiment is not entirely different from the words in the training data, but there are some words found in the training data such as

Table VII. THE ACCURACY OF TESTING DATA DOES NOT APPEARS IN TRAINING DATA.

No	Total Words	Correct Words	Incorrect Words	Accuracy %
1.	27	25	2	92,6
2.	18	16	2	88,9
3.	36	32	4	88,9
4.	42	38	4	90,5
5.	41	36	5	87,8
6.	32	25	7	78,1
7.	45	39	6	86,7
8.	54	47	7	87,0
9.	26	18	8	69,2
10.	33	25	8	75,8
Total				845,5
Average				84,5

conjunction words. The accuracy of this experiment is good but it still has errors in tagging the words. The most commonly found word class tag errors in Noun are due to personal pronouns, names of people, and place names.

Table VIII contains the accuracy obtained by entering words into the system from taking some of the sentences in the training data to be tested. The experiment by entering the same words as the training data is to find out how well the system is tagging if the sentences have been previously trained in the system. The test was carried out ten times and the highest accuracy was 96.2 % in the third trial. Although the testing scheme II uses the same sentences as the data train, the obtained accuracy has not yet reached 98 % because the accuracy of the model was set at around 91 %.

Table VIII. ACCURACY OF TESTING DATA SAME AS DATA TRAIN

No	Total Words	Correct Words	Incorrect Words	Accuracy %
1	43	40	3	93
2	17	16	1	94,1
3	26	25	1	96,2
4	34	32	2	94,1
5	22	20	2	90,9
6	31	29	2	93,5
7	29	27	2	93,1
8	33	31	2	93,9
9	29	26	3	89,7
10	31	28	3	90,3
Total				928,9
Average				92,9

Table IX is the result of the testing scheme III, carried out by inserting a different sentence with Krama Javanese language. Krama Javanese language are polite Javanese and the vocabulary used is different from the Javanese Ngoko; some words are similar such as the conjunctions word that are used both, table X contains collection for the unknown word of Javanese Krama. In the scheme III experiment, all words in Krama Javanese were included in the unknown words and it does not solve the OOV problems, while those were not included in the unknown words were conjunctions and punctuations that had been trained before. The experiment was carried out ten times, and the system could not tag them properly. The obtained accuracy was very small with the highest accuracy of 38 %.

Table IX. ACCURACY OF TESTING DATA FOR UNKNOWN WORDS KRAMA JAVENESE THAT ARE NOT SOLVE THE OOV PROBLEM

No	Total Words	Correct Words	Incorrect Words	Accuracy %
1	43	8	35	18
2	30	9	21	30
3	24	4	20	16
4	35	5	30	14
5	93	30	63	32
6	46	10	36	21
7	13	4	9	30
8	33	8	25	24
9	24	7	17	29
10	47	18	29	38
Total				252
Average				25,2

Table X. UNKNOWN WORDS KRAMA JAVENESE LANGUAGE

Input Words	Unknown Words
Nanging amargi kondisi radin cupet uga kahanan lalu lintas ingkang padet, korban nyenggol gandengan truk. Pawingkingipun, korban dhawah datheng arah kiwa mlebet ing kolong gandengan truk.	['amargi', 'kondisi', 'radin', 'cupet', 'kahanan', 'lalu', 'lintas', 'ingkang', 'padet', 'nyenggol', 'gandengan', 'truk', 'Pawingkingipun', 'dhawah', 'datheng', 'arah', 'kiwa', 'mlebet', 'kolong', 'gandengan', 'truk']
Pawingkingipun, korban dhawah datheng arah kiwa mlebet ing kolong gandengan truk. Badan korban lajeng ketlindes roda gandengan truk ngantos tilar ing panggen. Kecelakaan punika sempat ndamel kahanan lalulintas wonten jalan raya Pesantren macet. Kahanan lalu lintas enggal lancar saksampune korban uga truk gandeng ingkang kacelakan disingkirake dening panjejibahan.	['Pawingkingipun', 'dhawah', 'datheng', 'arah', 'kiwa', 'mlebet', 'kolong', 'gandengan', 'truk', 'Badan', 'lajeng', 'ketlindes', 'roda', 'gandengan', 'truk', 'ngantos', 'tilar', 'panggen', 'Kecelakaan', 'punika', 'sempat', 'ndamel', 'kahanan', 'lalulintas', 'wonten', 'jalan', 'raya', 'Pesantren', 'macet', 'Kahanan', 'lalu', 'lintas', 'enggal', 'lancar', 'saksampune', 'truk', 'gandeng', 'ingkang', 'kacelakan', 'disingkirake', 'panjejibahan']
Kahanan lalu lintas enggal lancar saksampune korban uga truk gandeng ingkang kacelakan disingkirake dening panjejibahan. Kasus kecelakaan ingkang nimpa pelajar niki taksih dipuntumindakake penyelidikan Unit Laka Polres Kediri kutha. Panjejibahan sampun numindakake olah TKP mawi nggambar kronologi kecelakaan ing aspal.	['Kahanan', 'lalu', 'lintas', 'enggal', 'lancar', 'saksampune', 'truk', 'gandeng', 'ingkang', 'kacelakan', 'disingkirake', 'panjejibahan', 'Kasus', 'kecelakaan', 'ingkang', 'nimpa', 'pelajar', 'niki', 'taksih', 'dipuntumindakake', 'penyelidikan', 'Unit', 'Laka', 'Polres', 'Kediri', 'Panjejibahan', 'sampun', 'numindakake', 'olah', 'TKP', 'mawi', 'nggambar', 'kronologi', 'kecelakaan', 'aspal'
Miturut cerito turun temurun, rikala tentara Kasultanan Demak ingkang dipun pimpin dening Sunan Ngudung kaliyan Sunan Kudus nyerbu dhateng pusering kerajaan Mojopahit.	['cerito', 'turun', 'temurun', 'rikala', 'tentara', 'Kasultanan', 'Demak', 'ingkang', 'dipun', 'pimpin', 'Sunan', 'Ngudung', 'kaliyan', 'Sunan', 'Kudus', 'nyerbu', 'dhateng', 'pusering', 'kerajaan', 'Mojopahit']

VI. CONCLUSION AND RECOMMENDATIONS

A. Conclusions

This study focuses on the application of the Hidden Markov Model (HMM) method for Part of Speech Tagging in Javanese Language. The highest accuracy obtained from the system is 92.6 % in the testing data that is different from the training data with solve the OOV problem. Testing the system by entering a sentence taken in part from the data train gets an accuracy of 96 %. This research also conducted an experiment using Javanese Krama as an input sentence, and the accuracy was 38 %. From the accuracy results that was obtained, the Hidden Markov Model method can be used but it is still not optimal for Part of Speech Tagging in the Javanese Ngoko language. Because it requires a lot of corpus data so that the accuracy results obtained can be maximum. From the test results by testing different data from the training data get a large number of unknown words, this shows that the HMM method for POS tags requires large corpus data too so it can be minimized the existence of unknown word.

B. Recommendations

Here are some suggestions related to the next POS Tag research for the Javanese language. For further research, it is expected to be able to use more corpus during training. In the corpus gives a special label to indicate all foreign words other than the Javanese Ngoko language, so that the system can learn and overcome the labeling for possible input sentences other than the Ngoko Javanese language. Make improvements to the accuracy of the model so that test accuracy can be optimized. Use another method for POS Tags in Javanese to get the best method for POS Tags in Javanese so that they can be compared between methods for POS Tags in Javanese.

REFERENCES

[1] Nitin Sabloak, Bebeto Agung Hardono, and Derry Alamsyah. Part-of-speech (pos) tagging bahasa indonesia menggunakan algiritma viterbi. *STMIK MDP*, 2016.

[2] Hafiz Ridha Pramudita, Ema Utami, and Armadyah Amborowati. Pengaruh part of speech tagging berbasis aturan dan distribusi probabilitas maximum entropy untuk bahasa jawa krama. *Jurnal Buana Informatika*, 7(4), 2016.

[3] Pujiati Suyata. Status isolek yogyakarta-surakarta dan implikasinya terhadap bahasa jawa standar, 2011.

[4] Ruli Manurung. Tutorial: Pengenalan terhadap pos tagging dan probabilistic parsing. In *Workshop Nasional INACL, Jakarta*, 2016.

[5] Alfan Farizki Wicaksono and Ayu Purwarianti. Hmm based part of speech tagger for bahasa indonesia. In *Fourth International MALINDO Workshop, Jakarta*, 2010.

[6] Kathryn Widhiyanti and Agus Harjoko. Pos tagging bahasa indonesia dengan hmm dan rule based. *Informatika: Jurnal Teknologi Komputer dan Informatika*, 8(2):68208, 2012.

[7] Ahmad Zuli Amrullah, Rudy Hartanto, and I Wayan Mustika. A comparison of different part-of-speech tagging technique for text in bahasa indonesia. In *2017 7th International Annual Engineering Seminar (InAES)*, pages 1–5. IEEE, 2017.

[8] R Sandra Yuwana, Asri R Yuliani, and Hilman F Pardede. On part of speech tagger for indonesian language. In *2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, pages 369–372. IEEE, 2017.

[9] Dilmi Gunasekara, WV Welgama, and AR Weerasinghe. Hybrid part of speech tagger for sinhala language. In *2016 Sixteenth International Conference on Advances in ICT for Emerging Regions (ICTer)*, pages 41–48. IEEE, 2016.

[10] S. L. Ibañez J. L. V Agawa, M. V. Z. Bernabe and J. M. Salcedo. A hmm – viterbi algorithm based part of speech tagger for tagalog.

[11] A Kelvin. Part of speech tagger untuk bahasa indonesia menggunakan konsep hidden markov model (hmm) dan algoritma viterbi, 2018. Online.

[12] Umriya Afini. Penerapan analisis morfologi untuk penanganan kata berimbuhan pada pos tagger bahasa indonesia berbasis-statistik. *Universitas Dian Nuswantoro*, 2016.

[13] Femphy Pisceldo, Ruli Manurung, and Mirna Adriani. Probabilistic part of speech tagging for bahasa indonesia. In *Third International MALINDO Workshop*, pages 1–6, 2009.

[14] Rusyidi, Mulyanto R.J, Supardiman, and W Sutadi. *Kosa Kata Bahasa Jawa*. Pusat Pembinaan dan Pengembangan Bahasa, Departemen Pendidikan dan Kebudayaan, 1985.