Malay Named Entity Recognition Based on Rule-Based Approach

Rayner Alfred, Leow Chin Leong, Chin Kim On, and Patricia Anthony

Abstract—**A Named-Entity Recognition (NER) is part of the process in Text Mining and it is a very useful process for information extraction. This NER tool can be used to assist user in identifying and detecting entities such as person, location or organization. However, different languages may have different morphologies and thus require different NER processes. For instance, an English NER process cannot be applied in processing Malay articles due to the different morphology used in different languages. This paper proposes a Rule-Based Named-Entity Recognition algorithm for Malay articles. The proposed Malay NER is designed based on a Malay part-of-speech (POS) tagging features and contextual features that had been implemented to handle Malay articles. Based on the POS results, proper names will be identified or detected as the possible candidates for annotation. Besides that, there are some symbols and conjunctions that will also be considered in the process of identifying named-entity for Malay articles. Several manually constructed dictionaries will be used to handle three named-entities; Person, Location and Organizations. The experimental results show a reasonable output of 89.47% for the F-Measure value. The proposed Malay NER algorithm can be further improved by having more complete dictionaries and refined rules to be used in order to identify the correct Malay entities system.**

Index Terms—**Information extraction, Malay named entity recognition, named entity recognition, rule-based.**

I. Introduction

Natural Language Processing (NLP) is one of the important fields in Computer Science. Basically, it analyzes text that is based on both a set of theories and a set of technologies [1]. NLP initially started at the late 1940s when machine translation was first used to decrypt enemy codes during World War II. However, not many researches in NLP were conducted until the 1980s. There are a lot of fields that apply the NLP technologies such as Information Retrieval, Information Extraction, Question-Answering and etc. [1]. Most recent studies focus on Information Extraction (IE).

There are three types of input files in IE which are structured, semi-structured or free text which is as shown as Fig. 1 [2]. Structured inputs refer to HTML pages while semi-structured inputs refer to XML pages and records. News articles are considered as unstructured input texts and they are written and understandable by human. News articles are hard to be understood by machines. A computer machine will not be able to comprehend the content of these articles. Nowadays, a huge volume of articles can be easily retrieved and extracted from websites. Hence, it would take a long time for human to manually process these articles in a short time. Besides that, the process of annotating articles manually often provides biased results. Hence, an automated process is needed and such process is known as Information Extraction.

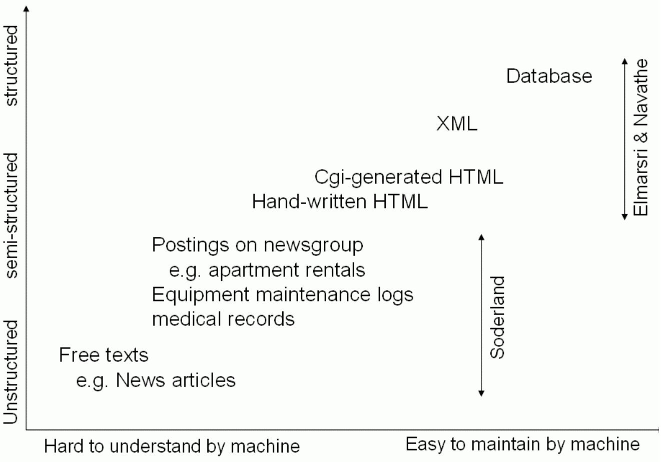


Fig. 1. Structurization of various documents [2].

IE is a process that extracts information from unstructured articles to provide more useful information. Given an article, a machine will learn on how to answer certain questions (e.g., How can we determine who is the CEO of a company? What is that name of the company?) One of the sub-tasks of IE is to help the process to identify and extract such information called named-entity and it is known as a Named Entity Recognition (NER) process.

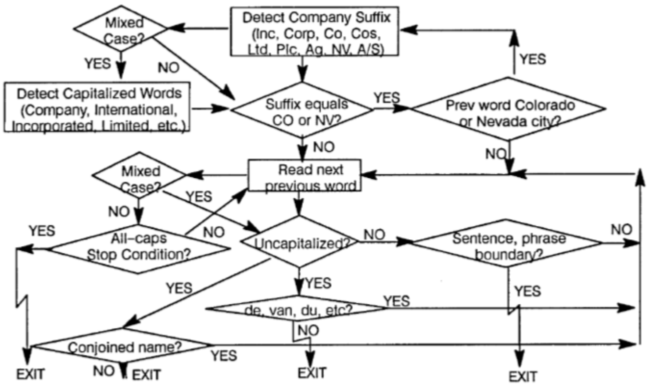

Mixed Case?
Detect Company Suffix (Inc, Corp, Co, Cos, Ltd, Plc, Ag, NV, A/S)
YES
NO
Detect Capitalized Words (Company, International, Incorporated, Limited, etc.)
Suffix equals CO or NV?
YES
Prev word Colorado or Nevada city?
YES
NO
NO
Mixed Case?
Read next previous word
YES
NO
YES
All–caps Stop Condition?
NO
Uncapitalized?
NO
Sentence, phrase boundary?
NO
YES
YES
de, van, du, etc?
YES
YES
NO
Conjoined name?
YES
EXIT
NO
EXIT
EXIT
EXIT

Fig. 2. Company name extraction [23].

A Named Entity Recognition process was a popular discussion at the Sixth Message Understanding Conference (MUC-6) [3], [4]. The NER process helps users to produce a

Manuscript received November 5, 2013; revised January 7, 2013. This work has been supported by the Long Term Research Grant Scheme (LRGS) project funded by the Ministry of Higher Education (MoHE), Malaysia under Grants No. LRGS/TD/2011/UiTM/ICT/04.

Rayner Alfred, Leow Chin Leong, and Chin Kim On are with the COESA, Universiti Malaysia Sabah, 88400 Kota Kinabalu, Sabah, Malaysia (e-mail: ralfred@ums.edu.my, dragon_july14@hotmail.com, kimonchin@ums.edu.my).

Patricia Anthony is with the Department of Applied Computing, Faculty of Environment, Society and Design, Lincoln University, Christchurch, New Zealand (e-mail: patricia.anthony@lincoln.ac.nz).

DOI: 10.7763/IJMLC.2014.V4.428

more meaningful corpus by identifying proper names in the corpus and classifying them into groups such as person, organization, locations and etc. For example, the query "Steve Job" should not check only on the word "Steve" or "Job". The word "Job" may lead to the process of searching for similar word such as "occupation". Hence it will lead to other meaning instead of "Steve Job," the co-founder of Apple Inc.

One of the early named entity studies was introduced by Lisa F. Rau [5]. Rau proposed an automated way to recognize company names from financial news. Company entities are extracted based on heuristics rules. Fig. 2 shows the algorithm of the proposed method.

The implementation of the NER algorithm for NLP is normally influenced by the domain of the studies. A domain-specific NER application may not be applicable for recognizing named-entities on other specific domains such as restaurant guides. For instance, AbGene [6], Abner [7] and BioNer [8] will not perform well in processing military articles as they are designed for different domains. In addition to that, different languages may require different techniques in recognizing the named entity. For instance, detecting the types of named entity for articles written in English language could easily be done by detecting the proper nouns. Proper nouns usually start with a capital letter. It is used to represent a unique named entity such as people, location, organization and etc. However, such methods may not be applicable to be applied for articles written in Arabic language as it does not contain such unique symbols that can be used to detect the named entity [9]. This is because most languages differ morphologically from other languages. In short, the implementation of the Named Entity Recognition depends upon the domain of studies and also the type of languages used.

There are a few NER systems that exist for various types of languages such as English, Indonesia, Arabic, Hindu and etc. However there is no existing system that is design to detect types of named entity in Malay language. Hence, in this paper, a rule-based Malay NER framework will be proposed that is designed to assist users in identifying types of named entity in order to improve the process of retrieving articles written in Malay language more effectively and efficiently.

This paper is organized as followed. Section II describes some of the works related to named entity recognition methods. Section III describes the general overview of the proposed rule-based named entity recognition for Malay language. Section IV outlines the experimental setup and Section V discusses the results obtained. Finally Section VI concludes this paper.

II. Types of NER

Algorithms for named-entity recognition (NER) systems can be classified into three categories; rule-based, machine learning and hybrid [10]. A Rule-Based NER algorithm detects the named entity by using a set of rules and a list of dictionaries that are manually pre-defined by human. The rule-based NER algorithm applies a set of rules in order to extract pattern and these rules are based on pattern base for location names, pattern base for organization name and etc. The patterns are mostly made up from grammatical, syntactic and orthographic features [10]. In addition to that, a list of dictionaries is used to speed up the recognition process. However, the types of dictionaries affect the performance of the NER systems and these dictionaries normally include the list of countries, major cities, companies, common first names and titles [11].

Next, a machine-learning NER algorithm normally involves the usage of machine learning (ML) techniques and a list of dictionaries. There are two types of ML model for the NER algorithms; supervised and unsupervised machine learning model. Unsupervised NER does not require any training data [12], [13]. The objective of such method is to create the possible annotation from the data. This learning method is not popular among the ML methods as this unsupervised learning method does not produce good results without any supervised methods.

Unlike unsupervised NER methods, supervised NER methods require a large amount of annotated data to produce a good NER system. Some of the ML methods that had been used for NER algorithm includes artificial neural network (ANN) [9], Hidden Markov Model (HMM) [14], Maximum Entropy Model (MaxEnt) [15], Decision Tree [16], Support Vector Machine [17] and etc. ML methods are applicable for different domain-specific NER systems but it requires a large collection of annotated data. Hence, this might require high time-complexity to preprocess the annotate data.

Finally, a hybrid named entity recognition algorithm implements both the rule-based and machine learning methods [18]. Such method will produce a better result. However, the weaknesses of the rule-based are still unavoidable in this hybrid system. A domain-specific NER algorithm may need to customize the set of rules used to recognize different types of named entity when the domain of studies is changed.

III. A Rule-Based Named-Entity Recognition Algorithm for Malay language

In this paper, a rule-based NER for Malay language will be proposed. In this work, a rule-based is applied instead of the machine learning technique due to the lack of annotated corpus resources for Malay language that can be used as a training data. Creating a large annotated dataset for Malay language is also time-consuming. The proposed rule-based NER for Malay language consists of three major steps. The first step is the tokenization. The purpose of the tokenization process is to split the sentences into tokens. For instance, the sentence "Pengerusi KMR telah sampai di Kuala Lumpur hari ini." will be converted into several tokens as shown in Table I. The sentence is split into words, punctuation and numbers.

The second step involves the part-of-speech tagging (POS) process. In order to retrieve the part-of speech tagging, a Rule-Based Part of Speech (RPOS) tagger has been implemented. RPOS tagger is a simple rule-based POS tagger for Malay languages that applies a POS tag dictionary and affixing rules in order to identify the word definition [19]. The flow of the RPOS tagger is shown in Fig. 3.

The rule-based NER for Malay language is basically implemented based on the rule-based POS tagging process for

Malay language and contextual features rules. The contextual features rules are studied and proposed for Iban and Indonesia Languages [20], [21] which are almost similar to Malay language. For instance, when the part-of-speech tag for the current word shows that the current word is proper noun, then a specific rule will be applied to this current word in order to determine whether it is an entity or not. In other words, the rules are built based on the POS-tagging contexts. In this work, these rules are designed to detect three major types of named entities that include a person, an organization and a location.

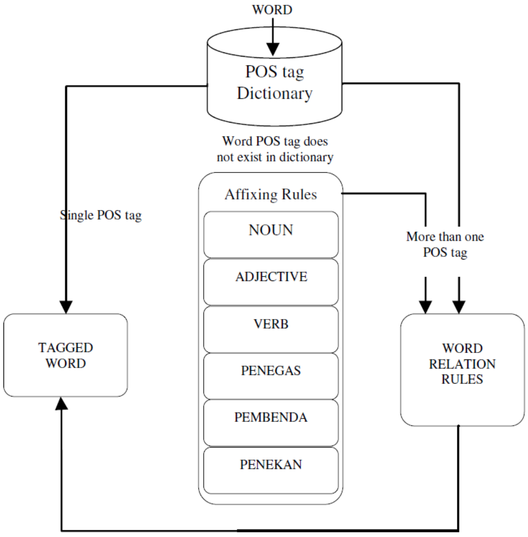


Fig. 3. The flow of the rule-based part of speech tagging for malay language.

For instance, the tokenized words will be initially evaluated using the POS tag dictionary. There are roughly more than 8,700 words in the tag dictionary which are retrieved from the Thesaurus Bahasa Melayu and stored manually in the POS tag dictionary. If the POS tag dictionary returns more than one tag results, the best tag will be chosen according to the predefined rules [19]. The rules that are used to determine the tag are shown in Table II. Table III shows the list of POS tag for Malay language words. However, if there is no word in the dictionary that can match the current word, then the affixing rules will be applied to determine whether the word is a noun, an adjective or a verb type of word as shown in Tables IV, V and VI. For instance, the result of this POS tagging process for the sentence "Pengerusi KMR telah sampai di Kuala Lumpur hari ini." is shown in Table I.

TABLE I: PART-OF-SPEECH TAGGING RESULT

Word	POS-tagging
Pengerusi	<POS>NNP</POS>
KMR	<POS>NNP</POS>
telah	<POS>AUX</POS>
sampai	<POS>VB</POS>
di	<POS>IN</POS>
Kuala	<POS>NNP</POS>
Lumpur	<POS>NNP</POS>
hari	<POS>RB</POS>
ini	<POS>NN</POS>
.	<POS>PNC</POS>

Fig. 4 shows the framework of the proposed Malay-NER. After classifying the POS-tagging of the tokens, tokens that are classified under proper nouns category will be applied into the rules with the exception for the location and person prepositions. At first, the articles will be checked against the organization suffixes. Organization suffixes are checked first rather than the person or location rules because there are people's names or locations' names that exist in a company name.

TABLE II: PART-OF-SPEECH TAGGING RESULT

Word Type	Valid Sequences of Word Types
Noun (NN)	adjective (JJ), adverb (RB), verb (VB), noun (NN), preposition (IN)
Verb (VB)	auxiliary (AUX), adverb (RB), noun (NN), penekan (PEN), pembenda (BND)
Adjective (JJ)	penguat (GUT), preposition (IN)
Adverb (RB)	verb (VB), preposition (IN), adjective (JJ), noun (AUX)
Direction (DR)	noun (NN), preposition (IN)
Preposition (IN)	noun (NN), verb (VB), adjective (JJ)
Auxiliary (AUX)	adjective (JJ), verb (VB), preposition (IN)
Cardinal number	noun (NN)
Penekan (PEN)	adverb (RB), noun (NN), conjunction (CC)
Pembenda (BND)	conjunction (CC), noun (NN)
Conjunction (CC)	noun (NN), verb (VB), preposition (IN), adjective (JJ)
Penguat (GUT)	adjective (JJ)
Interrogative (WP)	noun (NN), verb (VB)
Pangkal ayat (PNG)	noun (NN)

TABLE III: POS TAG LIST FOR MALAY

Word Type (English language)	Subtype (English language)	Subtype (Malay language)	Tag
Noun			NN
	Proper noun		NNP
Verb			VB
Adjective			JJ
Function	Conjunction	Kata hubung	CC
	Interjection	Kata seru	UH
	Interrogative	Kata Tanya	WP
	Command	Kata perintah	CO
		Kata pangkal ayat	PNG
	Auxiliary (Amplifier)	Kata bantu	AUX
		Kata penguat	GUT
	Particles	Kata penegas	RP
	Negation	Kata nafi	NEG
		Kata pemeri	MER
	Preposition	Kata sendi name	IN
		Kata pembenar	BNR
	Direction	Kata arah	DR
	Cardinal number	Kata bilangan	CD
		Kata penekan	PEN
		Kata pembenda	BND
	Adverb	Adverb	RB

For example, in the named-entity "Hong Leong Bank," "Hong Leong" might refer to a name of a person. If the person rules are used to identify the entity first, it will be identified as a named-entity for a person because the word "Hong" will be referred as a surname of a person. Hence, by checking against the organization suffixes, the word "Bank" will be identify as an organization entity. After that, the proper names before the word "Bank" will be recognized as the name of the bank.

After detecting and recognizing the organization entities, then the location preposition will be applied to identify the named-entity for locations. The list of preposition used for detecting location is listed in Table VII. A word will be checked against the preposition for a person if there is no location preposition identified. The flow goes by checking against the organization rules, location rules, person rules and with the existing entities.

TABLE IV: NOUN AFFIXING IDENTIFICATION RULES

Rules	Prefix	Next Character	Sequences of character	Suffix
1a	Pe	ny, ng, r, l and w	a-z	an
1b	Pem	b and p	a-z	an
1c	Pen	d, c, j, sy and z	a-z	an
1d	Peng	g, kh, h, k and vowel	a-z	an
1e	Penge	-	a-z (3 to 4 character)	an
1f	pel or ke	-	a-z	an
1g	Juru, maha, tata, pra, swa, tuna, eka, dwi, tri, panca, pasca, pro, anti, poli, auto sub, supra	-	a-z	-
1h	not started with me, meng, mem, menge, ber, be, di, diper	-	a-z	an, at, in, wan, wati, isme, isasi, logi, tas, man, nita, ik, is, al

TABLE V: ADJECTIVE AFFIXING IDENTIFICATION RULES

Rules	Prefix	Next Character	Sequences of character	Suffix
2a	ter, se, bi	-	a-z	-
2b	ke	-	a-z	an
2c	not starting with di and men	-	a-z	-

TABLE VI: VERB AFFIXING IDENTIFICATION RULES

Rules	Prefix	Next Character	Sequences of character	Suffix	
				May end with	
3a	me	ny, ng, r, l, w, y, p, t, k, s	a-z	-	-
3b	mem	b, f, p and v	a-z	kan and i	-
3c	men	d,c, j, sy, z, t and s	a-z	kan and i	-
3d	meng	g, gh, kh, h, k and vowel	a-z	-	-
3e	menge	-	a-z (3 to 4 character)	an	-
3f	memper or diper	-	a-z	kan or i	-
3g	ber	not r	a-z	kan or an	-
3h	bel	-	a-z	-	-
3i	ter	not r	a-z	-	-
3j	ke	-	a-z	-	an
3k	-	-	a-z	-	i or kan
3l	di or diper	-	a-z	kan or i	-

A. Rules for Identifying a Person-Entity

In this work, the person-entity is recognized based on the person’s titles and these person’s titles are identified based on the standard titles used in Malay and English language. If the word is a person title, then the rest of the proper noun word is known as a person’s name. These titles include “Yang Teramat Mulia,” “Yang Amat Berhormat” and “Dr”. For instance, in this sentence, “Dr. Tan Boon Keong telah tiba di Sabah hari ini,” the word “Dr.” is a title of a person. Hence “Tan Boon Keong” will be recognized as a person. Other than that, there are other pattern recognition methods that can be used to detect a person entity such as “A. Monhagen”. If the word starts with a single character followed by proper nouns then it is recognized as a person name. The example of such pattern is “M. Night Shyamalan”.

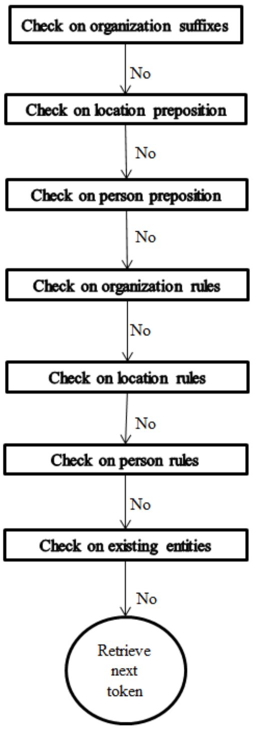

Check on organization suffixes
No
Check on location preposition
No
Check on person preposition
No
Check on organization rules
No
Check on location rules
No
Check on person rules
No
Check on existing entities
No
Retrieve next token

Fig. 4. The framework of Malay NER

When the person’s title is not included in the person’s name, then a different set of rules will be applied to identify the patterns of name. Since the name of a person is highly dependent on the ethnic group or the nationality of a person, a different set of rules will need to be defined to handle a person entity for the Malaysian people. For instance, Malaysian Chinese people usually start their names with their surname, Malaysian Malay and Indian people start with their first names and followed by their father’s name. Besides that, the gender of a person can also be determined based on the person’s name. For instance, for Malay people, they use “Bin” for male (e.g., Ali Bin Ahmad) and “Binti” for female (e.g., Helmi Binti Yunus) in their names. Indian people use” A/L” which stands for “Anak Lelaki” or “Son Of” in English and “A/P” which stands for “Anak Perempuan” or “Daughter of” in English for their names. In some cases, for Sarawak native, they use the word “Anak” in their naming convention. Hence, by detecting such patterns will assist the process of recognizing the person entity.

Besides that, a person entity can also be recognized by using the preposition features. These preposition features include “oleh” (e.g., means “by”). For instance, given a sentence, “Hadiah ini disedia oleh En. Lim” (e.g., This present is prepared by Mr. Lim), the person entity can be identified by looking at the word “oleh”.

B. Location Rule

The location entity can be identified by looking at the location’s prefixes and the usage of the “Di” preposition in the sentence. The words “Jalan,” “Lorong,” “Taman” and

"Persiaran," are commonly used for the location's prefixes. For instance, "Lorong Kinabalu," "Persiaran Damai" and "di Ranau" are some of locations named entity that can be identified as "Kinabalu," "Damai" and "Ranau".

C. Organization Rule

Prefixes and the suffixes of organizations' names may be used to identify the named entity for an organization. The prefix and suffix used for organization are shown in Table VI. For instance, "Syarikat Buku" is known as an organization because of the word "Syarikat" is identified as an organization's prefix. "Hong Leong Bank" is also known as an organization due to the word "Bank". Besides that, there is another pattern that can be used to recognize an organization entity which is "Persatuan Pelayaran Malaysia (MYA)". In this case, "MYA" will be recognized as an organization entity as it is an abbreviation of "Persatuan Pelayaran Malaysia".

Hence, it can be concluded that when a string of words, that appears before the substring that consists of "(" and ")" symbols, is identified as an organization entity, then the abbreviation that appears within the parentheses (e.g., MYA) will be considered as an organization entity too.

Other than checking on the rules, the system will also check the current tokens with existing identified entities. It is common in articles to have just the name of the person or an organization instead of the full name. Normally, the author will write the full name of the entities in the beginning of the articles and only refer the common name in the rest of the articles as the reader will understand that the common name referred to the same person mentioned earlier.

However, a system normally does not understand whether the current token has been identified as one of the named-entities or not. Therefore, the system will also check the current token against the existing annotated entities in order to classify this common name. For example, the entity "Automobile Sdn. Bhd." may be referred as "Automobile" in the rest of the article. The word "Automobile" would not be recognized as a new entity as it does not fulfill any of the rules. However, we understand that it refers to a company. Hitherto, by checking upon annotated entities, we can annotate "Automobile" as a company. Other than that, it also helps in the biased result. Based on the example given, it is possible that the company name consists of person name. By checking the words against existing entities will help us to determine the category of the current tokens.

Table VII shows the list of rules that had been applied for the three categories. There are some exceptions that need to be handled in recognizing all these types of named entity. This is because there are some named entities that do not fulfill any rules that are predefined. For example, the proposed rules do not handle the entity recognition process for the phrase "Jabatan Keselamatan dan Kesihatan Pekerjaan" in which it consists of the word "dan" (e.g., sometimes "&" is used) and the word "dan" is spelled in a small letter. In order to overcome such weaknesses, a list of dictionaries shall be used to handle these types of named entity that are difficult to be detected or recognized by using the proposed rules.

Table VIII shows the type of dictionaries that had been used in this work.

TABLE VII: LIST OF CONTEXTUAL RULES

Feature	Example
Location Prefix	Jalan, Bukit, Kampung
Preposition that usually followed by location	Di, ke
Organization prefix	Syarikat, Kelab, Persatuan
Organization suffix	Sdn. Bhd.,
Person prefix	Tan, Lim
Person middle	Bin, binti, a/p, a/l, anak
Person title	Dato Paduka, Tun
Preposition that usually followed by person	Oleh

TABLE VIII: TYPES OF DICTIONARIES

Dictionaries	Example
Location Prefix	Pekan, Padang, Pulau, Simpang, Gunung
Location	Dungun, Selangor, Sabah, Labuan, Sarawak
Person Title	Puan, Encik, Pn., En., Datin
Organization Prefix	Parti, Pertubuhan, Persatuan, Angkatan, Jabatan
Organization Abbreviation	MAA, KDCA, JPN, JPA, ATM
Organization Name	McGraw-Hill, McDonalds, Fujitsu
Organization Suffix	Sdn. Bhd., Berhad, Bank, Airlines

IV. EXPERIMENTAL SETUP

In order to evaluate the effectiveness of the proposed rule-based named-entity recognition algorithm for Malay language, four different categories articles had been retrieved from two local Malay websites (http://www.bernama.com/bernama/v7/bm/ and http://www.mstar.com.my/). There are a total of 155 articles retrieved in General category, 143 articles retrieved in Economic category, 35 articles retrieved in Politic category and lastly 30 articles retrieved from Sport category. The main purpose of this experiment is to identify the types of patterns that are failed to be identified by the proposed NER algorithm. The NER for these three types of named entity (e.g., person, location and organization) will be evaluated based on three measures which are Recall, precision and F-measure as proposed in MUC [20].

$$\mathrm{Re}\,call = \frac{Correct + 0.5 * Partial}{Possible} \tag{1}$$

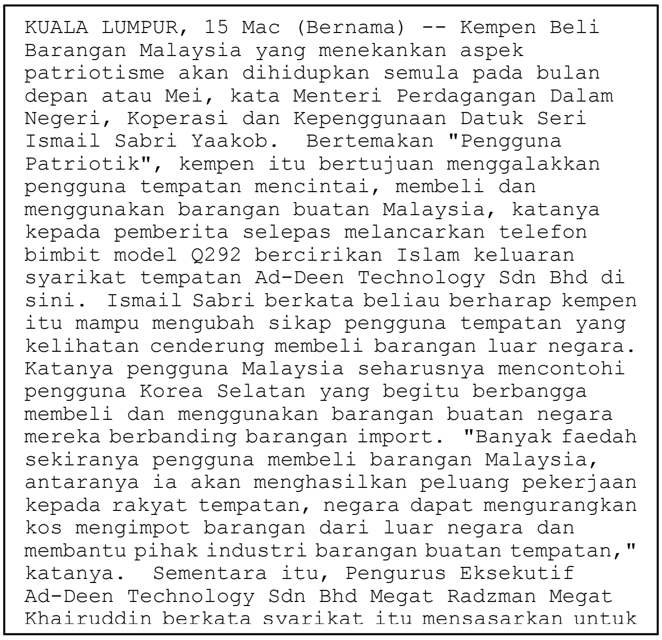
$$\mathrm{Pr}\,ecision = \frac{Correct + 0.5 * Partial}{Actual} \tag{2}$$

$$F - Measure = \frac{\mathrm{Re}\,call * \mathrm{Pr}\,ecision}{0.5 * (\mathrm{Re}\,call + \mathrm{Pr}\,ecision)} \tag{3}$$

In this work, in Equation (1), the term Correct represents the number of correct annotations produced by the proposed NER algorithm for Malay language and the Partial term shows the number of partially correct annotations. For instance, given an entity in two words "Barack Obama," the proposed NER algorithm should be able to identify these two words as a Person entity. However, if the proposed NER algorithm is only able to annotate either "Barack" or "Obama" as a Person entity, then it is called a partially correct annotation. A manually tagged annotation used for training purposes is called as "Possible" term. The term Actual shows the actual number of annotated entity that should be produced by the proposed NER algorithm for Malay language. In short, the produced annotated entity can be categorized into Correct, Partially Correct or Incorrect entity.

V. Results and Discussions

The results of the proposed NER for Malay language are comparable to other NER algorithm for other language [21]-[23] in which the obtained F-measure is 89.47 % with 94.44% of recall and 85% of precision rates. Fig. 5 shows one of the sample articles and the annotated results are shown in Fig. 6.

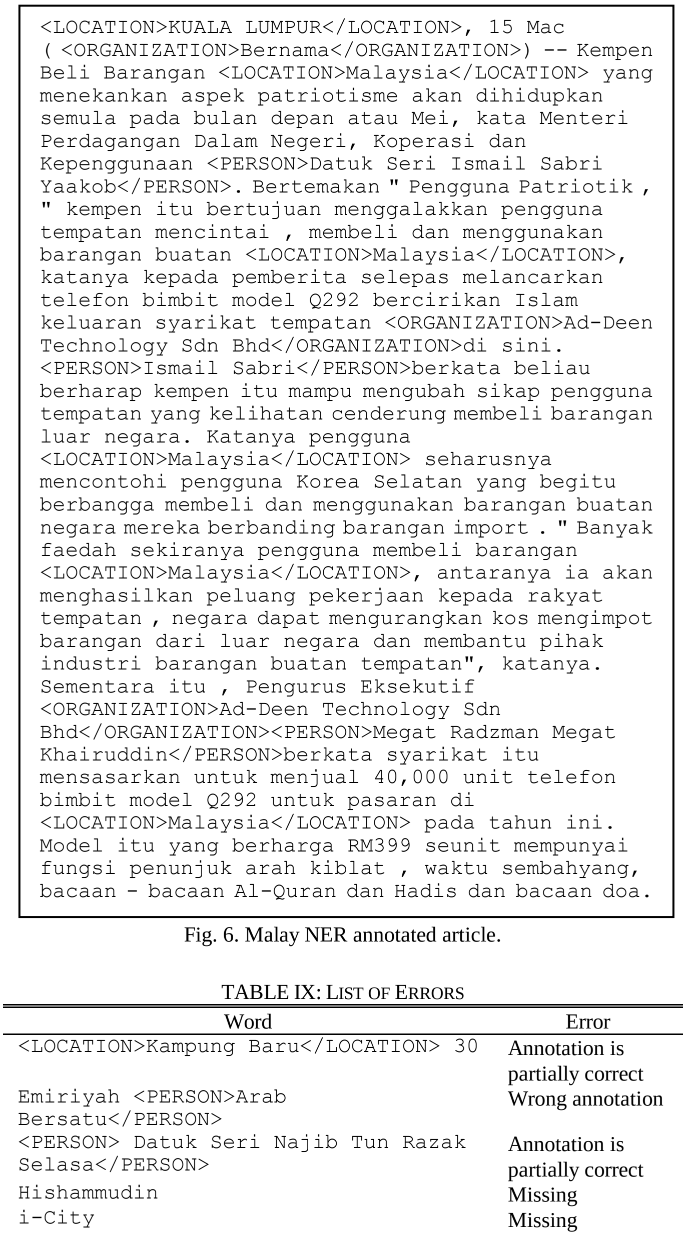


```
KUALA LUMPUR, 15 Mac (Bernama) -- Kempen Beli
Barangan Malaysia yang menekankan aspek
patriotisme akan dihidupkan semula pada bulan
depan atau Mei, kata Menteri Perdagangan Dalam
Negeri, Koperasi dan Kepenggunaan Datuk Seri
Ismail Sabri Yaakob.  Bertemakan "Pengguna
Patriotik", kempen itu bertujuan menggalakkan
pengguna tempatan mencintai, membeli dan
menggunakan barangan buatan Malaysia, katanya
kepada pemberita selepas melancarkan telefon
bimbit model Q292 bercirikan Islam keluaran
syarikat tempatan Ad-Deen Technology Sdn Bhd di
sini.  Ismail Sabri berkata beliau berharap kempen
itu mampu mengubah sikap pengguna tempatan yang
kelihatan cenderung membeli barangan luar negara.
Katanya pengguna Malaysia seharusnya mencontohi
pengguna Korea Selatan yang begitu berbangga
membeli dan menggunakan barangan buatan negara
mereka berbanding barangan import.  "Banyak faedah
sekiranya pengguna membeli barangan Malaysia,
antaranya ia akan menghasilkan peluang pekerjaan
kepada rakyat tempatan, negara dapat mengurangkan
kos mengimpot barangan dari luar negara dan
membantu pihak industri barangan buatan tempatan,"
katanya.  Sementara itu, Pengurus Eksekutif
Ad-Deen Technology Sdn Bhd Megat Radzman Megat
Khairuddin berkata syarikat itu mensasarkan untuk
```

Fig. 5. Malay article input example.

Nevertheless, the performance of the proposed NER can further be improved by re-formulating the rules used in these experiments. Table IX shown below indicates some of the incorrect/missing annotations and partially correct annotations obtained from the experiment.

Based on the errors produced by the proposed NER, it can be concluded that the proposed set of rules produced by the Malay named-entity recognition algorithm is not complete. It is also due to the fact that the lists of words stored in the dictionaries are not complete. For instance, an organization entity “U-Mobile” is successfully annotated as an Organization entity because the company name “U-Mobile” does not have any organization prefixes or organization suffixes. Besides that, this word does not exist in any of the dictionaries. Not all of the actual named entities (NEs) are written starts with a capital letter (e.g., “i-City”). Hence, this makes the process of NER more complicated. Other than that, some of the NEs are ambiguous. For example, the word “Medan” is mostly used to refer as a location in Indonesia. However, in this experiment, the term “Medan anak Nunying” is actually identified as a person entity. Other than symbols or non-capital letters, the proposed NER should also be able to handle numbering symbols. However, there are some location entities that contain numbering symbols such as “Kampung Baru 30”. The proposed NER only manages to detect “Kampung Baru” as a location entity instead of “Kampung Baru 30”. In short, most of the annotations that are made partially correct can be solved by analyzing the present rules in more details.



```
<LOCATION>KUALA LUMPUR</LOCATION>, 15 Mac
( <ORGANIZATION>Bernama</ORGANIZATION>) -- Kempen
Beli Barangan <LOCATION>Malaysia</LOCATION> yang
menekankan aspek patriotisme akan dihidupkan
semula pada bulan depan atau Mei, kata Menteri
Perdagangan Dalam Negeri, Koperasi dan
Kepenggunaan <PERSON>Datuk Seri Ismail Sabri
Yaakob</PERSON>. Bertemakan " Pengguna Patriotik ,
" kempen itu bertujuan menggalakkan pengguna
tempatan mencintai , membeli dan menggunakan
barangan buatan <LOCATION>Malaysia</LOCATION>,
katanya kepada pemberita selepas melancarkan
telefon bimbit model Q292 bercirikan Islam
keluaran syarikat tempatan <ORGANIZATION>Ad-Deen
Technology Sdn Bhd</ORGANIZATION>di sini.
<PERSON>Ismail Sabri</PERSON>berkata beliau
berharap kempen itu mampu mengubah sikap pengguna
tempatan yang kelihatan cenderung membeli barangan
luar negara. Katanya pengguna
<LOCATION>Malaysia</LOCATION> seharusnya
mencontohi pengguna Korea Selatan yang begitu
berbangga membeli dan menggunakan barangan buatan
negara mereka berbanding barangan import . " Banyak
faedah sekiranya pengguna membeli barangan
<LOCATION>Malaysia</LOCATION>, antaranya ia akan
menghasilkan peluang pekerjaan kepada rakyat
tempatan , negara dapat mengurangkan kos mengimpot
barangan dari luar negara dan membantu pihak
industri barangan buatan tempatan", katanya.
Sementara itu , Pengurus Eksekutif
<ORGANIZATION>Ad-Deen Technology Sdn
Bhd</ORGANIZATION><PERSON>Megat Radzman Megat
Khairuddin</PERSON>berkata syarikat itu
mensasarkan untuk menjual 40,000 unit telefon
bimbit model Q292 untuk pasaran di
<LOCATION>Malaysia</LOCATION> pada tahun ini.
Model itu yang berharga RM399 seunit mempunyai
fungsi penunjuk arah kiblat , waktu sembahyang,
bacaan - bacaan Al-Quran dan Hadis dan bacaan doa.
```

Fig. 6. Malay NER annotated article.

TABLE IX: List of Errors

Word	Error
<LOCATION>Kampung Baru</LOCATION> 30	Annotation is partially correct
Emiriyah <PERSON>Arab Bersatu</PERSON>	Wrong annotation
<PERSON> Datuk Seri Najib Tun Razak Selasa</PERSON>	Annotation is partially correct
Hishammudin	Missing
i-City	Missing
<LOCATION>AS</LOCATION> 350 B3	Wrong annotation
Ameika Syarikat	Missing
<PERSON>Jade Gallery</PERSON>	Wrong annotation
Di <LOCATIO>Papan Utama</LOCATION>	Wrong annotation
U-Mobile	Missing
S.Manikavasagam	Missing
Jenny @ Jita Eyir	Missing
<LOCATION>Medan</LOCATION> anak Nunying	Wrong annotation

VI. Conclusion

This paper has proposed the first effort to generate a NER algorithm for Malay language. Based on the results obtained, the proposed Malay NER algorithm requires some adjustments for improvements in the predefined rules and also in the dictionaries used for the named-entity recognition process. The main challenge of implementing an acceptable Malay NER is to keep updating all libraries used up-to-date. There should be an effective way to ensure that the list of dictionaries used is always updated. Thus, updating the dictionaries manually is not a good option. The creation of ontology technology for semantic web usage might be helpful in producing a better list of dictionary for organization or location entities. The morphological features of Malay language are so rich and complex and this also contributes to the difficulties of implementing an effective Malay NER algorithm.

For future works, more additional rules should also be implemented and tested to handle more complex Malay sentence structure (e.g., "Jenny @ Jita Eyir"). A Malay NER algorithm should also be able to detect named entity based on existing online knowledge-based in order to produce a more robust Malay NER system. Other named entities such as time, date and percentage should also be considered in implementing a more complete and effective Malay NER system in the future.

REFERENCES

[1] E. D. Liddy, "Natural language processing," in *Encyclopedia of Library and Information Science*, 2nd Ed. NY, Marcel Decker, Inc, 2001.

[2] C. H. Chang, M. Kayed, M. R. Girgis, and K. Shaalan, "A survey of web information extraction systems," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 10, pp.1411-1428, Oct. 2006.

[3] G. Ralph and S. Beth, "Message Understanding Conference-6: A Brief History," in *Proc. the 16th International Conference on Computational Linguistics (COLING)*, 1996, vol. 1, pp. 466-471.

[4] G. Ralph, "The NYU system for MUC-6 or where's the syntax?," in *Proc. Sixth Message Understanding Conference*, MUC-6, 1995, pp. 167-175.

[5] L. F. Rau, "Extracting Company Names from Text," in *Proc. Conference on Artificial Intelligence Applications of IEEE*, 1991.

[6] ABGene. [Online]. Available: ftp://ftp.ncbi.nlm.nih.gov/pub/tanabe/AbGene/

[7] Abner. [Online]. Available: http://pages.cs.wisc.edu/~bsettles/abner/

[8] S. Yu, Y. Eunji, K, Eunju, and G. L. Gary, "POSBIOTM-NER: A machine learning approach for bio-named entity recognition," in *Proc. the EMBO Workshop on Critical Assessment of Text Mining Methods in Molecular Biology*, 2004.

[9] F. M. Naji and O. Nazlia, "Arabic named entity recognition using artificial neural network," *Journal of Computer Science*, vol. 8, issue 8, ISBN 1549-3636, Science Publications, pp. 1285-1293, 2012.

[10] A. Mansouri, L. S. Affendy, and A. Mamat, "Named Entity Recognition Approaches," *International Journal of Computer Science and Network Security*, vol. 8, no. 2, pp. 339-344, 2008.

[11] W. Takahiro, G. Robert, and W. Yoricks, "Evaluation of an algorithm for the recognition and classification of proper names," in *Proc. the 16th International Conference on Computational Linguistics (COLING)*, vol. 1, 1996, pp. 418-423.

[12] C. Micheal and S. Yoram, "Unsupervised models for named entity classification," in *Proc. the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999, pp. 100-110.

[13] J. H. Kim, I. H. Kang, and K. S. Choi, "Unsupervised name entity classification models and their ensembles," in *Proc. the 19th International Conference on Computational Linguistics (COLING)*, 2002, vol. 1, pp. 1-7.

[14] M. B. Daniel, M. Scott, S. Richard, and W. Ralph, "Nymble: a high-performace learning name-finder," in *Proc. the Fifth Conference on Applied Natural Language Processing (ANLC)*, pp. 194-201, 1997.

[15] B. Yassine, R. Paolo, and M. B. Jose, "ANERsys: An Arabic Named Entity Recognition System Based on Maximum Entropy," in *Proc. the 8th International Conference on Computational Linguistics and Intelligence Text Processing (CICLing)*, 2009, pp. 143-153.

[16] B. Frederic, N. Alexis, and G. Franck, "Tagging Unknown Proper Names using Decision Trees," in *Proc. the 38th Annual Meeting on Association for Computational Linguistics (ACL)*, 2000, pp. 77-84.

[17] Y. C. Wu, T. K. Fan, Y. S. Lee, and S. J. Yen, "Extracting named entities using support vector machines," in *Proc. Knowledge Discovery in Life Science Literature, PAKDD 2006 International Workshop, KDLL 2006*, vol. 3886, Springer Berlin Heidelberg, pp. 91-103, 2006.

[18] S. Rohini, N. Cheng and L. Wei, "A Hybrid Approach for Named Entity and Sub-Type Tagging," in *Proc. the 6th Applied Natural Language Processing Conference*, 2001, pp. 247-254.

[19] R. Alfred, A. Mujat and J. H. Orbit, "A ruled-based part of speech (rpos) tagger for malay text articles," in *Proc. the 5th Asian Conference on Intelligent Information and Database System (ACIIDS)*, vol. 2, Springer-Verlag Berlin Heidelberg, 2013, pp. 50-59.

[20] B. Indra, B. Stephane, W. Gatot, A. H. Zainal, and A. A. N. Bobby, "Named entity recognition for the indonesian language: combining contextual," in *Morphological and Part-of-Speech Features into a Knowledge Engineering Approach*, A. Hoffman, H. Motoda and T. Scheffer, Eds. DS 2005, LNAI 3735, Springer Berlin Heidelberg, 2005, pp. 57-69.

[21] S. F. Yong, R. M. Bali, and Y. W. Alvin, "NERSIL: the named-entity recognition system for iban language," PACLIC, pp. 549-558, 2011.

[22] M. Ashraef, N. Omar, and M. Albared, "Arabic named entity recognition in crime documents," *Journal of Theoretical and Applied Information Technology*, vol. 44, no. 1, pp. 1-6, 2012.

[23] E. Ferreira, J. Balsa and A. Branco, "Combining rule-based and statistical methods for named entity recognition in portuguese," presented at Actas da 5 ^{a} Workshop em Tecnologias da Informa ção e da Linguagem Humana, 2007.

Rayner Alfred was born in Kota Kinabalu, Sabah. He completed a PhD in 2008 looking at intelligent techniques to model and optimize the complex, dynamic and distributed processes of knowledge discovery for structured and unstructured data. He holds a PhD degree in computer science from York University (United Kingdom), a Master degree in computer science from Western Michigan University, Kalamazoo (USA) and a Computer Science degree from Polytechnic University of Brooklyn, New York (USA).

Dr. Rayner leads and defines projects around knowledge discovery and information retrieval at Universiti Malaysia Sabah. One focus of Dr. Rayner's work is to build smarter mechanism that enables knowledge discovery in relational databases. His work addresses the challenges related to big data problem: How can we create and apply smarter collaborative knowledge discovery technologies that cope with the big data problem. He is a member of the Institute of Electrical and Electronic Engineers (IEEE) and Association for Computing Machinery (ACM) societies.

Leow Chin Leong is currently pursuing his master degree in computer science with the Center of Excellent in Semantic Agents under School of Engineering and Information Technology, in Universiti of Malaysia Sabah, Sabah, Malaysia. The author's research interests include text mining, natural language processing and information retrieval and extraction.

Chin Kim On received his PhD in artificial intelligence with the Universiti of Malaysia Sabah, Sabah, Malaysia. The author's research interests included gaming AI, evolutionary computing, evolutionary robotics, neural networks, image processing, semantics based visual information retrieval, agent technologies, evolutionary data mining and biometric security system with mainly focused on fingerprint and voice recognition.

Patricia Anthony received her PhD in computer science from the University of Southampton in 2003. She is currently working as a senior lecturer at the Department of Applied Computing, Lincoln University, New Zealand. Her research interest is in semantic agents and multi-agent systems and how these agents can interact with each other within an open domain to solve problems. She is also interested in investigating how agents can communicate with each other at the semantic level using semantic technology. To date, she has published more than 80 articles in the forms of journals, book chapters and conference proceedings. She is a member of IEEE, ACM and IACSIT.