Academic Information Service Chatbot Using HMM and AIML

Muhammad Affandes[1], Muhammad Juanda[2], Muhammad Fikry[3], Pizaini[4]
[1,2,3,4]Informatics Engineering, Faculty of Sciences and Technology,
Universitas Islam Negeri Sultan Syarif Kasim, Indonesia
[1]affandes@uin-suska.ac.id, [2]muhamad.juanda@gmail.com, [3]muhammad.fikry@uin-suska.ac.id, [4]pizaini@ uin-suska.ac.id

Abstract. UIN Suska Riau campus led to an escalation amount of data and information that must be maintained, such as academic information. UIN Suska Riau is responsible for managing and providing academic information to students and other academic communities. We can ask the Customer Care Center (C3) in Academic System or come directly to the PTIPD UIN Suska Riau office for academic questions. There still has limitations to serving existing questions submitted through C3 because officers can only serve during working hours both online and offline. Chatbots can be used to support the work of C3 officers in serving the questions asked. This system is built based on Named Entity Recognition (NER) using Artificial Intelligence Markup Language (AIML). We perform NER analysis using HMM. This study uses the contents of the academic manual as a base knowledge with 150 categories of questions and 30 answers that produce an accuracy of 55%.

Keywords: AIML, Named Entity Recognition, Hidden Markov Model, Chatbot

Received October 2022 / **Revised** November 2022 / **Accepted** November 2022

INTRODUCTION

The development of UIN Suska Riau as fast as an increasing number of civitas in UIN Suska Riau also causes an increase in the amount of data and information that must be managed, one of which is academic data. The campus is responsible for providing academic information to students and lecturers. That is why new students are usually given an academic guidebook and a website containing lecture rules. The management of student academic data such as student profiles and grade transcripts is currently managed by an academic information system, namely iRaise (Integrated Academic Information System).

Currently, C3 (Customer Care Center) or customer support is available for users to ask questions or problems regarding academics and the system, but this is considered less effective and less interactive in terms of time, namely the limitation of responding to questions that cannot be done outside working hours. Meanwhile, on the other hand, students did not only ask questions during working hours, few of them asked questions to C3 outside of working hours. This way we need chatbot to solve the problem.

Chatbot is a technology that support the role of C3 officers in responding to questions posed by students, especially questions that are general and frequently asked. Because chatbots generally provide a natural language interface, making it easier for users to interact with the chatbot. Chatbots can automatically serve user inquiries at any time and in huge numbers.

Chatbots can be part of Natural Language Processing (NLP), a system that can analyze text based on a set of theories and techniques. There are many fields such as information retrieval, information extraction, question and answer system [1].

One of the information extraction concepts from NLP is Named Entity Recognition (NER) which performs the information extraction process by marking or tagging each entity for unstructured data to be understood by computers. Christianto [2] has successfully applied NER in the learning and pattern recognition process and used Artificial Intelligence Markup Language (AIML) to provide relevant answers according to patterns or sentence keys commonly used in human language. While Morwal [3] has compared several methods, including CRF (Conditional Random Fields), MEMM (Maximum Entropy Markov Model), SVM

DOI: 10.24014/coreit.v8i2.19638

(Support Vector Machine), and HMM (Hidden Markov Model) and stated that the HMM method is better and gets better results.

NATURAL LANGUAGE PROCESSING

NLP originally started in the late 1940s when machine translation was first used to decrypt enemy codes. However, not much research on NLP was conducted until the 1980s. Many fields apply NLP technology, such as Information Retrieval, Information Extraction, and Question-Answering.

Named identity recognition (NER) is entity identification, chunking, and extraction. NER is an information extraction subtask that seeks to find and classify entities referred to in the text into predetermined categories such as names of people, organizations, locations, time expressions, amounts, monetary values, percentages, and so on [4].

AIML

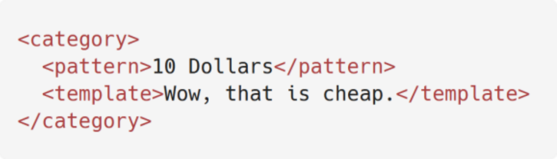
Artificial Intelligence Markup Language (AIML) is a format used to create documents about chatbot systems. AIML consists of data objects called AIML objects, which are units called topics and categories. A topic is a top-level element, having a name attribute and a set of categories associated with the topic.

Categories are the basic units of knowledge in AIML documents. Each category is a rule to match input and convert to output. This category consists of a pattern, which matches user input, and a template, which is used to generate chatbot answers. AIML was initially implemented by Wallace in ALICE [5].

ALICE is an artistic internet computer entity, which was first implemented. ALICE's knowledge of English conversation patterns is stored in AIML files.

In ALICE, there are three types of AIML categories [6], namely:

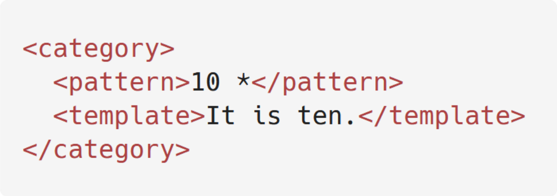
1. Atomic categories have patterns that do not have the '_' and '*' wildcard symbols.



```
<category>
  <pattern>10 Dollars</pattern>
  <template>Wow, that is cheap.</template>
</category>
```

Figure 1 Atomic categories pattern

   For example, in the above category, if the user enters '10 Dollars', then ALICE replies 'Wow, that is cheap.'.
2. Default categories are patterns that have the wildcard symbol '_' or '*'. Wildcard symbols match any input but differ in alphabetical order. Assuming the previous input is '10 Dollars', if the robot does not find the last category in atomic categories then, the robot tries to find a category with a default pattern like:



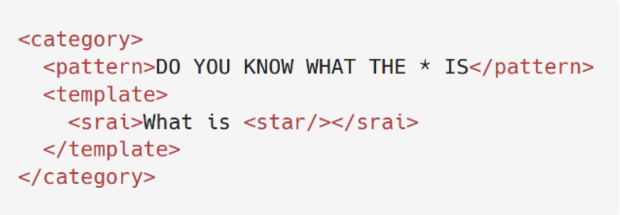
```
<category>
  <pattern>10 *</pattern>
  <template>It is ten.</template>
</category>
```

Figure 2 Default categories pattern

   So ALICE answered 'It is ten.'
3. Recursive categories : are templates that have the '<srai>' and '<sr>' tags, which refer to recursive subtraction rules. Recursive categories have many applications, such as symbolic subtraction, which reduces complex grammatical forms to simpler ones, sorting that divides the input into two or more subparts, and combining the responses for each. Furthermore, dealing with synonyms by

mapping out different ways of saying the same thing with the same answer. Examples of recursive categories are as follows:

a. Symbolic Reduction

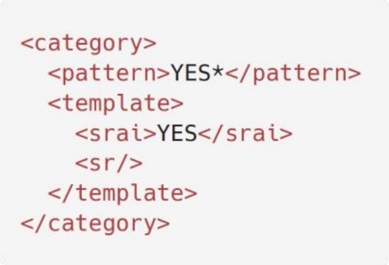


```
<category>
  <pattern>DO YOU KNOW WHAT THE * IS</pattern>
  <template>
    <srai>What is <star/></srai>
  </template>
</category>
```

Figure 3 Symbolic reduction pattern

This example <srai> reduces the input to the simpler "Apa itu *" form.

b. Divide and Conquer

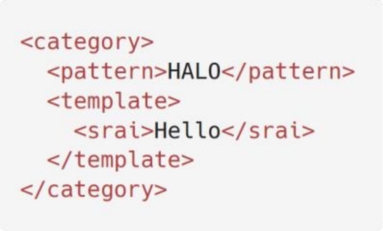


```
<category>
  <pattern>YES*</pattern>
  <template>
    <srai>YES</srai>
    <sr/>
  </template>
</category>
```

Figure 4 Divide and conquer pattern

The input is partitioned into two parts, "YES", and the second part '*' is matched with the <sr /> tag namely: <sr /> = <srai> <star /> </srai>.

c. Synonyms



```
<category>
  <pattern>HALO</pattern>
  <template>
    <srai>Hello</srai>
  </template>
</category>
```

Figure 5 Synonyms pattern

That is input another form, which has the same meaning.

The following is a pattern matching technique to produce the appropriate answer [2]:

1. Fetch categories based on the length of the input keyword.
2. Match patterns by category.
3. If the match percentage is 100%, then the process is complete.
4. Otherwise, match the longer one and shorter 1 (if any) pattern.
5. The highest match percentage will be used as the result.

Hidden Markov Model (HMM) is a generative model. This model assigns joint probabilities to pair observations and label sequences. Then the parameters are trained to maximize the possible combination of the training sets.

Definition of HMM formally According to [7] as follows:

1. = (A, B,). Here, A represents the transition probability. B represents emission probability and represents start probability.
2. A = aij = (Number of transitions from state to sj) / (Number of transitions from state si).
3. B = bj(k) = (Amount of time in state j and observed symbol k) / (expected amount of time in state j).

METHODOLOGY

This stage is done so that there is no confusion in understanding any problems that will arise in the future. These include:

1. Determining the research background
2. Determine the objectives, limitations, and scope of the research
3. Looking for information about literature and materials related to research.

Question collection data will be used as a source of knowledge on chatbots. It contains academic information at UIN Suska Riau and is made in conversation scenarios obtained from student academic guidebooks.

Table 1 Question and answer collection

Question	Answer
Apa itu iRaise? Apa kepanjangan iRaise?	iRaise adalah Integrated Academic Information System.

In addition to the question collection data, Indonesian language corpus data is needed that already has a tag for each word. This data is taken from previous research.

Named Entity Recognition (NER) is the central part of the entity classification process in chatbot analysis. In general, this stage helps learn the knowledge to tag each entity, which will help assist the AIML process in creating patterns for chatbots.

Named Entity Recognition (NER) has several stages, namely:

1. Text preprocessing
   Text preprocessing is a process for processing text data into data that is easier to process when the primary process is carried out. The steps for doing text preprocessing are:
   a. Case folding is changing all uppercase letters into lowercase form. All characters other than letters are discarded.
   b. Tokenization is the process of breaking sentences based on the constituent words.
2. Feature Extraction
   Feature extraction is the process of extracting attributes from training data or text data that has been processed.
3. Classifier
   The classifier is a model used to classify attributes that have been analyzed with specific methods based on each constituent word. This classifier process uses the Hidden Markov Model (HMM) method to assign entities to sentences.

Chatbot analysis is an overview of how the chatbot process responds:

1. Data input
   Input data or input is the key to determining the answer following the expected results. Good input data is data written in standard language and written in the correct order.
2. Preprocessing
   This process is carried out to change the input data from the user so that it is easy to process by the system. This stage is carried out like the text preprocessing process in NER analysis.
3. Entity Classification
   The Entity Classification is a process of providing entity information in a word or sentence will be carried out. At this stage, the NER process will be carried out. This process is helpful for the pattern recognition process (pattern learning).
4. AIML

AIML is the primary process of the chatbot system in responding to the user, providing responses, and using knowledge based as a basis for conducting conversations.

5. Answer extraction
   Answer extraction is the process used to extract the answers given to the user. The answer given to the user is done by returning the template tag from the appropriate category tag.

The functional analysis of the application to be built will use the Business Process Modeling and Notation (BPMN) diagram.

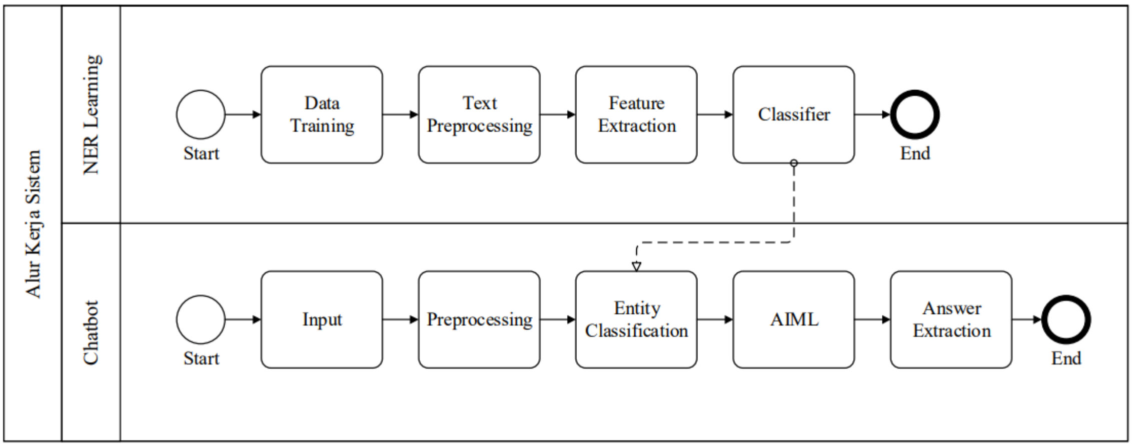


Figure 6 Chatbot app workflow

RESULT AND DISCUSSION

At the NER stage, an analysis of the tagging process or tagging of the question text or input data will be carried out based on the training data. This NER process will use calculations according to the Hidden Markov Model method. Sample data will be used to simplify the writing process. The data that will be used are as follows:

1. Question Data,
   Data collection of questions is input data that contains a collection of questions and answers. Every single question will be generated into 5 different question patterns. So the total data is 150 question patterns with 30 answer categories. The document collection data will be the initial knowledge-based data on the chatbot application.

```
Apa itu iRaise?
```

2. Corpus Data
   The Indonesian tagged corpus data is obtained from previous studies [8]. This data is 1 million Indonesian words that have been tagged which will be used as training data or training data. The samples used in making the HMM model are as follows:

```
apa/WP sistem/NN itu/DT
untuk/IN itu/DT iraise/NN
iraise/NN sistem/NN akademik/NN
apa/WP itu/DT iraise/NN
akademik/NN
```

The steps for conducting a NER analysis are as follows:

1. Text Preprocessing
   The text is processed before calculating for tagging the entity at this stage. The tokenization process is carried out on the training data. For tokenization, the testing data only deletes and does

not need attention to the sentence structure. The tokenization results for the training data can be seen in Table 2.

Table 2 Tokenization results

Question	Result
Apa itu iRaise?	apa itu iraise

2. Feature Extraction
   Performed to extract the features obtained from the training data at the NER stage, these features will be used as stated in the HMM calculation. To find states from the previous training data sample, it is as follows:
   Input : Annotated Tagged Corpus
   Output : State Vector or state direction
   States : [WP, NN, DT, IN]

3. Classifier
   This stage determines the category of states to get the observation values that we input. We will enter the observation value: "apa itu iraise". To find out the hidden states in the observation of the calculation model using the HMM method.

Analysis with Hidden Markov Model steps are as follows:

1. Initial Probability
   The first stage is to find the initial probability value from the training data.

Table 3 Initial probability matrix

WP	NN	DT	IN
0.5	0.25	0	0.25

2. Transition Probability Matrix
   To get the value of the transition probability will be calculated using the equation in the form of a matrix as follows:

$$(\text{A}) = \frac{total\, Ti\; to\; Tj}{total\, Ti}$$

3. Emission Probability Matrix
   To get the value of the emission probability will be calculated using the equation in the form of a matrix as follows:

$$(B) = \frac{total\; word\; occurrence\; as\; tag}{total\; tag\; occurrences}$$

Table 4 Emission probability matrix

	apa	sistem	itu	untuk	iraise	Akademik
WP	1	0	0	0	0	0
NN	0	0.28	0	0	0.42	0.28
DT	0	0	1	0	0	0
IN	0	0	0	1	0	0

From the above calculation to determine the value of the observation "what is iraise", then which has the highest value probability approach with the output ('WP','DT','NN').

The chatbot analysis stages will give an idea of how this chatbot can run:

1. Data Input
   The input data is the question data that the user wants to ask the chatbot system. The data that can be entered is data in free text. For example, the user enters the question: "Apa itu iRaise?".

2. Preprocessing
   After the user enters the question he wants to ask, the system will take the question and carry out the preprocessing process. In general, this process is the same as in the text preprocessing process in NER analysis, but case-folding will also be carried out for questions from users.
   The following is the result of the process at the preprocessing stage for questions on the chatbot.

3. Entity Classification
   The process of providing entity information in a word or sentence will be carried out. The NER process will be carried out. This process is helpful for the pattern recognition process (pattern learning).

Table 5 Entity classification

Word	Tag
apa	WP
itu	DT
iraise	NN

4. AIML
   AIML is the primary process of the chatbot system in responding to the user, providing responses, and using knowledge based as a basis for conducting conversations.
   A pattern recognition process or pattern matching is carried out to provide answers to the user's question "apa itu iraise".
   The stages of the pattern matching process from the data above are as follows:
   a. Fetch categories based on the length of the keyword found in the input. Based on the input data above, the length of the keyword "apa itu iraise", what was found was 3. Table 5 questions with the same keyword length were questions number 1 and 2.
   b. Perform pattern matching based on the categories that have been taken. The pattern of user input data is "WP, DT, NN" from the data knowledge base pattern for question 1 is "WP, DT, NN" and the pattern for question 2 is "WP, NN, VBI". From these data, it can be calculated that the percentage value of the match pattern for question 1 is 100%, and the second question is 33.33%.
   c. If the match percentage shows 100%, the pattern matching process is complete and returns the category. Where that shows a value of 100% is question 1.
   d. For percentage results that are not equal to 100%, the process will continue by recalculating the pattern matching process using a longer and shorter pattern than the pattern found on the input.
   e. Based on checking the pattern with the same length, the longer pattern one, and the shorter one being 1, the pattern matching process will return the category with the most significant percentage match value provided that the match value is more than 75%.

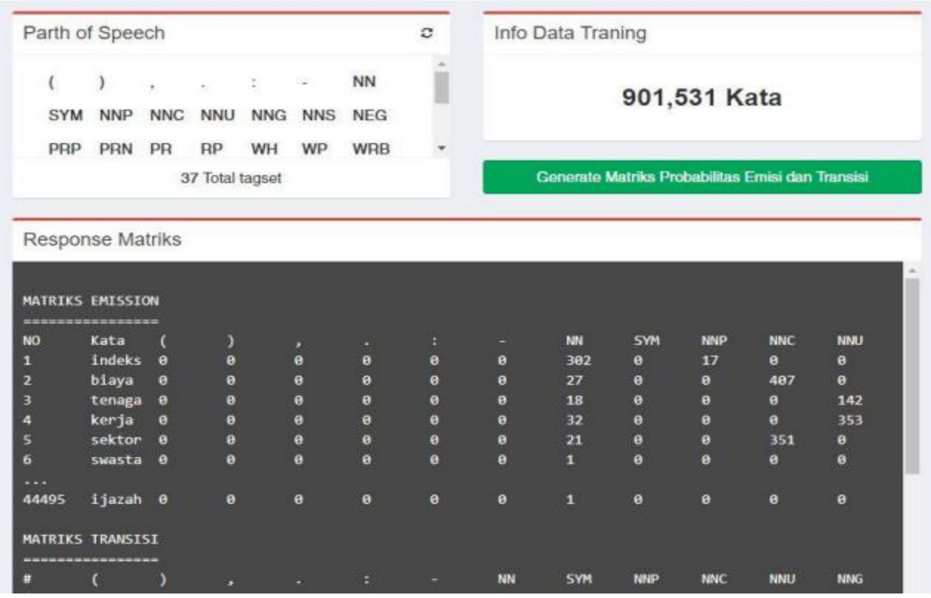
5. Answer Extraction
   Answer Extraction is the process used to extract the answers provided to the user. The answer given to the user is done by returning the template tag from the appropriate category tag. Answer extraction is a process that will extract and provide answers to users. In general, this process will read the tag template from the tag category obtained from the previous process.

Table 6 Answer extraction

Category	Template
Apa itu iRaise?	iRaise adalah Integrated Academic Information System

Application

The Generate Model page is the index page or the first page displayed when this application is run. As expected at the design stage, the system can display all existing tags, information on the amount of training data, generate a matrix button to calculate the matrix, and the response of the matrix calculation results.


Parth of Speech

( ) , . : - NN
SYM NNP NNC NNU NNG NNS NEG
PRP PRN PR RP WH WP WRB

37 Total tagset

Info Data Traning

901,531 Kata

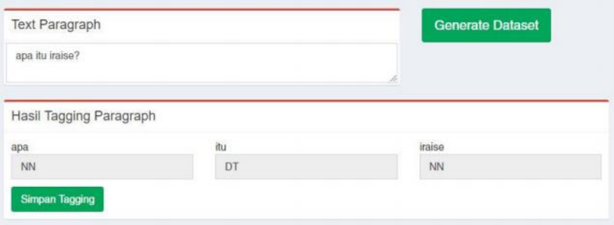
Response Matriks

```
MATRIKS EMISSION
================
NO     Kata    (     )     ,     .     :     -     NN    SYM   NNP   NNC   NNU
1      indeks  0     0     0     0     0     0     302   0     17    0     0
2      biaya   0     0     0     0     0     0     27    0     0     407   0
3      tenaga  0     0     0     0     0     0     18    0     0     0     142
4      kerja   0     0     0     0     0     0     32    0     0     0     353
5      sektor  0     0     0     0     0     0     21    0     0     351   0
6      swasta  0     0     0     0     0     0     1     0     0     0     0
...
44495  ijazah  0     0     0     0     0     0     1     0     0     0     0

MATRIKS TRANSISI
================
#      (     )     ,     .     :     -     NN    SYM   NNP   NNC   NNU   NNG
```

Figure 7 Generated model

On the Import Dataset page, the user can enter paragraph text which is then carried out by the dataset generation process. Then the system can display results in the form of words that have been tagged or not yet available in the dataset; users can also select a tagset for words that have not been registered in the dataset and then save it.


Text Paragraph

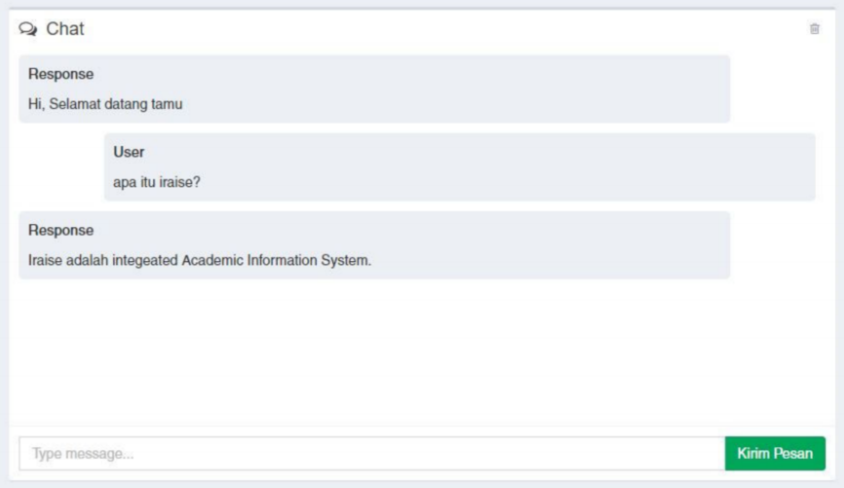
apa itu iraise?

Hasil Tagging Paragraph

apa: NN
itu: DT
iraise: NN

Figure 8 Import dataset

The application can display a form to enter questions from users, tag each word, enter a template or answer to the question and then save it. The chatbot page is the core page where users can communicate with conversation robots or chatbots. The chatbot will respond to the question when the user enters a question.


Chat

Response
Hi, Selamat datang tamu

User
apa itu iraise?

Response
Iraise adalah integeated Academic Information System.

Figure 9 Chatbot interface

From the black box testing that has been carried out, it can be concluded that the information service chatbot application with AIML and NER was successfully executed following the purpose of application analysis and design.

From test results, it can be seen that the system answered questions as expected, amounting to 83 of 150 questions, and for incorrect or inappropriate answers, that were expected to be 67 out of 150. The percentage accuracy value for correct answers was 55% from this amount.

The system's low accuracy of the answers is caused by several factors, namely, irrelevant answers caused by the number of sentences being tagged too few so that many questions have the same pattern.

The algorithm in the program directly returns the answer if it finds the same pattern or pattern without comparing it with other categories. The questions that were not recognized were caused by the absence of a pattern that matched the questions given. It can be increased by increasing the number of question categories.

CONCLUSION

The conclusions that researchers can draw are as follows:

1. The information service chatbot system using Artificial Intelligence Markup Language and Named Entity Recognition was successfully built according to the analysis and design stages.
2. The percentage results obtained from the tests carried out on the questions in the knowledge-based have an accuracy value of 55%.
3. Based on the percentage of accuracy from the test, this system can be applied as a source of information for users with continuous optimization.

REFERENCES

[1] E. D. Liddy, “Natural Language Processing. In Encyclopedia of Library and Information Science,” Encyclopedia of Library and Information Science. 2001.

[2] D. Christianto, E. Siswanto, and R. Chaniago, “Penggunaan Named Entity Recognition dan Artificial Intelligence Markup Language untuk Penerapan Chatbot Berbasis Teks,” J. Telemat., 2016.

[3] S. Morwal, “Named Entity Recognition using Hidden Markov Model (HMM),” Int. J. Nat. Lang. Comput., 2012.

[4] E. Marsh and D. Perzanowski, “MUC-7 Evaluation of IE Technology: Overview of Results,” in Proceedings of the Seventh Message Understanding Conference (MUC-7), 1998.

[5] H. Henderson, “Artificial intelligence : mirrors for the mind.” Chelsea House, New York NY, 2007.

[6] B. Abu Shawar and E. Atwell, “Chatbots: are they really useful?,” LDV-Forum Zeitschrift für Comput. und Sprachtechnologie, 2007.

[7] L. R. Rabiner, “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,” Proc. IEEE, 1989.

[8] A. F. Wicaksono and P. Ayu, “HMM Based Part-of-Speech Tagger for Bahasa Indonesia,” in Prooceedings of 4th International MALINDO (Malay and Indonesian Language) Workshop, 2010.