TMLAI TRANSACTIONS ON MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE
SOCIETY FOR SCIENCE AND EDUCATION UNITED KINGDOM
Volume 7 No. 1
ISSN 2054-7390

Named Entity Recognition for Characteristic of Medical Herbs Using Modified HMM Approach

Lailil Muflikhah, Agung Setiyono, Nurul Hidayat
Faculty of Computer Science; Brawijaya University; Malang, Indonesia
lailil@ub.ac.id; setiyoajiagung@gmail.com; ntayadih@ub.ac.id

ABSTRACT

The amount of articles in medicinal herbs is very huge. It is performed with unstructured format so that it takes time to get information as reader's need. Therefore, this research purposes to recognize the name entity of article from internet in order to increase information retrieval or other analysis data purposes. Named entity recognition is one of the goals of information extraction which is to identify the name and characteristics of the herbs. This paper is propose the modified method of Hidden Marcov Model (HMM) with Viterbi algorithm. In this method, it is enclosed gazetteer list for labeling name and location of data training to construct HMM. The data sets are taken from three web sites including: miliaton, aliweb, and plants. As a result, the performance is achieved at average precision value of 0.93, recall of 0.83 and f-measure of 0.85.

Keywords: Hidden Marcov Model, Gazetteer, Viterbi, Named entity, Medicinal herbs

1 Introduction

The number of websites about health is more and more increase various information including about health. Recently, medicinal herbs is main issue on medical analysis. Many people assume that there is no side effect to consume as medicine. Website is a site to share information through internet. The growth of information in internet is very huge, then it need to classify by information extraction based on herbs' characteristic.

Named Entity Recognition (NER) is a subtask of information extraction to identify and classify named entities in document text into pre-defined user's categories such as name of plant, location, characteristic, and advantages. In general, there are two methods of NER, i.e. rule based approach and machine learning based approach. The first method is need information from domain expert to construct many patterns. However, the second method, the pattern is generated from training data. Therefore, this research is proposed to combine the both methods.

The related research of NER is conducted by Todorovic, et al. that Hidden Marcov Model (HMM) in English text disregard to the grammar is achieved at accuracy of 91,71% [1]. HMM is a machine learning method with sequence probability model to solve the problem. It is need a set of training data for labeling words as POS-Tagging. The other research is conducted by Alfred et al., [2]. It is about named entity recognition for Malay article using rule based approach, but the accuracy is achieved of 89.47%. Furthermore, the NER

DOI: 10.14738/tmlai.71.6086
Publication Date: 23rd February, 2019
URL: http://dx.doi.org/10.14738/tmlai.71.6086

is applied to health article using Support Vector Machine, a machine learning method, and the accuracy rate is 90% [3].

2 Research Methods

This research is applied to the modified HMM with Viterbi algorithm for identifying named entity in news article of health. The method is involved rule based approach. Generally, the steps of named entity recognition system is shown in Figure 1, the block diagram. The first step is preprocessing data, including tokenizing only. In this steps, it is also put labelling for named entity of training set. The rule method is address to support the HMM construction. It is applied to find the pattern of word sequence, especially for ambiguity of term. Then, the next step is to construct the HMM model of term sequence. And as a result is the named entity of herb.

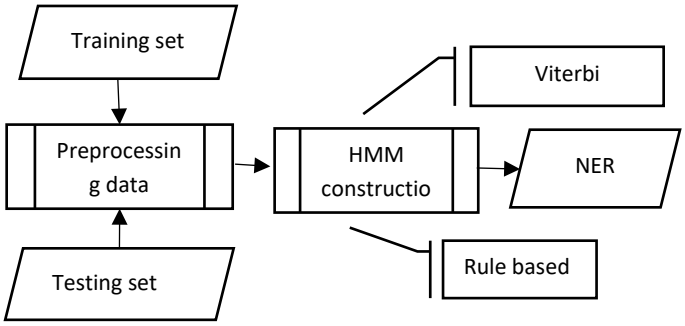

Figure 1. Block diagram of Named Entity Recognition

2.1 Text Mining

Text document has unstructured format which need to be transformed in structured format in order to get useful meaning of information and it is known as text mining. Information extraction is process to get information from document text [4]. This method is applied after preprocessing step, including: tokenizing, filtering, and stemming. It is address to index the term and make easy to create the HMM model and constructing the gazetteer in rule based.

2.2 Preprocessing Data

Preprocessing is a preliminary step which involves tokenization, stop words removal, stemming and misspellings words normalization. It involves tokenization, stop words removal, stemming and normalization including query expansion. Tokenization is a process removing punctuation, numbers, and characters other than the alphabet [5]. It is also conducted case folding, which is changing all capital letters into lowercase. Then, stop words removal or filtering is removing uninformative words referring to the existing stop word dictionary. Meanwhile, stemming is a process to convert every words to its root. This process is done by removing affixes such as prefix, infix and suffix.

2.3 Named Entity Recognition (NER)

Named entity recognition is information extraction derivation which is identify the type of word (entity) in document. In this research, it is used to identify name, substance and characteristics of herbal plants. The identification is proposed to organize the document in order to make easy for information retrieval. NER is an important tool in almost all of the Natural Language processing applications such as Information Retrieval (IR), Information Extraction (IE), Question Answering (QA), Machine Translation (MT) and

Automatic Summarization (AS) etc. NER can be defined as a two stage problem:- Identification of Proper Noun and classification of the Proper Noun into a set of classes such as Plant names, Location names, Substances (chemical). These words are collectively defined as "ENAMEX" by MUC-6[6]. Thus NER can be said as the process of identifying and classifying the tokens into the above predefined classes.

Basically, there are two approaches of NER, i.e. rule based approach and machine learning based approach. The first approach is concerned to manual rules or regular expression. Many rules based NER contains Lexicalized Grammar, Gazetteer list, and List of triggered words. Then, the machine learning based approach is concerned to pre-defined method, such as: Hidden Marcov Models (HMM), Decision Trees, Maximum Entropy (ME), SVM and CRF. Therefore, this research is proposed combined the both approach. The gazetteer list is used to support the HMM method.

2.4 Hidden Marcov Model (HMM)

Hidden Marcov Model (HMM) is one of NER approaches which based on statistical method and is depend on the sequence of word. Therefore, this research is developed and applied to statistical approach and combined to rule based using gazetteer list. HMM is a statistical method that involves the sequence probability of term-document. The method is a machine learning method using sequence probability approach to solve the problem. Each word is put pre-labelling of POS-Tagging to construct markov model [1] as in Equation 1.

$$\lambda = (A, B, \pi) \quad (1)$$

where: λ is marcov model, B is emission probability, and π is initial probability. There are five tuples in modelling HMM, including observed state (O), hidden state (Q), transition probability matrix (A), emission probability matrix (B) and initial state probability (π) [7].

1. Observed state (O)

   In the observed state, it is made with the symbols O = O_1, O_2, O_3, ... O_n observed state, the observed modeling is observed.

2. Hidden state (Q)

   Hidden State is a state that is hidden and cannot be observed, symbolized by Q = Q_1, Q_2, Q_3, ... Q_n.

3. Transition Probability Matrix (*A*)

   Transition probability is an probability to move from state i to state j. It is symbolized by A = a_{01}, a_{02}, a_{n1}, ... a_{nm}; a_{ij}, the number of transition probability matrix is Q x Q.

4. Emission Probability Matrix (B)

   Emission probabilities are an probability to move the state i with the O_t (Observed State) time requirement. Symbolized by B = b_i (O_t) the number of emission probability matrix is Q x O.

5. Initial State Probabilities (π)

   The initial probability is symbolized by π. In named entity recognition that is the number of names of word entities, if the chance for each name of the entity word to be added will be worth one.

2.5 Viterbi Algorithm

Viterbi algorithm is an algorithm to optimally find the sequence of hidden state from real problem. It is used Viterbi trellis, reverse counting or recursive. The Viterbi algorithm is implemented to find the most likely tag sequence in the state space of the possible tag distribution based on the state transition probabilities [8] .The Viterbi algorithm allows us to find the optimal tags in linear time. The idea of the algorithm is only the most probable of all the state sequences to be considered.

2.6 Gazetteer List Method

Gazetteer list method is a rule based approach to construct the list of for different Named Entities and then applies search operations to classify the names [9]. In this research, the method needs two types of input to collection of gazetteer, one for Name of herb and second for Location of herb. The list is automatically created from scanning the document with label name or location.

3 Result and Discussion

Furthermore, the comparison of performance between HMM and proposed method to recognize the named entity of characteristic for herbal plant is shown on Table 1. As general, the validity of modified HMM is better than the HMM method only. However, recognition of name and location is unreachable 100% (1). This indicates that there is ambiguity of two or more word to be recognized in the same named entity.

Table 1. Performance of accuracy result

Named Entity	Proposed method			HMM		
	Precision	recall	f-measure	Precision	recall	f-measure
NAME	0.847	0.639	0.655	0.748	0.422	0.448
LOC	0.929	0.895	0.881	0.927	0.774	0.776
SUBST	0.992	0.895	0.929	0.992	0.812	0.860
FUNC	0.953	0.908	0.922	0.926	0.829	0.842

Then, the 2nd scenario is to know the effect of the number of training set in HMM model construction and gazetteer list of rule for terms sequence. The experimental result shows that the more number of the training set, the higher of performance is. Also, the proposed method has the higher performance than the HMM method as shown at Figure 2. (a), (b), and (c).

(a) Precision

(a) Recall

(a) F-measure

Figure 2. Performance of the number of training set

As general, the performance result of the proposed method is higher than the HMM using viterbi algorithm only. It is shown the bar chart of comparison for the both method as in Figure 3.

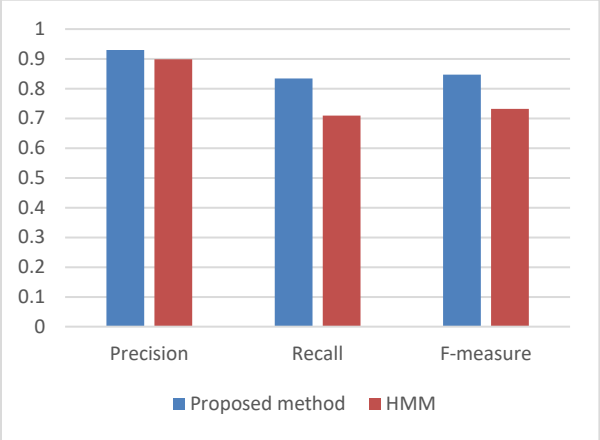

Fig.3. Comparison of average accuracy result

4. Conclusion

The proposed method which involving rule based to HMM method has been applied to the named entity recognition for herbs article. The entity is including name of herbs, substance, location and their function. The rule is applied into constructing HMM process for labelling name and location in term sequences of training set. The rule is address to reduce the ambiguity of words. It is impact for modelling in HMM. Generally, the performance result of the proposed method is better than the conventional method of HMM. The accuracy rate is also depend on the number of training set. The more number of data set, the higher of the performance is.

REFERENCES

[1] Alfred, R., Leong, L.C., Kim On, C., Antony, P., et.al. (2013). Named Entity Recognition for Malay Articles. Lecture Notes in Computer Science (LNCS, Volume 8346. pp:288-299).DOI: 10.1007/978-3-642-53914-5_25

[2] Todorovic, B.T., Rancic, S.R., Markovic, L.M., Mulalic, E.H., Dan Ilic, V.M. (2008). *Named Entity Recognition And Classification Using Context Hidden Markov Model* . Symposium On Neural Network Applications In Electrical Engineering. Neurel-2008.

[3] Suwarningsih, W., Supriana, I., and Purwarianti, A.(2014). *Imner Indonesian Medical Named Entity Recognition*. 2nd International Conference On Technology, Informatics, Management, Engineering & Environment.

[4] Sumathy, K.L., Chidambaram, M. (2013). *Text Mining: Concept, Applications, Tools and Issues- An Overview.* International Journal of Computer Applications (00975-8887). Volume 80(4).

[5] Fauzi, M.A., Arifin, A.Z. and Yuniarti, (2017). An Arabic Book Retrieval using Class and Book Index Based Term Weighting. *International Journal of Electrical and Computer Engineering (IJECE). Volume 7*(6)

[6] David Nadeau, Satoshi Sekine , "A survey of named entity recognition and classification" National Research Council Canada / New York University

[7] Lin, J. Dan Dyer, C. (2010). Data-Intensive Text Processing with Mapreduce. Available https://Lintool.Github.Io/Mapreducealgorithms/Mapreduce-Book-Final.Pdf

[8] Viterbi, A. (1967). Optimum decoding algorithm for convolutional codes. In IEEE Trans. Info. Theory (supported by AFOSR)

[9] Padmaja Sharma, Utpal Sharma, Jugal Kalita. (2011), "Named Entity Recognition: A Survey for the Indian Languages".