NAMED ENTITY RECOGNITION in Indonesian Text for Automatic Question Generator

Rusliani[1], Ken Kinanti Purnamasari[2]
[1,2] Universitas Komputer Indonesia
Jl. Dipati Ukur 102, 112, 114, 116 Bandung
E-mail : liaein.ein@gmail.com[1], guinev_gs@yahoo.co.nz[2]

ABSTRACT

Question liquid plant discussed the task automatically arouse questions from various kinds of input as text, database, or flow semantics. In previous research, automatic question generation using templates based on syntactic analysis tlah successfully done for sentence ask "who", "what", "where", and "the Chart". This research resulted in a still-low accuracy applies of 14,53%. This is due to the detection of the existence of a name, yet the ambiguity that has not been resolved, the dictionary of bias is used in is still not complete, and the search function syntax that is not yet perfect. In the problem, it can make on this research use the NER for it can make detecting with the method name HMM. The process that consists of preprocessing is done; tokenisasi tokenisasi words, sentences, and post tags. Then the data are trained in methods of HMM. In this study, the configuration is done there are 3 (three), apply the configuration class word against the word, the word against Word class and Word, and the word against Word class, words before and after it. Based on the test results using the data as much as 897 tag entity, obtained the result value of the recall of 0.0043, precision of 0.1249, and f-a measure of relevance, and accuracy 0.0083 labeling 12.49% of words and the relevance of the question of 19,71%.

Keywords : *named entity recogniton* (NER), *hidden markov model* (HMM), *part-of-speech* (POS) tag, *automatic question generator.*

1. INTRODUTCION

Question generator is a task that automatically evokes questions from a wide variety of input as text, database, or flow semantics [1]. A lot of research that's been done before on generating questions with different methods either in Indonesia or text speak in United Kingdom.

In previous research, automatic question generation using templates based on syntactic analysis has been successfully done to sentence ask "who", "what", "where" and "when". This research resulted in a still-low accuracy namely of 14,53%. It is caused due to the detection of the existence of a name, yet the ambiguity that still has not been resolved, the dictionary used is still not complete, and the search function a rudimentary syntax [2]. In this study, Named Entity Recogniton used to resolve the name of the detection on the application of the automatic question generator.

Named Entity Recognition (NER) is the principal component in many applications of NLP (Natural Language processing) such as question answering, information extraction, detection, clustering of topics, and compaction [3][4][5]. NER can be used to identify and detect entities such as people's names, locations, organizations, time, and others [1][4][5]. Much research that is already using the NER in a variety of languages including bahasa Indonesia [6][7], Turkey [2], Malaysia [3], Arabic [8], Persian [9], India [1], and Korea [10] with a different method.

Algorithms for NER are classified into 3 categories; rule-based, machine learning, and hybrid [4]. In this study, the algorithm that is used is one of machine learning i.e. Hidden Markov Model (HMM). This is because HMM is easy to understand, implement, and analyzed by the system when compared to rule-based and maximum entropy [1].

Research previously conducted NER, i.e. research conducted by Muhammad Dartheritage about NER speak Indonesia by using method of HMM produce accuracy of 88.4% [34], the value indicates that NER with the proven reliability of HMM method. Therefore, in this study will be used for generating NER questions automatically.

Based on the description on this research will be done on implementation of Named Entity Recognition method using Indonesia-language Text Hidden Markov models for Automatic Question Generator.

2. CONTENTS OF REEARCH

The input data used in this study was divided into two, namely data for training and testing in the form of data files with the *.txt format. The expected input is text speak in Indonesia which has a grammar that is raw. The following is the form of input data are used, sourced from the University of Gajah Mada.

Fig. 1. Data Input

Pengamat politik dari Universitas Gadjah Mada, Arie Sudjito, menilai, keinginan Ketua Umum Partai Golkar Aburizal Bakrie untuk maju kembali sebagai ketua umum merupakan pemaksaan kehendak. Menurut dia, ada kesan bahwa Aburizal menggunakan segala cara untuk memuluskan jalannya kembali menduduki Golkar 1. Hal ini, kata Arie, berpotensi menimbukan perpecahan di kalangan kader Golkar di daerah. "Tentu ini akan merugikan Golkar karena Golkar akan pecah dan konflik di daerah dipastikan akan terjadi," ujar Arie, saat dihubungi Kompas.com, Selasa (2/12/2014). Arie mengatakan, sikap Aburizal tersebut ditentang oleh sebagian kalangan di internal Golkar yang melakukan perlawanan dengan

2.1. Process

NER systems stage is divided into two stages, namely the stage of training and testing. The training phase begins with users uploading data to the trainer of text corpus contains NER speak Indonesia. Then the text will be processed, the process is called tokenisasi. Tokenisasi does is tokenisasi against the sentences and words. After going through the process of tokenisasi, the Word will be classified according to its type (verbs, nouns, prepositions, etc.), this stage is called the POS tags. After it goes into the stage of modeling by using Hidden Markov Model in which there is a probability calculation of initial operation, the probability of a transition, and the probability of emission. After the operation the calculation is complete, the viterbi algorithm then went into to find conclusions based on modeling has been done before. At this stage of testing, data entered in the form of test data. Test data is the corpus of NER speak Indonesia. Training data, test data whatsoever through the stage's just preprocessing the data will be tested with the modeling that was done on the training process. After that, the results of the testing will be processed again to stage the stirring questions so generate output in the form of questions and answers.

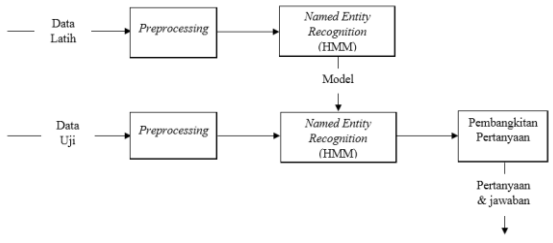


Fig. 2. Alur Proses

2.2. Preprocessing

Preprocessing in this research consists of input data, sentences tokenizing, word tokenizing, and post tags.

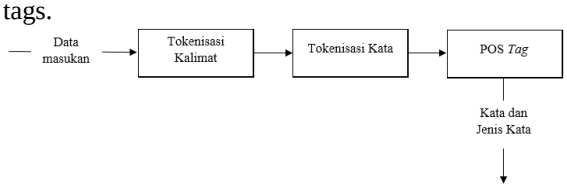


Fig. 3. *Preprocessing*

2.3. Tokenizing of Sentences

Text to be input in the form of a collection of a few sentences, so that the required tokenisasi the sentence to separate the sentences collection. Tokenisasi sentence is carried out by means of separate sentences with a period (.) as a delimeter. Here's a tokenisasi sentence of data input.

Table 1. Tokenizing of Sentences

Num.	Sentences
1	Pengamat politik dari Universitas Gadjah Mada, Arie Sudjito, menilai, keinginan Ketua Umum Partai Golkar Aburizal Bakrie untuk maju kembali sebagai ketua umum merupakan pemaksaan kehendak
2	Menurut dia, ada kesan bahwa Aburizal menggunakan segala cara untuk memuluskan jalannya kembali menduduki Golkar 1
3	Hal ini, kata Arie, berpotensi menimbukan perpecahan di kalangan kader Golkar di daerah.
4	"Tentu ini akan merugikan Golkar karena Golkar akan pecah dan konflik di daerah dipastikan akan terjadi," ujar Arie, saat dihubungi Kompas.com, Selasa (2/12/2014)
5	Arie mengatakan, sikap Aburizal tersebut ditentang oleh sebagian kalangan di internal Golkar yang melakukan perlawanan dengan membentuk Presidium Penyelamat Partai Golkar

2.4. Tokenizing of Words

After all sentences have been tokenized, sentences needs to tokenized in words.

Tabel 2. Tokenisasi Kata

No.	Kata		
1	Pengamat	keinginan	kembali
	politik	Ketua	sebagai
	dari	Umum	ketua
	Universitas	Partai	umum
	Gadjah	Golkar	merupakan
	Mada	Aburizal	pemaksaan
	Arie	Bakrie	kehendak
	Sutedjo	untuk	
	menilai	maju	
2	Menurut	menggunakan	kembali
	dia	segala	menduduki
	ada	cara	Golkar
	kesan	untuk	1
	bahwa	memuluskan	
	Aburizal	jalannya	
3	Hal	menimbulkan	Golkar
	ini	perpecahan	di
	kata	di	daerah
	Arie	kalangan	
	berpotensi	kader	
4	Tentu	pecah	ujar
	ini	dan	Arie
	akan	konflik	saat
	merugikan	di	dihubungi
	Golkar	daerah	Kompas.com
	karena	dipastikan	Selasa
	Golkar	akan	2/12/2014
	akan	terjadi	
5	Arie	sebagian	perlawanan
	mengatakan	kalangan	dengan
	sikap	di	membentuk
	Aburizal	internal	Presidium
	tersebut	Golkar	Penyelamat
	ditentang	yang	Partai
	oleh	melakukan	Golkar

2.5. Part-of-Speech Tagging

POS Tagging is classification kind words. POS Tagging system used on this is the POS Tagging of Alfan [11]. So is generated as follows.

1.

Pengamat	politik	dari	Universitas	Gadjah
NN	**NN**	**IN**	**NN**	**NN**
Mada	Arie	Sudjito	menilai	keinginan
NNP	**NNP**	**NNP**	**VBT**	**NN**
Ketua	Umum	Partai	Golkar	Aburizal
NN	**JJ**	**NN**	**NN**	**NN**
Bakrie	untuk	maju	kembali	sebagai
NNP	**IN**	**VBI**	**RB**	**IN**
ketua	umum	merupakan	pemaksaan	kehendak
NN	**JJ**	**VBT**	**NNP**	**NN**

2.

Menurut	dia	ada	kesan	bahwa
VBT	**PRP**	**VBI**	**NN**	**SC**
Aburizal	menggunakan	segala	cara	untuk
NN	**VBT**	**NN**	**NN**	**IN**
jalannya	kembali	menduduki	Golkar	1
NNG	**RB**	**VBT**	**NN**	**CDP**

3.

Hal	ini	kata	Arie	berpotensi
NN	**DT**	**VBT**	**NNP**	**NN**

4.

menimbulkan	perpecahan	di	kalangan	kader
VBT	**NN**	**IN**	**NN**	**NN**
Golkar	di	Daerah		
NN	**IN**	**NN**		

Tentu	ini	akan	merugikan	Golkar
NN	**DT**	**MD**	**VBT**	**NN**
karena	Golkar	akan	pecah	dan
CC	**NN**	**MD**	**VBT**	**CC**
di	daerah	dipastikan	akan	terjadi
IN	**NN**	**VBT**	**MD**	**VBI**
ujar	Arie	saat	dihubungi	Kompas.com
VBT	**NNP**	**SC**	**VBT**	**NNP**
Selasa	2/12/2014			

5

Arie	mengatakan	sikap	Aburizal	tersebut
NNP	**VBI**	**NN**	**NN**	**DT**
ditentang	oleh	sebagian	kalangan	di
NN	**IN**	**CDI**	**NN**	**IN**
Golkar	yang	melakukan	perlawanan	dengan
NN	**SC**	**VBT**	**NN**	**IN**
membentuk	Presidium	Penyelamat	Partai	Golkar
VBT	**NN**	**NN**	**NN**	**NN**

Fig. 4. *POS Tag*

2.6. Training

The input data is processed so that every word in the text is marked with the appropriate entities. This process is done manually [7]. The following is an example of the tagging entity is used.

1.

Pengamat	politik	dari	Universitas	Gadjah
OTHER	**OTHER**	**OTHER**	**ORG**	**ORG**
Mada	Arie	Sudjito	menilai	keinginan
ORG	**PER**	**PER**	**OTHER**	**OTHER**
Umum	Partai	Golkar	Aburizal	Bakrie
OTHER	**ORG**	**ORG**	**PER**	**PER**
untuk	maju	kembali	sebagai	ketua
OTHER	**OTHER**	**OTHER**	**OTHER**	**OTHER**
umum	merupakan	pemaksaan	kehendak	
OTHER	**OTHER**	**OTHER**	**OTHER**	

2.

Menurut	dia	ada	kesan	bahwa
OTHER	**OTHER**	**OTHER**	**OTHER**	**OTHER**
Aburizal	menggunakan	segala	cara	untuk
PER	**OTHER**	**OTHER**	**OTHER**	**OTHER**
memuluskan	jalannya	kembali	menduduki	Golkar
OTHER	**OTHER**	**OTHER**	**OTHER**	**ORG**

1

	OTHER				
3.	Hal	ini	kata	Arie	berpotensi
	OTHER	**OTHER**	**OTHER**	**PER**	**OTHER**
	menimbulkan	perpecahan	di	kalangan	kader
	OTHER	**OTHER**	**OTHER**	**OTHER**	**OTHER**
	Golkar	di	daerah		
	ORG	**OTHER**	**OTHER**		
4.	Tentu	ini	akan	merugikan	Golkar
	OTHER	**OTHER**	**OTHER**	**OTHER**	**ORG**
	karena	Golkar	akan	pecah	dan
	OTHER	**ORG**	**OTHER**	**OTHER**	**OTHER**
	konflik	di	daerah	dipastikan	akan
	OTHER	**OTHER**	**OTHER**	**OTHER**	**OTHER**
	terjadi	ujar	Arie	saat	dihubungi
	OTHER	**OTHER**	**PER**	**OTHER**	**OTHER**
	Kompas.com	Selasa	2/12/2014		
	OTHER	**TIME**	**TIME**		
5	Arie	mengatakan	sikap	Aburizal	tersebut
	PER	**OTHER**	**OTHER**	**PER**	**OTHER**
	ditentang	oleh	sebagian	kalangan	di
	OTHER	**OTHER**	**OTHER**	**OTHER**	**OTHER**
	internal	Golkar	yang	melakukan	perlawanan
	OTHER	**ORG**	**OTHER**	**OTHER**	**OTHER**
	dengan	membentuk	Presidium	Penyelamat	Partai
	OTHER	**OTHER**	**ORG**	**ORG**	**ORG**
	Golkar				
	ORG				

Fig.5. Tagged Entity

After the data is labeled manually [34], the next entry to the training process. The training process in the system is used to get the value of the initial probability, probability matrix of transition probability matrix and emissions. Based on the research of Sudha Morwal, Nusrat Jahan, and Deepti Chopra [1] the initial probability can use equation (1).

(1)

$$\pi = \frac{tag\ entitas\ di\ awal\ kalimat}{jumlah\ kalimat}$$

Tabel 3. Start Probability

PER	LOC	ORG	TIME	QTY	OTHER
1/5= 0.2	0	0	0	0	4/5= 0.8

After getting the initial probability, is seeking a transition probability based on the research of Sudha Morwal, Nusrat Jahan, and Deepti Chopra [1] can use the equation (2).

(2)

$$A = \frac{jumlah\ urutan\ tag\ dari\ t_i\ ke\ t_j}{jumlah\ tag\ t_i}$$

Table 4. Transition Probability

	PER	LOC	ORG	TIME	QTY	OTHER
PER	0.2222	0	0	0	0	0.7778
LOC	0	0	0	0	0	0
ORG	0.1538	0	0.4615	0	0	0.3846
TIME	0.5	0	0	0.5	0	0
QTY	0	0	0	0	0	0
OTHER	0.0548	0	0.1096	0.0137	0	0.8082

After the initial probability and the probability of transition is complete, the next step that is looking for research based on the emission probability from Sudha Morwal, Nusrat Jahan, and Deepti Chopra [1] can use the equation (3).

(3)

$$B = \frac{kelas\ kata\ dari\ tag\ entitas}{jumlah\ tag\ entitas}$$

In this study there are three grade sequence of configuration Word will do. The configuration is as follows.

1. the word on Word Classes
2. class said on those words and in previously said
3. class said on the word, a word before, and afterwards

Tabel 5. Emition Probability Configuration 1

	PER	LOC	ORG	TIME	QTY	OTHER
NNP	0.6667	0	0.0769	0	0	0.0274
VBI	0	0	0	0	0	0.0137
NN	0.3333	0	1	1	0	0.0548
DT	0	0	0	0	0	0.0137
IN	0	0	0	0	0	0.0274
CDI	0	0	0	0	0	0.0137
JJ	0	0	0	0	0	0.0137
SC	0	0	0	0	0	0.0137
VBT	0	0	0	0	0	0.0137

2.7. Decoding

Testing on this system using the Viterbi algorithm, by equation (4), (5), (6) and (7).

1. Inisialisation

$$\delta(i) = \pi_i b_t(O_1) \quad (4)$$

$$\varphi_1(i) = 0$$

2. Rekurtion

$$\delta_t(j) = \max_{1 \le 1 \le N} \left[\delta_{t-1(i)} a_{ij}\right] b_j(O_t) \quad (5)$$

$$\psi_t(j) = \operatorname{argmax}_{1 \le 1 \le N} \left[\delta_{t-1(i)} a_{ij}\right]$$

3. Termination

$$P^* = \max_{1 \le I \le N} \left[\delta_{T(i)}\right] \quad (6)$$

$$Q^*T = \operatorname{argmax}_{1 \le I \le N} [\delta_{T(i)}]$$

4. Backtracking

$$Q_t^* = \varphi_{t+1}(Q_{t+1}^*) \quad (7)$$

Of those processes, it can be inferred that the state is possible with the greatest opportunities for observation {*Arie, mengatakan, sikap, Aburizal, tersebut, ditentang, oleh, sebagian, kalangan, di, internal, Golkar, yang, melakukan, perlawanan, dengan, membentuk, Presidium, Penyelamat, Partai, Golkar*} is {*PERSON, OTHER*}.

2.8. Question Generator

After each word entity tag is detected, the next process is to find a syntactic functions such as subject, predicate, object, and description. Question will be raised using sentence templates based on syntactic functions that have been found. The process of generation of questions in this research is divided into two processes namely syntactic analysis to find the function syntax and generation questions to generate questions based on template syntax.

In previous studies, the search function syntax is done by dividing the clauses that are predicative. In this study, the search function syntax is done based on the POS tags that have been acquired earlier. So to find the clauses that are predicative, simply by finding kind words (verbs).

Table 6. Question Generator

Sentences	Question Generatiion
Arie mengatakan, sikap Aburizal tersebut ditentang oleh sebagian kalangan di internal Golkar yang melakukan perlawanan dengan membentuk Presidium Penyelamat Partai Golkar	1. Siapa yang mengatakan sikap Aburizal tersebut ditentang oleh sebagian kalangan di internal Golkar? 2. Apa yang ditentang oleh sebagian kalangan di internal Golkar? 3. Apa yang sebagian kalangan lakukan pada Aburizal? 4. Apa yang dibentuk oleh sebagian kalangan di internal Golkar?

3. Results

Performance testing was done by calculating the recall, precision, and f-measure in the equation (8), (9), (10).

(8)

$$Recall = \frac{Correct \ + \ 0.5 * Partial}{Possible}$$

(9)

$$Precision = \frac{Correct \ + \ 0.5 * Partial}{Actual}$$

(10)

$$F - measure = \frac{Recall * Precision}{0.5 * (Recall + Precision)}$$

Table 7. Pengujian Perfomansi

Recall	Precision	F-measure
0.0043	0.1249	0.0083

The resulting accuracy of data being tested as many as 28992 tag with 12333 relevant tag. Total questions that can be generated by the system is as much as 137 questions with 27 of them are relevant questions in accordance with the text input.

Tabel 8. Pengujian Akurasi

	Total	**Relevan**	**Akurasi**
Tag	28992	12333	**42,54%**
Pertanyaan	137	27	**19,71%**

4. CONCLUSION

Based on the results of research that has been done, it can be noted that the automatic question generator using NER better compared to automatic question generator without using NER, generating accuracy 19.71% of accuracy previously i.e. 14,53%. Named Entity Recogniton already able to help increase accuracy in automatic question generator. Suggestions for further research is the added ambiguity on NER so it can reduce the words are

ambiguous. A combination of various methods can also be done to find out which method is most appropriate for the NER.

REFERENCES

[1] S. Morwal, N. Jahan, & D. Chopra, "Named Entity Recognition using Hidden Markov Model (HMM)", *International Journal on Natural Language Computing (IJNLC)*, Vol. 1 No. 4, 2012.

[2] M. Raihan. I, "Pembangkit Pertanyaan Otomatis dari Teks Berbahasa Indonesia Berdasarkan Template Sintaksis", Bandung, 2016.

[3] S. K. Metin, T. Kisla. T, & B. Karaoglan, "Named Entity Recognition in Turkish Using Association Measures", *Advanced Computing: An International Journal (ACIJ)*, Vol. 3 No. 4, 2012.

[4] R. Alfred, L. Leong, On. C. K., & P. Anthony, "Malay Named Entity Recognition Based on Rule-Based Approach", *International Journal of Machine Learning and Computing*, Vol. 4 No. 3, 2014.

[5] Handi, "Text Pre-Processing pada Text To Speech Synthesis System untuk Penutur Berbahasa Indonesia", ITS Surabaya, 2011.

[6] I. Budi, S. Bressan, G. Wahyudi, Z. Hasibuan, & B. Nazief A, "Named Entity Recognition for the Indonesian Language: Combining Contextual, Morphological and Part-of-Speech Features into a Knowledge Engineering Approach", 2005.

[7] M. Fachri, "Pengenalan Entitas Bernama Pada Teks Bahasa Indonesia Menggunakan Hidden Markov Model", Yogyakarta, 2014.

[8] Y. Benajiba & P. Rosso, "Arabic Named Entity Recognition Using Conditional Random Fields".

[9] M. K. Khormuji, & M. Bazrafkan, "Persian Named Entity Recognition based with Local Filters", *International Journal of Computer Applications* (0975 8887), Vol. 100 No.4, 2014.

[10] C. N. Seon, Ko. Y., Kim. J. S, & J. Seo, "Named Entity Recognition using Machine Learning Methods and Pattern-Selection Rules".

[11] A. F. Wicaksono & A. Purwarianti, "HMM Based Part-of-Speech Tagger for Bahasa Indonesia", *Fourth Int. MALINDO Workshop*, 2010.