Building Balinese Part-Of-Speech Tagger Using Hidden Markov Model (HMM)

I Gde Made Hendra Pradiptha[a1], Ngurah Agus Sanjaya ER[a2]

[a]Informatic Departement, Udayana University
Bali, Indonesia
[1]hendrapradiptha98@gmail.com
[2]agus_sanjaya@unud.ac.id

Abstract

Part-of-Speech tagging or word class labeling is a process for labeling a word class in a word in a sentence. Previous research on POS Tagger, especially for Indonesian, has been done using various approaches and obtained high accuracy values. However, not many researchers have built POS Tagger for Balinese. In this article, we are interested in building a POS Tagger for Balinese using a probabilistic approach, specifically the Hidden Markov Model (HMM). HMM is selected to deal with ambiguity since it gives higher accuracy and fast processing time. We used k-fold cross-validation (with k = 10) and tagged corpus around 3669 tokens with 21 tags. Based on the experiments conducted, the HMM method obtained an accuracy of 68.56%*.*

Keywords: *Part-of-Speech, POS Tagger, Balinese, Corpus, Tagset, Hidden Makarov Model, HMM.*

1. Introduction

Part-of-Speech tagging or word class tagging is a process to label a word class in a word in a sentence [1]. Part-of-Speech (POS) tagging is part of Natural Language Processing (NLP). The word class labels obtained from the POS Tagging process in documents can help the development of various other NLP applications such as Language models, Information Retrieval, Text Summarization, Machine Translation, and others. So that Part-of-Speech Tagging becomes one of the important studies in the field of NLP [2].

Manual word-class labeling is a labor-intensive, expensive, and time-consuming activity. Therefore, we need an approach that can do POS tagging automatically. In the last few decades, various approaches to automated POS Tagging have been developed. In general, these approaches can be divided into three, namely: rule-based, stochastic or probabilistic, and transformation-based or hybrid [3]. In a rule-based approach [4], POS Tagger will label word classes based on linguistic rules that are manually constructed by experts. A rule-based approach may be difficult to apply because constructing all linguistic rules in a language will not be easy. In the probabilistic approach [3], [5], the POS tagger will label the most likely word class, based on the probability value obtained from the manually labeled corpus. Meanwhile, in a Transformation-based approach [6], POS Tagger uses a combination of a rule-based and probabilistic approach.

Previous research on POS Tagger, especially for Indonesian, has been carried out using various approaches and obtained high accuracy. Several methods have been used to develop POS Tagger in Indonesian, including Maximum Entropy [3] with an accuracy value of 88.43%, Rule-based [4] with an accuracy value of 79%, HMM [5] with an accuracy value of 96.5%, Brill Tagger [6] with an accuracy value of 89.70%.

Meanwhile, in Indonesia, people do not only use Indonesian. Indonesian people also use the local language as the language of communication. Ethnologue [7] states that Indonesia has 720 regional languages, where 710 are living and 12 are extinct. The Balinese language is one of the regional languages that still living today with more than 3 million speakers spread across the islands of Bali, Nusa Tenggara, and South Lombok. However, research and resources on NLP for Balinese are still insufficient to date. Therefore, we wants to do research on NLP, especially in developing POS Tagging for Balinese. From the various existing approaches, we used the Hidden Markov Model (HMM) approach. HMM is selected to deal with ambiguity since it gives higher accuracy and fast processing time.

2. Reseach Methods

The process of developing POS Tagger for Balinese using HMM through preprocessing stages, training the HMM model, then predicting the word class for each word. The preprocessing stage used is tokenization. The results from the preprocessing will be used in the HMM model training process. The HMM model training receives input in the form of tokens and the corresponding tag and produces an HMM model. Then predict the word class of each input word based on the model that has been made. The following is a diagram of the POS Tagger development process for Balinese using HMM which can be seen in Figure 1.

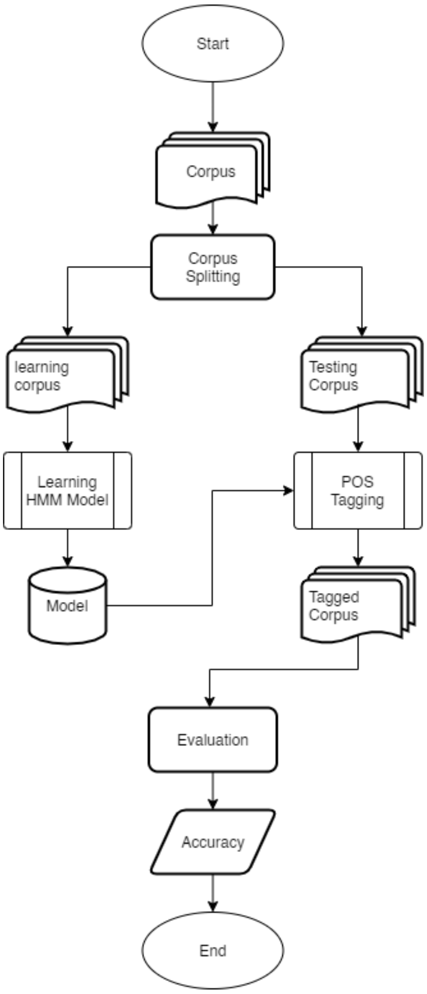

Figure 1. Research Design Flowchart

2.1. Data Collection

The dataset or corpus used in this study is secondary data. The corpus is compiled from examples of basic sentence patterns in Balinese from books [8]–[10] and Balinese language news sourced from the internet. In the corpus, there are 500 sentences and about 3500 words that have been manually tagged. The class of Balinese words used in this study is adjusted to the Penn Treebank POS tagset. There are a total of 20 classes of Balinese words used in this study. Table 1 is a list of POS tagset with descriptions and examples.

Table 1. Tagset for bahasa Bali

Num.	Tag	Description	Example
1	CC	Coordinating conjunction, also called coordinator.	Lan, tur, muah
2	CD	Cardinal number.	Abesik, dadua, seket, karo belah, 7916, 0,255.
3	DT	Determiner / article.	I, Ni, Ipun.
4	FW	Foreign word.	*Online, handphone.*
5	IN	Preposition.	di, ka, uli, ring.
6	JJ	Adjective.	ageng, selem, manis
7	MD	Modal and auxiliary verb.	suba, sampun, mangda.
8	NEG	Negation.	sing, nenten, eda.
9	NN	Noun.	baju, jaler, toko.
10	NND	Classifier, partitive, and measurement noun.	ukud, katih, ijas.
11	NNP	Proper noun.	Surabaya, Denpasar, Singaraja.
12	PR	Pronoun.	Tiang, Ragane, cai, ento, ia, niki
13	RB	Adverb.	Teken, olih, lakar
14	RP	Particle.	ja, ke, teh.
15	SC	Subordinating conjunction, also called subordinator.	sawireh, sane, krana
16	SYM	Symbol.	+, #, $, IDR, +, %, @
17	UH	Interjection.	nget, jeg, pih, ih, beh, aduh, aruh.
18	VB	Verbs.	meli, memunyi, mulih.
19	WH	Question.	sire, kenapi, nyen, dija.
20	X	Unknown	
21	Z	Punctuation	. ! ? : ; ( ) “ ‘

2.2. Data Preprocessing

Data preprocessing is the process of preparing a dataset that is used to become data that can be processed at a later stage. In this study, the preprocessing stage used was tokenization. Tokenization is the process of breaking a string into its smallest form which is called a token. Tokens can be in the form of characters, words, sentences, or paragraphs depending on the research needs. In this research, the token in question is the word and punctuation.

2.3. Hidden Markov Model (HMM)

Hidden Markov Model (HMM) is one of the most popular approaches used to solve sequence labeling problems such as POS Tagging. HMM is statistical modeling of a system that can determine hidden parameters from observable parameters. In POS Tagging HMM allows us to talk about the Markov Model observed events (such as the words we see in the input) and hidden events (such as the part-of-speech tag) which we consider to be causal factors in our probabilistic model [1]. In general, the components contained in HMM are:

a. Set of N states
   $Q = q_1 q_2 \dots q_N$
b. Transition probability matrix
   $A = a_{11} \dots a_{ij} \dots a_{NN}$
c. Sequence of *T* observations
   $O = o_1 o_2 \dots o_T$
d. Sequence of observation likelihoods
   $B = b_i(o_i)$
e. Initial probability distribution
   $\pi = \pi_1, \pi_2, \dots, \pi_N$

HMM POS Tagger selects the appropriate label order by maximizing the following equation (1):

$$P(tag|previous\ n\ tag) * P(word|tag) * \tag{1}$$

More specifically, HMM uses the following formula which can be used to find the appropriate label sequence from certain word order.

$$\Pr(t_{i,n}\,, w_{1,n}) \approx \prod\nolimits_{i=1}^{n} (\Pr(t_i|t_{i-k,i-1}) * \Pr(w_i|t_i) \tag{2}$$

The current probability of the t_i tag depends on the t_{i-1} tag and the probability of the current word depends only on the current tag (t_i).

2.4. Evaluation

In this study, the results of the tagging of the testing data will be tested to determine the level of accuracy obtained. To calculate the accuracy value, it is done by counting the number of words that are labeled correctly with the total number of words in the testing data using equation (3).

$$Accuracy = \frac{Number\ of\ Correct\ word\ tag}{Total\ number\ of\ words\ in\ test\ data} \tag{3}$$

3. Result and Discussion

To find out the performance of the POS Tagger that was built, a testing or evaluation stage was carried out to obtain the resulting accuracy. The test was carried out using a Balinese language corpus consisting of 500 sentences and around 3669 tokens. In this test, we use k-fold cross-validation (with k = 10) to divide the corpus in the learning and testing process.

3.1. Balinese Tagged Corpus

In the Balinese corpus, we use the simplest tagged corpus format, which is the form of word/tag. Figure 2 provides some examples of a sentence in the corpus used in this research.

```
Dugas/NN icange/PR kema/VB ia/PR konden/NEG ngenah/VB ditu/RB ./Z
Daweg/NN titiange/PR mrika/VB ipun/PR dereng/NEG makanten/VB drika/RB ./Z
Icang/PR lakar/MD luas/VB ka/IN Denpasar/NNP ./Z
Bapak/DT Perbekel/NN pacang/MD ka/IN Badung/NNP ./Z
Ida/DT Pedanda/NN jagi/MD ngweda/VB ./Z
Senjatane/NN ento/IN tusing/NEG ja/NN bakal/MD mintulin/VB ./Z
Bakal/MD anggon/VB apa/WH ngalih/VB penyalin/NN ?/Z
Ibi/NN sanja/NN nandes/MD icang/PR cegut/VB kuluk/NN ./Z
Kayang/NN jani/NN ia/PR enu/MD nyilih/VB pipis/NN di/IN bank/NN ./Z
Ia/PR enu/MD ngrengkeng/VB kayang/IN jani/NN wireh/SC motorne/NN pasilihanga/VB ./Z
```

Figure 2. Balinese tagged corpus example

3.2. Tagging Example

In the tagging process, the system will accept input in the form of unlabeled text and produce text that has been tagged with the corresponding tag. In Table 2, we can see some of the results of tagging using the HMM method.

Table 2. Tagging example

Untagged Text	Adinne I Mendra jemet bin dueg di sekolahan . I Kasda lan adinne suba luas ka Tabanan . Titiang utawi ipun sane iwang . Sang Arjuna kacritayang seda utawi lina ring payudan punika . Dumadak ipun rahajeng utawi tan kapialangan ring marga mangda prasida rauh mriki .

	Sasampune kaicen jinah antuk biang ipun , raris ipun gageson lunga ka sekolahan . Sesubane nrima pipis uli memenne laut ngenggalang ia majalan ka sekolahan .
Tagged Text	Adinne/PR I/DT Mendra/NNP jemet/JJ bin/SC dueg/JJ di/IN sekolahan/NN ./Z I/DT Kasda/NNP lan/CC adinne/PR suba/RB luas/VB ka/IN Tabanan/NNP ./Z Titiang/PR utawi/CC ipun/PR sane/SC iwang/JJ ./Z Sang/DT Arjuna/NNP kacritayang/NN seda/VB utawi/CC lina/VB ring/IN payudan/VB punika/NN ./Z Dumadak/MD ipun/PR rahajeng/JJ utawi/CC tan/NEG kapialangan/VB ring/IN marga/NN mangda/SC prasida/RB rauh/VB mriki/NN ./Z Sasampune/RB kaicen/VB jinah/NN antuk/IN biang/NN ipun/PR ,/Z raris/SC ipun/PR gageson/RB lunga/VB ka/IN sekolahan/NN ./Z Sesubane/RB nrima/RB pipis/NN uli/IN memenne/NN laut/SC ngenggalang/RB ia/PR majalan/VB ka/IN sekolahan/NN ./Z

3.3. Performance and Result

In this study, we conducted tests using k-fold cross-validation (with k = 10). This means that the corpus will be divided into k folds. Then, every 1 fold will be used in the evaluation process and the rest (k - 1 folds) will be used in the learning process. The Table 3 shows the experimental results and the accuracy of each fold.

Table 3. POS tagger evaluation with 10-fold cross-validation

Fold	No. of Words	True	Accuracy
1	341	240	70.38%
2	433	303	69.97%
3	419	318	75.89%
4	370	212	57.29%
5	299	220	73.57%
6	284	193	67.95%
7	319	250	78.26%
8	389	269	69.15%
9	391	241	61.63%
10	424	261	61.55%
	Average		68.56%

4. Conclusion

From the research conducted, it can be concluded that the Hidden Markov Model method can be used as a POS Tagger for Balinese. From tests carried out using the 10 fold cross-validation method, HMM obtained an average accuracy of 68.56%. In this study, the limitations of the size and variation of the corpus used are still the main obstacles.

Future suggestions for this research are as follows: Evaluating the Bali Part-of-Speech tagset used. Developing a larger and more diverse Balinese language corpus, so that POS Tagger can recognize more diverse words and be able to obtain higher accuracy. Building a Balinese POS Tagger using other methods such as Rule-Based, Maximum Entropy, and Brill Tagger, to find out the best method for Bali POS Tagger

References

[1] D. Jurafsky and J. H. Martin, “Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition,” vol.

1, 2019, doi: 10.1515/zfsw.2002.21.1.134.

[2] N. Sabloak, “Part-of-Speech (POS) Tagging Bahasa Indonesia Menggunakan Algoritma Viterbi,” no. x, pp. 1–11, 2016.

[3] R. S. Yuwana, A. R. Yuliani, and H. F. Pardede, “On part of speech tagger for Indonesian language,” *Proc. - 2017 2nd Int. Conf. Inf. Technol. Inf. Syst. Electr. Eng. ICITISEE 2017*, vol. 2018-Janua, no. October, pp. 369–372, 2018, doi: 10.1109/ICITISEE.2017.8285530.

[4] F. Rashel, A. Luthfi, A. Dinakaramani, and R. Manurung, “Building an Indonesian rule-based part-of-speech tagger,” *Proc. Int. Conf. Asian Lang. Process. 2014, IALP 2014*, pp. 70–73, 2014, doi: 10.1109/IALP.2014.6973521.

[5] Muljono, U. Afini, C. Supriyanto, and R. A. Nugroho, “The development of Indonesian POS tagging system for computer-aided independent language learning,” *Int. J. Emerg. Technol. Learn.*, vol. 12, no. 11, pp. 138–150, 2017, doi: 10.3991/ijet.v12.i11.7383.

[6] E. R. Setyaningsih, “Part of Speech Tagger Untuk Bahasa Indonesia Dengan Menggunakan Modifikasi Brill,” *Din. Teknol.*, vol. 9, no. 1, pp. 37–42, 2017.

[7] D. M. Eberhard, G. F. Simons, and C. D. Fennig, *Ethnologue: Languages of the World*, Twenty-Thi. Dallas, Texas: SIL International, 2020.

[8] I. W. Bawa and I. W. Jendra, *Struktur Bahasa Bali*. Jakarta: Pusat Pembinaan dan Pengembandan Bahasa, Departemen Pendidikan dan Kebudayaan, 1981.

[9] I. W. Bawa, I. G. K. Anom, Margono, I. B. U. Naryana, and I. N. Medra, *Sintaksis Bahas Bali*. Jakarta: Pusat Pembinaan dan Pengembandan Bahasa, Departemen Pendidikan dan Kebudayaan, 1983.

[10] K. Ginarsa, M. Denes, A. M. Mbete, I. G. K. Ardhana, and I. K. Merta, *Kata Tugas Bahasa Bali*. Jakarta: Pusat Pembinaan dan Pengembandan Bahasa, Departemen Pendidikan dan Kebudayaan, 1984.