INFORMATION EXTRACTION OF SCIENTIFIC DOCUMENTS USING GENERALIZED HIDDEN MARKOV MODEL

Bhakti Junrio[1], Ken Kinanti Purnamasari[2]

[1.2]Universitas Komputer Indonesia
Jl. Dipati Ukur No.122-116, Lebakgede, Coblong, Bandung, West Java 20132
E-mail: bhaktijunrio15@gmail.com[1], ken.kinanti@email.unikom.ac.id[2]

ABSTRACT

Extraction of information can be applied to a wide variety of cases, one of them to detect any components on the scientific paper documents. Document scientific papers usually have a fairly diverse formats. The diversity that make it difficult to detect each of the components in it. In the case of scientific papers document the information extraction related been much research done previously, one of which documents the thesis scientific papers in Indonesian language. In the study carried out tests on 50 documents that were created from the years 2013 and 2017, obtained an average accuracy of 57%. This study was conducted using an algorithm Generalized Hidden Markov Models *(GHMM).* Previous research has been done using algorithms GHMM in the case of web information extraction and the introduction of DNA, *GHMM* has superior accuracy than other algorithms. Implementation *GHMM* to the document information extraction scientific papers based on testing that has been performed on the document data 40 scientific papers thesis, obtained an average score of 64% accuracy. Obtaining the token-class accuracy due to misspelling or typo of the document conversion and also use weighting feature.

Keywords : Information Extraction, *Generalized Hidden Markov Model*, *GHMM*, Document scientific papers, thesis.

1. INTRODUCTION

Extraction is a field of science information for data processing which aims to transform the data into structured information [1]. Another purpose of the information extraction is to get the facts of an event, entity and connectedness of the data set [2] The processed data can be text or also a variety of other data. The approach taken in the case of one of them using the information extraction machine learning, because in the process of development does not require much time, but the need for models and also a large training data.

In this study, document information extraction performed on the thesis scientific papers, where there are several documents retrieved that document cover and abstract. Previous research has been done in the case of document information extraction thesis scientific papers in Indonesian language. In the study carried out by taking cover and abstracts of scientific papers document the Indonesian-language thesis and obtained the value of the accuracy of testing three documents, namely 100%. As for the testing of 50 documents obtained an average accuracy of 57% [3]. The same study but using different algorithms, namely LVQ, obtained an average accuracy of 39% were carried out on 40 of documents data [4].

Research by the information extraction case on scientific papers thesis Indonesian language use *GHMM* algorithm has never been applied before. In the English-language journal *GHMM* algorithm is never applied to the case in addition to the extraction of information, and also in the case of information extraction. In the case of Web information extraction is the extraction of information by testing a web page obtained an accuracy of 85.5% [5]. Whereas in the case of the introduction of human genes contained in the DNA obtained an accuracy of 85% [6].

Based on the above research *GHMM* note that the algorithm can be applied to various cases. Yet their use GHMM algorithms for information extraction case the thesis scientific papers document the Indonesian language, into consideration its possible to do testing. This study aims to establish an information extraction system on the thesis scientific papers document the Indonesian language and also to determine the resulting accuracy of the information extraction cases of scientific papers document paper uses an algorithm *GHMM* Indonesian language. Limitation of documents used in this study is to document scientific papers thesis at the University Computer Indonesia (UNIKOM).

2. RESEARCH CONTENT

On the part of the contents of this study explain a few things that include research methods, scientific papers document paper, system architecture, tokenization, data segmentation, feature extraction, algorithms GHMM, accuracy and error, and the test results.

2.1 Research Methods

Stages of the research methods used in this research is quantitative method. The stages are carried out here starting from the study of literature, collection of datasets, software development and examination. Block diagram of research methods presented in the following figure.

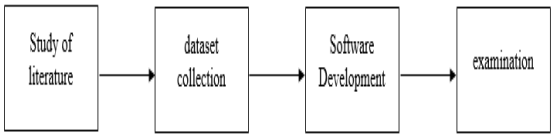

Figure 1. Block Diagram Research Methods

2.2 Scientific Documents

Document is written or report that presents the results of research or assessment of an issue that are written or printed and can be used as proof of a scientific paper [7]. Scientific work must same to the rules and ethics of science that confirmed and adhered to by the people. The data, conclusions and other information contained in scientific papers are used as a reference (reference) for other scientists in conducting research or further studies.

Thesis is a scientific paper that suggests the author's opinion based on what others are supported by data and facts to complete the requirements to obtain a bachelor's degree from a college [8].

Research will be made using scientific papers document the thesis on the cover sheet and abstract as the data that aims to take the information contained therein, because in a scientific paper documents there are different categories of the assortment. On the cover sheet has a category title of the study, this type of research, sentence submission, name, nim, courses, faculty, university, and also year. While on a heading of research abstract (abstract), by name, nim, the contents of the abstract, and keywords.

2.3 Natural Language Processing (NLP)

Natural language processing or commonly known as NLP lot of its tools. NLP in many cases is very important to get good accuracy. As in detection stemming basic words and word type detection (Marking PO). In many cases the use of NLP much did one study about stemming and POS tagging. in penenganannya NLP sometimes have problems one of which in the case above is the lack of corpus in Indonesia and incomplete rules on stemming the available research, first published by Julia Beth Lovins in 1968 [11].

2.4 System Architecture

Information extraction system using GHMM (Generalized Hidden Markov Model) has several processes from the input data input traning and testing, training and testing of data preprocessing in which there are the segmentation process line, tokenisasi said, feature extraction. The results of preprocessing the training data will be used in testing at GHMM. The picture of the information extraction system using GHMM is as follows.

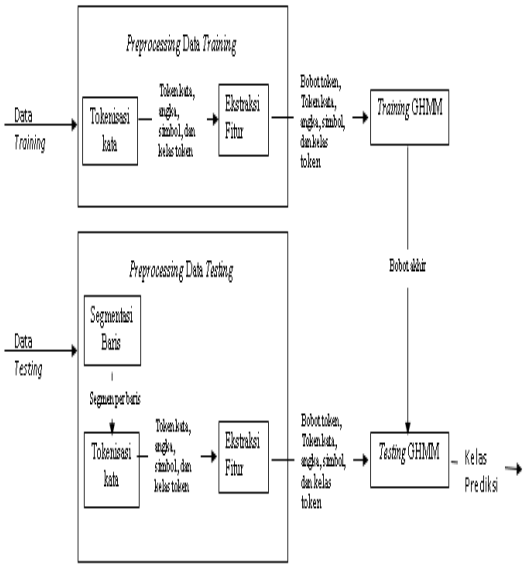

Figure 2. System Architecture

Figure 2 is a system overview in this study. Preprocessing stages required training data is data in the form of type .CSV file. CSV file contains the data set that has been labeled from each category. Tokenization used to separate every word in the sentence into tokens that can be weighted through the process of feature extraction.

On the Step to testing data preprocessing, what is needed is data in the form of type .TXT file. .TXT files will be processed into the stages of segmentation lines, tokenizationi, and feature extraction. At each word line segment will be divided according to the lines formed by the array. Tokenization divide or separate each word of the results of data segmentation. Feature extraction is a gift token feature on every word.

2.5 Tokenization

Tokenization is a cutting or splitting the string of each sentence to be broken up into single words that stand alone. Separation sentence into word by word by utilizing the space. Here is an example of tokenization in Indonesian language.

Table 1. Tokenisasi

text input	word
MAPPING SYSTEM ELECTRICAL SUBSTATION BASED Dekstop IN PLN UPJ Cileungsi	MAPPING
	SYSTEM
	ELECTRICAL
	SUBSTATION
	BASED
	Dekstop
	IN
	PLN
	UPJ
	Cileungsi

2.6 Data Segmentation

Data segmentation aims to remove rows and columns to produce a data segment that has become per align. Data segmentation performed on the data preprocessing stages of testing the input as a text file in the form of txt. The following block diagram of the data segmentation.

Table 2. Data Segmentation

Sentence	Segmentation Results
MAPPING SYSTEM ELECTRICAL SUBSTATION BASED Dekstop IN PLN UPJ Cileungsi ESSAY	MAPPING SYSTEM ELECTRICAL SUBSTATION BASED Dekstop IN PLN UPJ Cileungsi ESSAY

2.7 Feature Extraction

Is a weighting value in the features contained in each token word from the tokenization.

Extraction feature of this research is to provide a special feature on a word tokens that can then be recognized by the system and will be given the value of the weight on the word. Selection of the right features can increase accuracy in labeling.

Extraction of features used in this study as many as 15 pieces features, refer to the document information extraction research scientific paper using LVQ by Firdamdam [4]. The explanation of feature extraction as follows.

,

Table 3. Feature Extraction

No.	name Features	Information
1	INITCAPS	Recognize each token that begins with a capital letter.
2	ALLCAPS	Recognize each token that all capital letters.
3	CONTAINSDIGIT	Recognize each token containing a digit.
4	ALLDIGIT	Recognize each token that is all digits.
5	CONTAINSDOTS	Recognize each token containing a point
6	lowercase	Recognize each token that is all lowercase.
7	punctuation	Recognize each token containing certain signs such as periods, commas, colons, semicolons, brackets, and the exclamation mark.
8	EIGHTDIGIT	Additional features of this study, this feature is devoted to recognizing tokens with digits with a length of 8 digits
9	WORD	Additional features of this study, this feature is devoted to give weight to class JENIS_PENELITIAN token, KALIMAT_PENGAJUAN.
10	LINE_START	Recognizing the token position at the beginning of the array index.
11	LINE_IN	Recognizing the token position at the middle of the array index.
12	LINE_END	Recognizing the token position at the end of the array index.
13	PERSON	Token recognize a person's name
14	ORGANIZATION	Recognizing token an organization
15	YEAR	Recognizing the characteristics token years.

Any existing features will be assigned a weight value of 1 or 0, if the token word including one of the above features will be given a weight of 1, otherwise it will be given a weight of 0. Here is a picture of the training data feature extraction and feature extraction of data testing.

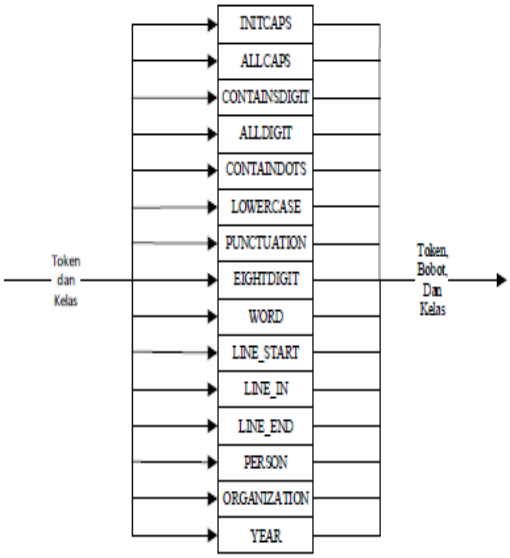

Figure 3. Block Digaram Feature Extraction

2.8 Algorithms GHMM

Generalized HMM including the development of HMM (hidden Markov model). Generalized hidden Markov models is a method that based probability interval, which is probability interval used to capture the uncertainty of the calculations produced by the usual HMM. The results given by GHMM in some complex cases can improve the accuracy resulting from the merger GHMM existing errors during the data collection phase of learning and recognition features.

GHMM need the training stages on the input data and features classes and word tokens. The very beginning of the training is to look for the parameters of GHMM, there are 3 main parameter above its value must be known beforehand that are included in the parameters of λ = (A, B,**π).** In general, the steps in the process depicted as in Figure 4.

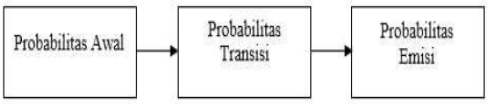

Figure 4. Schematic Probability GHMM

2.8.1 training GHMM

1. Initial probability

Initial probability is a chance that there is in a particular state as the beginning of a situation. Where π_i is the result of the probability q_i that the value of the hidden conditions on the word i, and S_i the total number of existing class to class to i. The following equation at 2.1 [9].

$$\pi_i = P\ (q_i = S_i\ ) \qquad (2.1)$$

Where :

π_i = Initial probability of a word at the beginning of class to -i

i = The word at the beginning of the i-th grade

P = Probability

q_i = Hidden conditions on the word at the beginning of the i-th grade

S_i = The total number all of classes until to-i

2. Transition probabilities

The probability of chance is a condition for switching from the first grade to the next grade. q_j The next class is hidden condition when the condition q_i of the current class. Transition probabilities can be denoted by A = {aij}. The following equation at 2.2 [9].

$$a_{ij} = P\ (q_i \text{ at } t+1 \mid q_j \text{ at } t) \qquad (2.2)$$

Where :

a_{ij} = Probability of transition from i to j

i = is the current class

j = is the next state

t = is the state that will be passed

P = is the probability

q_j = Hidden conditions on the circumstances next

q_i = Hidden conditions on the current state

t + 1 = is the order of the index-t + 1

3. probability Emissions

Is the probability of a probability distribution of the state of observation which produces a new state. Emission is obtained from the number of occurrences of the features of the class and divided by the total number of token words that contain features that are sought. Suppose that the emissions will be the condition of each token words that are from class 1 to 16, are included into allcaps if included then rated the weights 1, do check until the early grades have all been processed and added together, then divided by the number of tokens containing allcaps in all classes.

Inside GHMM emission probabilities denoted by B = (b_ (I) (O_ (t))). 2.3 The following equation [9].

$$b_j\ (O_t\ ) = P\ (O \text{ on } t \mid q_j \text{ on } t) \qquad (2.3)$$

Where :

$b_j\ (O_t\ )$ = The probability of emission into the state j t observersi All

P = Is the probability

O = Is the observation conditions

t = Is the word index to t

q_j = Is the Features

4. Observations Sequence, Upper and Lower

Observations Sequence will be calculated on the overall value of the results of the initial probability, transitional and emissions. λ is a container of values that have been sought earlier, onπ, B, A in Equation 2.1, 2.2 and 2.3. while the observation sequence O is the observation sequence muncuknya opportunities forπ, B, A, can be seen in equation 2.4 [9]. Observations upper is the upper limit of the elements contained on the observations of λ which is the value of the initial probability, transitional and emissions, while observations lower the lower limit of the elements contained on the observations of λ, can be seen in equation 2.4, 2.5, 2.6 [9].

Observation Sequence P ($O \mid \lambda$) (2.4)

Observation Sequence upper ($\bar{O} \mid \lambda$) (2.5)

Observation Sequence lower ($\underline{O} \mid \lambda$) (2.6)

Where :

O = Observation

λ = The value of the initial probability, transition and emission

$\bar{O}$ = Is the upper observation

$\underline{O}$ = Is the upper observation

5. Probability generalization initial

Generalization initial probability here to find the value of the lower and upper limits of the elements contained in the results obtained from the calculation of the initial probability of the equation 2.1 [9]. The equation for the lower and upper at 2.7 and 2.8 [9]:

$$\overline{\pi_i} = \overline{P}\,(q_1 = S_i)\,(1 \le i\ N \le) \quad (2.7)$$

$$\underline{\pi_i} = \underline{P}\,(q_1 = S_i)\,(1 \le i\ N \le) \quad (2.8)$$

Where :

$\overline{\pi_i}$ = Initial probability upper

$\underline{\pi_i}$ = lower initial probability

P = Probability

q_i = Hidden conditions on the word at the beginning of the i-th grade

S_i = The total number of classes until all i

6. Generalizing the Transition Probability

Generalization transition probabilities here to find the value of the lower and upper limits of the elements contained in the results obtained from the calculation of the transition probabilities of the equation 2.2 [9]. The equation for the lower and upper at 2.9 and 2:10 [9]:

$$A = a_{ij} = p\,(q_{t+1} = S_j \mid q_t = S_i) \quad (1 \le i, j \le N) \quad (2.9)$$

$$a_{ij} \ge 0, \sum_{j=1} a_{ij} = 1\ (1 \le i, j \le N)$$

$$\overline{A} = \overline{a_{ij}} = p\,(q_{t+1} = S_j \mid q_t = S_i) \quad (1 \le i, j \le N) \quad (2.10)$$

$$a_{ij} \ge 0, \sum_{j=1} a_{ij} = 1\ (1 \le i, j \le N)$$

Where :

$\underline{a_{ij}}$ = Value of the upper generalization transition probabilities

$\underline{a_{ij}} =$ = Value lower generalization of the transition probabilities

q_t = a hidden state +1

S_j = is The next class

S_i = is class

7. Probability generalization Emissions

Generalisai emission probabilities to find the lower limit value and the upper limit of the elements contained in the results obtained from the calculation of the probability of emission of the equation 2.3 [9]. The equation for the lower and upper at 2.11 and 2.12 [9]:

$$\underline{B} = \underline{B}j\,(k) = p\,(o_t = v_k \mid q_t = S_j) \quad (2.11)$$

$$\overline{B} = \overline{Bj}(k) = p\,(o_t = v_k \mid q_t = S_j)\,(2.12)$$

Where :

$\underline{Bj}$ (K) = probability of lower emissions into the state j t observersi All

$\overline{Bj}(k)$ = j to the emission upper probability on the state observersi All t

P = Is the probability upper

0 = Adah observation conditions

V = Was featured on the class

vk = Is the next feature

2.8.2 testing GHMM

Viterbi algorithm is commonly used to resolve issues related to coding, can also to complete the other fields. GHMM Viterbi algorithm in handy to find the value of the hidden state sequence of a sequence of observations maximum or optimal value. In Viterbi there are several stages, the following are the stages in the viterbi.

1. Initialization

The first stage is the initialization. At this stage it will be searched starting value resulting from the Viterbi. The equation is at 2.13 and 2.14 [9] [10]. Initialization lower formulated with:

$$\delta_1^l(i) = \pi_i b_i(\underline{o_1}) \ (1 \le i \le N) \qquad (2.13)$$

Initialization upper formulated with:

$$\delta_1^u(i) = \pi_i b_i(\overline{o_1}) \ (1 \le i \le N) \qquad (2.14)$$

$$\psi_1(i) = 0 \ (1 \le i \le N)$$

Where :

$\delta_1^l(i)$ = Is a lower initial initialization conditions
$\delta_1^u(i)$ = Is the initial condition of the initial upper
π_i = Is the initial probability
b_i = Is the probability of emission
$\underline{O}$ = Lower observation
$\overline{O}$ = upper observation

2. Induction

In the induction phase will count the results obtained from the initialization, the calculation is done as much as the existing class. Here is the formula are looking for induction. Here equationd 2.15 and 2.16 [9]:

Lower induction formulated with:

$$\delta_1^1(j) = \max_{1 \le i \le N} \{\delta_{t-1}^1(i)\, a_{ij}\}\, b_j\left(\underline{o_t}\right)$$
$$(2 \le t \le T,\ 1 \le j \le N) \qquad (2.15)$$

Induction upper formulated with:

$$\delta_1^u(j) = \max_{1 \le i \le N} \{\delta_{t-1}^u(i)\, a_{ij}\}\, b_j(\overline{o_t})$$
$$(2 \le t \le T,\ 1 \le j \le N) \qquad (2.16)$$

$$\psi_1(j) = argmax \left\{ \min_{1 \le i \le N} (pro\ \delta_{t-1}^1(i)\, a_{ij}),\ \min_{1 < i < N} (pro\ \delta_{t-1}^u(i)\, a_{ij} \right\}$$

Where :

$\delta_1^l(j)$ = Initial each induction results lower
$\delta_1^u(i)$ = Initial each upper induction results
a_{ij} = Is the probability of transition
b_i = Is the probability of emission
$pro\ \delta_{t-1}^1$ = Penginisialisasian result of induction minus 1
$\psi_1\ (j)$ = Saving the results of induction

3. Terminations

At the stage of termination will take to get results or best value. Here is the formula to find the termination. At 2.17 the following equation [9].

$$P^* = \max_{1 \le i \le N} \{\delta_T^l(i), \delta_T^u(i)\}, \qquad (2.17)$$

$$q_T^* = argmax \left\{ \min_{1 \le i \le N} (pro\delta_T^l(i)), \min_{1 \le i \le N} (pro\delta_T^u(i)) \right\}$$
,

Where :

P^* = Probability of best value
$\max_{1 \le i \le N}$ = Result of the maximum value of the recursion
q_T^* = Hidden Conditions
$argmax$ = The contents resulting from nilaimaksimal

4. Backtracking

At the stage of backtracking, will check back in the best order resulting from the termination. Here is the calculation for Backtracking at 2.18 equations [9].

$$Q_t^* = \psi_{t+1}(Q_{t+1}^*), t = T-1, T-2, \dots, 1 \qquad (2.18)$$

Where :

Q_t^* = Hidden Condition
$\psi_{t+1}(Q_{t+1}^*)$ = Matrix Storage hidden state
q_T^* = Hidden = Conditions
T = Number T minus 1

2.9 value Accuracy

Values obtained by measuring accuracy token clarified the truth value, divided by the number of tokens, and multiplied by 100%. Here's the formula calculation accuracy value.

$$Akurasi = \frac{Keseluruhan\ data\ benar}{Keseluruhan\ testing\ data}\ x\ 100\%$$

2.10 value Error

The error value is the value obtained when some classes are not clarified properly.
Here's the formula calculation error.

$$Error = \frac{Keseluruhan\ data\ salah}{Keseluruhan\ testing\ data}\ x\ 100\%$$

2.11 Test result

The test results with the data of 40 is done with the concept of class tokens are as follows. Testing the accuracy values shown in Table 4.

Table 4. Accuracy Results

No.	Document name	accuracy
1	1.txt	65%
2	2.txt	77%
3	3.txt	43%
4	4.txt	72%
5	5.txt	58%
6	6.txt	66%
7	7.txt	66%
8	8.txt	73%
9	9.txt	72%
10	10.txt	53%
11	11.txt	72%
12	12.txt	72%
13	13.txt	65%
14	14.txt	52%
15	15.txt	72%
16	16.txt	80%
17	17.txt	52%
18	18.txt	55%
19	19.txt	63%
20	20.txt	71%
21	21.txt	53%
22	22.txt	56%
23	23.txt	74%
24	24.txt	60%
25	25.txt	69%
26	26.txt	65%
27	27.txt	61%
28	28.txt	66%
29	29.txt	55%
30	30.txt	74%
31	31.txt	57%
32	32.txt	54%
33	33.txt	73%
34	34.txt	60%
35	35.txt	64%
36	36.txt	68%
37	37.txt	69%
38	38.txt	57%
39	39.txt	69%
40	40.txt	70%
	Average	**64%**

2.12 Test Result

The value of accuracy and error testing to validate the token-class on comparable data and prediction data. Validation is done between data_testing.csv of the results of plotting outside the system as the original data and the test data class as a data token predictions. The high value of the low accuracy and error obtained have several causes. The following analysis of the results of testing accuracy and error with token-class concept. Analysis of test results with token-class concept:

1. Because the input data used by the system as a text file, then there are limitations on the features used.
2. Low accuracy value obtained for the testing documents are some of the symbols that irregular dihasilkam by previous conservation process.
3. Based on the observations made, the impact which affects the accuracy of impairment is always caused by ORGANIZATION feature devoted to gave the weighting in the category Studies Program, Faculty and University.

3. CONCLUSION

Based on the test system functionality and accuracy measurements have been done using 40 the data testing, system information extraction algorithm using Generalized Hidden Markov Model (GHMM) has successfully dibagun indicated by the accuracy of the token-class with accuracy correctly by 64% and error by 36%.

In this study still has some shortcomings, so that the value of the accuracy of the token-class concept can not be obtained with the maximum. Thus, some suggestions will be presented for the future

development of the information system using machine learning, including:

1. Using the input data with a .docx or .html format in order to be able to use some additional features such as detecting bold, italic, underline, font style and some others.
2. Need demonstrated by comparison GHMM algorithm parameters in the extraction of information on scientific papers document.
3. Needed checking and correcting for spelling mistakes or typos on the PDF conversion results.

BIBLIOGRAPHY

[1] D. Jurafsky and JH Martin, "Chapter 21 - Information Extraction," in Speech and Language Processing, 2017, pp. 1-31.

[2] E. Susanti and K. Mustafa, Ektraski Information Web Pages Using bootstrapping approach in Ontology-Based Information Extraction, vol. 9, pp. 111-120, 2015.

[3] D. Mustaqwa, "Implementation of Information Extraction In Final Text Documents Using Rule Based Method," Thesis, University Computer Indonesia, Bandung, West Java, in 2018 ..

[4] F. Sasmita, "Document Information Extraction Essay Learning Algorithm Using Vector Quantiation," 2018.

[5] MEB Prasetyo, "Basic Theory of Hidden Markov Model," 2011.

[6] K. David, H. David and EF H, "A Generalized Hidden Markov Model for the Recognition of Human Genes in the DNA," 1996

[7] S. Lestanti and AD Susana, Document Archiving System Master And Servant Using Mixture Modeling Methods Based WEB, vol. X, 2016

[8] A. Wahyuni, Cryptography Application for Development of E-Documents With Hybrid Methods: Biometric Signature and DSA (Digital Signature Algorithm), pp. 1 to 123.2011

[9] XY Feng, HM You and W. yan, "A Generalized Hidden Markov Model and Its Applications In Recognition of Cutting States," 2016.

[10] K. David, H. David and EF H, "A Generalized Hidden Markov Model for the Recognition of Human Genes in the DNA," in 1996.

[11] KK Purnamasari and IS Suwadi, "Rule-Based Part Of Speech Tagger For Indonesia Language," IOP Confrence Series: Materials Science and Engineering, vol.407, no. 012 151, pp. 1-4, 2018.