International Conference on Information Technology Systems and Innovation (ICITSI) 2014
Bandung-Bali, 24-27 November 2014
ISBN: 978-1-4799-6526-7

Rule-Based and Machine Learning Approach for Event Sentence Extraction in Indonesian Online News Articles

Taufik Fuadi Abidin
Department of Informatics, College of Science, Syiah Kuala University
Banda Aceh, Aceh, 23111, Indonesia
Email: taufik.abidin@unsyiah.ac.id

Rahmad Dimyathi [1)] and Ridha Ferdhiana [2)]
[1)] Department of Mathematics, College of Science, Syiah Kuala University
[2)] Department of Statistics, College of Science, Syiah Kuala University
Banda Aceh, Aceh, 23111, Indonesia
Email: ridha.ferdhiana@unsyiah.ac.id

***Abstract*—With the rapid maturity of internet and web technology over the last decades, the number of Indonesian online news articles is growing rapidly on the web at a pace we never experienced before. In this paper, we introduce a combination of rule-based and machine learning approach to find the sentences that have tropical disease information in them, such as the incidence date and the number of casualty, and we measure its accuracy. Given a set of web pages in tropical disease topic, we first extract the sentences in the pages that match contextual and morphological patterns for a date and number of casualty using a rule-based algorithm. After that, we classify the sentences using Support Vector Machine and collect the sentences that have tropical disease information in them. The results show that the proposed method works well and has good accuracy.**

***Index Terms*—Event sentence extraction, support vector machine, accuracy measure.**

I. Introduction

Different types of tropical diseases, such as dengue fever, malaria, chikungunya, lymphatic filariasis, and leprosy are usually found in tropical countries like Indonesia. According to a report, published by the Indonesian Ministry of Health in 2011, chronic filariasis cases almost reach 12,000 in a last decade [1]. Research also shows that in 2004, the largest number of malaria incidence in South-East Asia countries was found in Indonesia, Thailand, and Korea [2].

This paper presents a combination of rule-based algorithm that incorporates contextual and morphological components and machine learning algorithm to find and extract event sentence in Indonesia online news articles. Event sentence extraction (EE) is part of Information Extraction (IE) that finds sentences with specific information such as the date of an event, the number of casualty, the name of an organization, and the location of an event [3]. Such information or sentences can be very useful for a repository and a monitoring system if they are extracted correctly from the articles.

Although EE is not widely studied as name entity recognition (NER), several prominent works have been initiated to study and evaluate methods for EE in a specific language, event, or topic [3][4][5]. In this work, we introduce a rule-based algorithm that uses contextual and morphological components to find sentences that contain occurrence date, number of casualty, or both. After that, we classify the sentences using Support Vector Machine and collect those classified as "positive".

A number of sentences in previously classified web pages in tropical disease topic [6] were used as a data source. The web pages were manually read and annotated. The sentences that have the occurrence date or the number of casualty, or both were labeled {+1}, and those that have no occurrence information or the number of casualty were labeled {-1}. Our contributions are:

1. We introduce a rule-based approach to identify the sentences that have the occurrence date, the number of casualty, or both and build SVM classifier to classify the sentences and determine which of the sentences really have the date of occurrence and the number of casualty caused by tropical disease incidents.
2. We evaluate the accuracy of the proposed approach by measuring the precision, recall, and f-measure.

The paper is organized as follows: Section 2 discusses related works. Section 3 describes the proposed approach. Section 4 reports the results, and section 5 concludes the work.

II. Related Work

Research to study and evaluate methods for event sentence extraction has been done by researchers around the world. Yun [3] presented an event sentence extraction method in Korean newspapers by acquiring various features such as verbs, nouns, noun phrases, 3W (Who, When, Where). They computed weights of sentences using those features to determine which to extract. Gao [4] presented a two-stage method to extract injurious incidents information and the relationship among them. They used semi-CRFs model to extract from news the full name of people appeared in the news, location information, the organization mentioned in the news, time, and the number of injury. They selected the most frequent verbs as key description verbs to capture the relationship of the event such as *crash*, *erupt*, and *outbreak*. They claimed that their method works well and mines out reasonable relationships among the events.

Jindal [5] proposed a methodology to extract events and temporal expressions from clinical text. They showed a strategy to enforce consistency constraints on attributes of events that are close to one another. Reference [7] is one of our preliminary works in developing SVM models to find sentences containing date and number of casualty.

III. Proposed Approaches

Our event sentence extraction approach consists of two steps. First, a rule-based approach is used to identify the event sentence that may have the number of casualty and the date of tropical disease incidence. Contextual and morphological components are incorporated in this first step. Contextual are reference words that form a specific pattern of an entity. In a sentence, contextual words are commonly written near the targeted entity. Contextual words for the date of an incidence are different to those for the number of casualty. A few examples of contextual words for the date and the number of casualty are listed in Table 1 and 2 respectively.

Second, the candidates of event sentence are then classified using SVM classifier, trained, and evaluated beforehand. Positive sentences, labeled as {+1}, are collected because they may contain the occurrence date or the number of casualty.

TABLE 1
Contextual Components for Date (Time)

Label	Description	Examples
DATE	Words to specify a date	Tanggal (day), bulan (month), tahun (year)
DAY	Name of days	Senin (Monday), Selasa (Tuesday), …
MONTH	Name of months	April, Maret (March), Jan, Feb, Mar, ...

TABLE 2
Contextual Components for Number of Casualty

Label	Description	Examples
NOMINAL	Nominal figures	Satu (one), dua (two), tiga (three), …
UNIT	Unit of casualty	Warga (residents), orang (people), …
TOTAL	Number of casualty	Jumlah (total), seorang (a person), ...
VICTIM	Describe the word of casualty	Penderita (patients), pengidap (affected person), …

Morphology is a major component in the grammar that is concerned with the rules of the word formation [8]. The morphology for the number of casualty is formally written as a combination of digits and contextual words such as *dua korban (two victims)* or *2 orang meninggal (2 people dead)*, whereas the morphology for a date is commonly written as a combination of digits and strings in specific formats such as *dd/dd/dddd*, *dd-dd-dddd*, *dd/dd/dd*, *dd-dd-dd*, *dd name-of-month dddd*, and some other forms.

We constructed the numerical features of a sentence by taking the ratio of 1-gram, 2-grams, and 3-grams words' weights over the total number of n-grams words in that sentence that are found in the dictionary of each class [9]. Hence, there are 6 numerical features constructed for each sentence, described as follows:

1) The *1-gram negative feature* is the sum of 1-gram weight found in a negative dictionary divided by the total number of 1-gram word in the sentence.
2) The *2-grams negative feature* is the sum of 2-grams weight found in a negative dictionary divided by the total number of 2-grams words in the sentence.
3) The *3-grams negative feature* is the sum of 3-grams weight found in a negative dictionary divided by the total number of 3-grams words in the sentence.
4) The *1-gram positive feature* is the sum of 1-gram weight found in a positive dictionary divided by the total number of 1-gram word in the sentence.
5) The *2-grams positive feature* is the sum of 2-grams weight found in a positive dictionary divided by the total number of 2-grams words in the sentence.
6) The *3-grams positive feature* is the sum of 3-grams weight found in a positive dictionary divided by the total number of 3-grams words in the sentence.

After the features are constructed, SVM classifies the sentences. Although SVM is basically a linear classifier, by using a kernel trick, it can also solve non-linear problems [9]. Research has shown that SVM often performs well for classification tasks and its accuracy usually beats other classifiers [6].

The accuracy of the proposed approach is measured by precision, recall, and F-measure. Precision (P) measures the ratio of objects correctly assigned to a particular class (true positives) and the total number of objects assigned to that class (true positives + false positives), whereas Recall (R) measures the ratio of objects correctly assigned to a particular class (true positives) and the actual number of objects belong to that class. F-measure computes the harmonic mean of Precision and Recall.

$$P = \frac{TP}{TP + FP} \tag{1}$$

$$R = \frac{TP}{TP + FN} \tag{2}$$

$$F = \frac{2 \cdot P \cdot R}{P + R} \tag{3}$$

IV. Experimental Results

A. *Dataset*

A collection of sentences derived from 1,863 tropical disease web pages [6] was used as the dataset. The sentences were split into two groups. The first group is a set of sentences that has the date of occurrence or the number of casualty or both. This group is labeled {+1}. The second group is a set of sentences that has no event information. They are labeled as {-1}.

An example of a sentence that contains an incidence date is: *"Dinkes Kota Depok berencana menggelar pengobatan massal kaki gajah di 6 kecamatan dengan sasaran 1,2 juta penduduk karena kota ini telah menjadi daerah endemis kaki gajah sejak tahun 2005*", where as an example of a sentence that contains the number of casualty is: *"Jumlah penderita filariasis di Kabupaten Kediri, Jawa Timur mencapai 16 kasus sehingga daerah ini termasuk endemis filariasis"*. Both sentences are positive samples. In this work, we used 131 positive and 142 negative sentences for the date of occurrence problem and used 372 positive and 138 negative sentences for the number of casualty problem. Table 3 recaps the number of sentences used in training phase.

TABLE 3
DISTRIBUTION DATASET BY CLASS LABELS

Dataset	**Class**	**Number of Sentences**
Date of Occurrence	+1	131
	-1	142
Total		**273**
Number of Casualty	+1	372
	-1	138
Total		**510**

Two different SVM classifiers were trained to categorize the sentences. The first classifier is aimed to determine whether the sentence contains the occurrence date, while the second SVM classifier is aimed to identify whether the sentence contains the number of casualty. We did not split the two datasets into fixed portion of training and testing sets during training phase, instead we applied *k*-fold cross-validation to split the datasets into *k* complementary subsets. One subset was used for performing the analysis and the other subset was used for validating the analysis. K-fold cross-validation is a good method to use in training phase to avoid over fitting issue. In this experiment, we used *k=10*.

B. *Extracting Event Sentence for Date of Occurrence*

In order to find the best SVM classifier for incidence date, we trained SVM using linear, polynomial, and radial kernel on the datasets listed in Table 3 and applied *k*-fold cross-validation. We measured the accuracy of SVM classifiers by looking at their F-measure values. The average of F-measure for 10-fold cross-validation is 80.27%. The best model was achieved using polynomial kernel and the accuracy is 93.33%. We further evaluated our approach by classifying 1,000 additional sentences. The confusion matrix of the evaluation is summarized in Table 4. Examples of positive sentences found:

*"Berdasarkan data di Dinkes Aceh, sejak **Januari** hingga **Agustus 2009**, tercatat setidaknya 2095 kasus cikungunya yang penderitanya tersebar di beberapa kabupaten dan kota di Aceh".*

*"Sembilan kasus yang sama ditemukan kembali pada **tahun 2009** dan 7 kasus pada **2010**".*

TABLE 4
CONFUSION MATRIX IN EVALUATING SVM FOR DATE OF OCCURRENCE

		Prediction	
		Positive	Negative
Actual	Positive	306	134
	Negative	25	535

$$\text{Precision} = \frac{306}{306+25} = 0.92,\ \text{Recall} = \frac{306}{306+134} = 0.70$$

$$\text{F-measure} = \frac{2 \cdot 0.92 \cdot 0.70}{0.92+0.70} \text{ x } 100\% = 79.51\%$$

The experimental results using 10-fold cross-validation and 1,000 new sentences for extracting event sentence for date of occurrence show that the overall accuracy of SVM classifier reaches to 79.51%, and it yields precision and recall up to 92% and 70% respectively.

C. *Extracting Event Sentence for Number of Casualty*

The process of extracting event sentence for number of casualty begins with a simple rule. First we identify and collect all sentences with contextual words, as shown in Table 2 and identify sentences that contain nominal numbers or digits in them. Then, the sentences are classified using the best trained SVM model to double check whether the sentences really have the number of casualty. Examples of positive sentences extracted:

*"Sejak Januari hingga akhir Maret 2011, rumah sakit telah merawat sebanyak **27 orang** pasien flu burung".*

*"Sejak deman berdarah merebak, sebanyak **24 pasien** positif dirawat, walau beberapa diantaranya sudah diperbolehkan pulang karena kondisi sudah membaik".*

We trained SVM using three different kernels, i.e. linear, polynomial, and radial kernel. The datasets listed in Table 3 were used and the accuracy of SVM classifiers was observed. The results show that using 10-fold cross-validation, the average F-measure is 90.83% and the best SVM model was trained using polynomial kernel, with the accuracy of 96.55%. We further evaluated our approach by classifying 1,000 new additional sentences. The confusion matrix is summarized in Table 5.

TABLE 5
CONFUSION MATRIX IN EVALUATING SVM FOR NUMBER OF CASUALTY

		Prediction	
		Positive	Negative
Actual	Positive	695	92
	Negative	111	102

$$\text{Precision} = \frac{695}{695+111} = 0.86, \text{ Recall} = \frac{695}{695+92} = 0.88$$

$$\text{F-measure} = \frac{2 \cdot 0.86 \cdot 0.88}{0.86+0.88} \text{ x } 100\% = 86.99\%$$

The experimental results using 10-fold cross-validation and 1,000 new sentences show that the overall accuracy of SVM classifier, trained for this purpose, is 86.99% with precision of 86% and recall of 88% respectively.

D. Design of Flow to Extract and Renew Event Entities

A rule-based algorithm to identify the event information using contextual and morphological components and SVM classifier to classify sentences containing the date and the number of casualty of tropical disease incidences have been evaluated and developed.

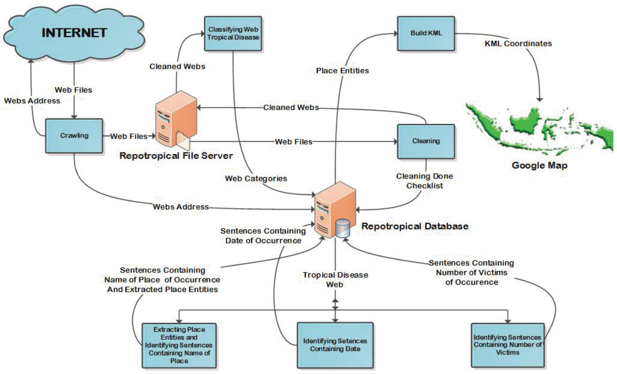
INTERNET
Classifying Web Tropical Disease
Build KML
KML Coordinates
Cleaned Webs
Place Entities
Web Files
Webs Address
Cleaned Webs
Crawling
Web Files
Web Files
Cleaning
Repotropical File Server
Google Map
Web Categories
Cleaning Done Checklist
Webs Address
Repotropical Database
Sentences Containing Date of Occurrence
Sentences Containing Name of Place of Occurrence And Extracted Place Entities
Tropical Disease Web
Sentences Containing Number of Victims of Occurence
Extracting Place Entities and Identifying Sentences Containing Name of Place
Identifying Setences Containing Date
Identifying Sentences Containing Number of Victims

Fig. 1. Flow to periodically extract and update tropical disease incidence entities.

Fig. 1 shows a flow to periodically extract and update tropical disease event information from the web. The process starts by finding and crawling relevant Indonesian news articles using a set of query seeds and Google Site Search API. After that, the articles are cleaned and classified. Those that fall into tropical disease category are further examined to get the location of the incidence [7] and to determine the sentences that contain the date of occurrence and the number of casualty as proposed in this paper. The tropical disease incidence information are then stored in MySQL server for further reference and they are organized in a keyhole markup language notation for expressing geographic visualization in Google Earth.

V. CONCLUSION

The proposed method is composed of the following two steps: (1) determining candidates of sentence using a combination of rule-base algorithm and contextual and morphological components; (2) the candidates of sentences are further classified using SVM to double check whether the sentences are event sentences. The evaluation results show that the accuracy of SVM to classify the sentences that have the incidence date and the number of casualty are 79.51% and 86.99% respectively.

VI. ACKNOWLEDGMENT

This work was supported by the Directorate General of Higher Education, Ministry of Education and Culture, Indonesia through the National Strategic Grant 2014, contract number 172/UN11.2/LT/SP3/2014.

REFERENCES

[1] Indonesian Ministry of Health, “Neglected Tropical Diseases in Indonesia: An Integrated Plan of Action,” 2011.

[2] Behrens, et al., “The Incidence of Malaria in Travellers to South-East Asia: Is Local Malaria Transmission a Useful Risk Indicator?” Malaria Journal, vol. 9, no. 1, pp. 266, 2010.

[3] B. Yun, T. Kim, “Event Sentence Extraction in Korean Newspapers,” Computational Linguistics and Intelligent Text Processing Lecture Notes in Computer Science, vol. 2588, pp. 580-583, 2003.

[4] C. Gao, Y. Song, P. Jia, “A Two-Stage Extraction Method for Events and Their Relationship in Injurious Incidents Monitoring,” Journal of Information and Computational Science, vol. 8, no.13, p. 2791–2798, 2011.

[5] P. Jindal, D. Roth, “Extraction of Events and Temporal Expressions from Clinical Narratives,” Journal of Biomedical Informatics, vol 46, pp. S13-S19, 2013.

[6] T.F. Abidin, R. Ferdhiana, H. Kamil, “Learning to Classify Tropical Disease Web Pages from Large Indonesian Web Documents”, In Proc. of the 4th International Conference on Computer and Electrical Engineering, pp. 14–15, 2011.

[7] T.F. Abidin, R. Ferdhiana, H. Kamil, “Automatic Extraction of Place Entities and Sentences Containing the Date and Number of Victims of Tropical Disease Incidence from the Web”, Journal of Emerging Technologies in Web Intelligence, vol. 5, no. 3, pp. 302-309, August 2013.

[8] A. Spencer, Morphological Theory: An Introduction to Word Structure in Generative Grammar. Oxford & Cambridge, 1991, pp. xviii + 512.

[9] T.F. Abidin, A. Misbullah, M. Subianto, “Determining Features of Web Documents and Building a Web Classifier using SVM,” AISS: An International Journal of Research and Innovation, vol. 3, no. 10, pp. 401–408, 2011.

[10] Joachims, “Making Large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning,” B. Scholkopf , C. Burges and A. Smola (ed.), MIT Press, 1999.