A Concise Review of Named Entity Recognition System: Methods and Features

M. Ikhwan Syafiq¹, M. Shukor Talib², Naomie Salim³, Habibollah Haron⁴, and Razana Alwee⁵

¹,²,³,⁴,⁵School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia, 81310, Johor Bahru, Johor, Malaysia.

E-mail: shukor@utm.my

Abstract. Named Entity Recognition (NER) is an elementary tool for all application areas in Natural Language Processing (NLP) such as Automatic Summarization, Information Extraction, Information Retrieval, Text Mining, Machine Translation, Question Answering, and Genetics. NER is a task to discover and categorises the named entities (‘atomic elements’) in the text into predefined classes such as the names of persons, organizations, locations, terminologies of time, quantity and etc. Different languages may have different morphologies and thus involve dissimilar NER procedures. For example, an Arabic NER system cannot be practically used in processing Malay texts due to the different morphological features. The morphological features of every language are rich and complex and donates to the difficulties of implementing an actual method to develop the accurate NER system. In this paper, we review on three main techniques that commonly used to develop an NER system well-known as Rule-Based, Machine Learning, and Hybrid approach. This paper also highlights the features of each technique.

1 Introduction
Access to information extraction (IE) is a process to obtain relevant information from unstructured sources whether using classification or bridging methods. To get structured information on the article, the information process needs to through several phases, namely pre-processing of data, extraction and classification. One of the most widely used extraction techniques is identifying Names Entity (NE) based on unstructured text in the articles. However, texts are hard to understand by computers. A machine will not be capable to understand the content of texts. At the moment, an enormous capacity of documents can be easily regained and extracted...
from internet. But, it would proceed a long time for people to physically process the documents in a short period time and the procedure of annotating texts manually frequently provides a partial results [1].

With a rapid growth of NER system, the capacity of data in many languages is accessible on the web, this make the procedure of extracting and classifying the NE is very useful [2]. However, domain specific, for example, DNA, biochemical, and etc., might not be valid for identifying named entities on other particular domains for example food guides. Also, identifying the categories of named entity for texts written in English language not applicable for other languages. Typically in the English language have proper nouns that normally begin with a capital letter. This proper nouns always been used to signify a unique named entity such as location, organization and person. Besides that, such methods might not be valid to be applied to texts written in Chinese and Arabic language as it does not hold capital letter that can be led to identify the NE. This is because most languages have rich morphological features. In short, the development of the NER system be subject upon on the domain of studies and also the type of languages used.

1.1 Named Entity Recognition

The word “Named Entity”, nowadays extensively used in NLP field, was invented for the initial phase in the Sixth MUC conference (MUC-6) [3]. In other words, information extraction is a systematic procedure to solve the problem in the NLP area. The confusion of NE created the named entity system very challenging and complicated, and most of the NER fail to reach human level performance. Although, extracting information, for example classifying relationships between a single type of named entity; first need to require the extraction of entity’s name in the text i.e., NER task. This process is important and its accomplishment defines the effectiveness of other information extraction tasks. NER focuses to detecting the rations of the text that state to specific entities. Due to the complexity of NER system, exists several steps to implement NER system;

i. **Pre-processing step**: To detect the boundaries of named entity in the text by tokenize and splitting the sentences. Larger chunks of article can be tokenized or splitting into one sentence and one sentence can be tokenized into a single word [4].

ii. **Feature processing step**: NLP has many complexities in language completion, generating a design to match the potential linguistic understandings of the single word facts involves the pre-processing on physical input, for example allocating Parts-of Speech (PoS) tagging and morphological features.

iii. **Named entity recognition step**: This step is to identify the NE and allocates it to a class or groups of entity.

iv. **Solving normalization or recognition errors step**: Normalization mapping NE and place all words on the same footing, and keep consistently [4]. This step to help in solving problems occur from distinctions in the synonymous word as well as the confusing abbreviations.

2 Methods
2.1 Rule Based Approach

Rule based are a method that relies on either heuristic or regular expression rules to classify the named entity. Also, rule based used in combination of language resources such as dictionaries, gazetteer lists, or list markers to classify the entity’s name. This approach recognizes the NE by consuming a conservative of procedures and a list of dictionaries that are made by hand predefined by humans. The system using a rule based approach spreads over a customary of rules in order to extract design. The designs are generally made up from linguistic, syntactic and orthographic patterns. Also, rule-based approach is recognized in the form of either determinate state regular expressions or transducers. The followings are the main features and methods that can be performed in rule-based NER systems. Table 1 below illustrated the common features used in rule-based NER system.

| Feature            | Feature definition                                                                 |
|--------------------|------------------------------------------------------------------------------------|
| Gazetteer          | A gazetteer is a list of distinct NE, and it holds a particular list of named for the specific category of named entity class. For instance, Kuala Lumpur is in location class while Mansor is in a person names class with a gazetteer list. Gazetteers also made as dictionaries or whitelists word [5]. |
| Blacklist (Filter) | Filters or blacklist is used as a rejection element for words or string of words that are acknowledged to be an unacceptable name entity and to refuse such incorrect named entity which could not be sentences/words. For instance, (A Singapore prime minister Lee Hsien Loong), here the phrase “prime minister” is unacceptable. |
| Trigger Words (Keywords) | Keywords or trigger words are regular words that usually come after or before named entity. These trigger words are used to recognize a named entity and can gross many forms such as a noun list or verb list, for instance word (drive) [6]. |
2.2 Machine Learning Technique

Machine learning technique (ML) in the viewpoint of text analytics is an established of statistical approaches for categorizing some characteristic of a document, for example, sentimentality, PoS tag, named entity, and etc. [7]. In NER system using machine learning techniques, there are several models and features that are used to identify the named entity. Features of ML NER system are important material or descriptors for words characteristics. Also, feature engineering is a leading critical task of NER for entirely classifiers. The common features of NER system using ML technique are illustrated in Table 2.

| Feature | Feature definition |
|---------|--------------------|
| Word Length | Word Length can be used to determine NE, for example, if the word less than three or two words, it is not considered as NE. |
| Special Marker | To recognize some special marker/symbols in the text [8]. |
| Word Suffix/Prefix | This feature used for pattern identical to capture word suffix and prefix of word length up to N [9]. However, prefix/suffix rarely comes as a named entity. |
| Stop word | This feature removes common words that cannot be part of NE. |
| Part-of-Speech (PoS) | PoS tagging features can be used to recognize part-of-speech class words (e.g. proper nouns, nouns, verbs, etc.). |
| Morphology-based | The languages extracted from a group of morphology-based features. This feature is significant and used extensively in NER. |
| Infrequent word | Infrequent words are gained by scheming the word frequency in the used corpus during the training segment and then choosing the cut-off frequency to form the binary method. |
| Word-Left/Right | Word-Left/Right feature used to investigate neighbour words (left/right) of length up to N, for example, NE tags or PoS tags from NER system that used as a feature [7]. |

ML technique can be specified as a standard that is well-designed to other text and article named as supervised approach, or probably will be a set of method that perform through a huge collection of files to annotated meaning that named unsupervised approach. Feature engineering in ML technique is a leading critical task of named entity recognition system for words categorizations [7]. Also, ML technique is more accomplished when compared to rule based approaches because the structure can be trained and also can perform in various domains. Machine learning can be allocated into three well-known methods: unsupervised learning, semi-supervised learning and supervised learning. The most regularly implemented are supervised learning (SL) method which characterize the NER problem as a
grouping task and involve the accessibility of huge annotated datasets. Table 3 illustrated the main model in ML NER systems.

Table 3. Common model in machine learning NER system

| Learning Method   | Model                          |
|-------------------|--------------------------------|
| Supervised        | Conditional Random Field (CFR)  |
|                   | Hidden Markov Model (HMM)       |
|                   | Decision Tree (DT)              |
|                   | Maximum Entropy (ME)            |
|                   | Support Vector Machine (SVM)    |
|                   | Artificial Neural Networks (ANNs)|
| Unsupervised-Learning | Clustering                   |
| Semi-Supervised   | Bootstrapping                  |

The machine learning NER system aims to transform the identification problem into classification problem. In principle, the ML system identify and recognize NE into detailed NE classes such as persons, organization, locations, and etc. [10]. Most recent studies in NE for all major languages use a ML technique, also called statistical approach. Machine learning also have been widely used to identify named entity classes from annotated texts.

### 2.3 Hybrid approach

The hybrid approach is a technique used more than one named entity recognition approach in order to develop a good characteristics from each method. The term hybrid is concerned with the combination of rule based approach and dictionary-based approaches, this should not be confused with the hybrid technique used when two or more machine learning algorithms are combined to build the classification model. Such technique will produce an improved result and outcome [11]. On the other hand, the weaknesses of the rule based are still unescapable in the hybrid classification. A domain specific named entity system may possibly modify the usual of rules used to identify dissimilar forms of NE when the domain of trainings is reformed [12]. The way of the processing system may start from the rule-based approach to the machine learning approach or vice versa. Therefore, to solve the problem of identifying a precise entity terms, the rule based approach can be used to addressing the problem of overlapping entities, while, the machine learning techniques will measure strengthen of the co-occurrences between multi-word entities [13]. Hybrid NER systems also can be implemented using Local Grammars (LGs) that representing the contextual rules in linguistics approach by combining with any machine learning models [14]. This hybrid method seeks to discover the potential of the linguistics approaches to a specific type of language in NER systems.

### 3 Conclusions
NER is considered to be the most important task for many information extraction systems. Based on review above, we can conclude that the NER has been growing rapidly with various techniques and features. Many researchers have been proposed NER system for various domains using established methods such as rule-based, ML, and hybrid approaches, however, more investigation efforts are still needed to implement because we have a lack of resources on particular languages such as, Indian, Chinese, Malay, and many other languages, also on the specific domain, such as biomedical, DNA, criminal, food guide and etc. Hence, there is a necessity for researchers to employ a new brand handcrafted method that can utilize and perform for many languages and domains in the NER field.

Acknowledgements
This work was financially supported by a Tier 1 grant, reference number (PY/2017/01337), and Faculty of Engineering, Universiti Teknologi Malaysia (UTM), Skudai, Johor.

References
1. R. Alfred et al 2014 Int. J. Mach. Learn. Comput. 4 300.
2. Ajees A. P. et al 2018 Int. Conf. Data Sci. Eng. 1.
3. K. B. Cohen et al. 2017 BMC Bioinformatics 18 1.
4. S. Eltyeb et al 2014 J. Cheminform. 6 1.
5. R. E. Salah et al 2017 Int. J. Adv. Sci. Eng. Inf. Tech. 7 815.
6. W. Zaghouani et al 2010 Adapting a resource-light highly multilingual 563.
7. R. E. Salah et al 2017 Int. J. Adv. Sci. Eng. Inf. Tech. 7 511.
8. A. Abdul-hamid et al 2010 Proc. 2010 Named Entities Work 110.
9. A. F. Wicaksono et al 2010 Prooc. 4th Int. MALINDO (Malay Indones. Lang. Work. 1.
10. T. Dasgupta et al 2017 Proc. Conf. Web Intell. 541.
11. M. Basaldella et al 2017 J. Biomed. Semantics 8 0.
12. M. Alruily et al 2009 Proc. - Int. Conf. Dev. eSystems Eng. DeSE 2009 153.
13. S. S. Balgasem et al 2018 Proc. 2017 6th Int. Conf. Electr. Eng. Informatics Sustain. Soc. Through Digit. Innov. 1.
14. J. P. C. Pirovani et al 2015 Adv. Intell. Syst. Comput. 4452.