Journal of ICT in Education (JICTIE)
ISSN 2289-7844 / Vol. 2 / 2015 / 1 – 14

Malay Named Entity Recognition: A Review

Farid Morsidi[1], Sulaiman Sarkawi[2], Suliana Sulaiman[2], Siti Asma Mohammad[2], Rohaizah Abdul Wahid[2]
[1]Faculty of Art, Computing & Creative Industry, Universiti Pendidikan Sultan Idris;
[2]Computing Department, Faculty of Art, Computing & Creative Industry, Universiti Pendidikan Sultan Idris
farid_m2mfan@yahoo.com.my[1], {sulaiman, suliana, siti_asma, rohaizah}@fskik.upsi.edu.my[2]

Abstract

The Named Entity Recognition (NER) field had been thriving for more than 15 years. NER could be defined as a process that recognizes named entities, such as the names of persons, organizations, locations, times, and quantities. The research field of NER generally emphasizes on the extraction and classification of mentions for rigid designators. This ranged from text, such as proper names, biological species, temporal expressions, and so on. NER has been utilized in many sectors, for example ranging from inquiries to morphological syntax, besides information extraction. However, most of the work had been delegated on limited domains and textual genres such as news articles and web pages. Techniques used during the processing of English text cannot be used to process Malay-related terminology. This is due to the different morphological usage of a particular language. Finding co-references and aliases in a text can be reduced to the same problem of finding all occurrences of an entity in a document. This paper proposes approaches that have been applied in the fields of NER that is in Malay, or partially related to it, in order to detect proper nouns within Malay documents. This paper also discusses the various researches done in an effort to produce high-quality training data for Malay corpus via appropriate NER algorithms and methods aside from highlighting the key points needed in improving the current NER studies.

Keywords Named entity recognition, natural language processing, Malay, fuzzy rule-based, information retrieval (IR), information extraction (IE), artificial intelligence, fuzzy relational calculus

INTRODUCTION

Named Entity Recognition, or abbreviated as NER is a significant field constituting the field of Natural Language Processing (NLP). The term "Named Entity" is proposed during the Sixth Message Understanding Conference

(MUC-6) (Alfred et. al., 2014). During those times, the Message Understanding Conference (MUC) is emphasized on Information Extraction (IE) tasks, where structured information from various company activities besides defence-related activities is extracted from unstructured text. These unstructured texts were derived from credible academic sources, such as newspaper articles. During the process of categorizing the task, individuals noticed that it is important to categorize information units such as names, including person, organization and location names, and numerical expressions including time, date, money, and percent expressions (Soo-Fong et. al., 2011). The identification reference to these entities in text was recognized as "Named Entity Recognition and Classification (NERC)".

The synonymous field of NER and NLP are associated with each other due to its nature of language process analysis. NER is important for NLP in terms of document retrieval, morphological analysis, and information extraction (Ananiadou et. al., 2010; Rayner et. al., 2014). As performed by Ruch in Turian, Ratinov & Bengio (2010), NLP tools can be used to improve the processing of clinical records. In clinical search, a named entity extractor was used to identify between patient's and physician's names. Within Information Retrieval (IR), NER improved the detection of relevant documents (Wang, Liu & Sun, 2012). Related researches done to recognize named entities in other languages include English (Nothman et. al., 2013), Chinese (Soo-Fong et. al., 2011), Arabic, Indonesian, and Indian (Oudah & Shaalan, 2012). A good proportion of work for NER research area was dedicated towards the study of English but a larger section of it addresses language independence and multilingualism problems. An example of this includes Chinese, which is studied in an abundant literature (Gifu & Vasilache, 2014; Kral, 2014; Soo-Fong et. al., 2011), and Arabic (Abdul-Hamid & Darwish, 2010; Rayner, Mujat & Orbit, 2013) has started to receive a lot of attention during large-scale projects such as Global Autonomous Language Exploitation (GALE). Rayner et. al. (2014) utilized a rule-based approach for the purpose of developing Malay NER. For almost all conventional NER research, several key elements were identified as primary information units such as names, including person, organization and location names, as well as numeric expressions including time, date, money, and percentage expressions (Ananiadou et. al., 2010). For Rayner's research, several dictionaries were used to handle named entities such as Person, Location, and Organization. Correct rules need to be applied in order to obtain a reasonable output. All dictionaries also must be up to date so that precise results could be achieved (Cao, Tang & Chau, 2012).

An important question at the inception of the Named Entity Recognition task was whether machine learning techniques were important at all, aside from the relevance of simple dictionary lookup would be suffice in order to produce optimal performance (Don, 2010). For some situations, a small gazette is enough to provide good accuracy and recall values (Powers, 2011). These data that were accumulated via pre-processing methods such as web scraping and crawling were stored and categorized according to their ontological

features, namely their aliases, typing, identifiers, noun structures, and keyword significance (Soo-Fong et. al., 2011). Named entities in a document, be it on the web content, articles, and text corpora, possess lexical characteristics known as word. The main categories in NER include *people, organization,* and *location*. Although there exist several minor categories that had been the accumulating factor of varsities of word to be gathered into one equal group, these categories remain under the shadow of the three major categories. Another factor that supports the implementation of Named Entity Recognition into data clustering is because fundamental clustering procedure does not apply the fact that an entity may exist under different aliases or terminologies (Carvalho, Batista & Coheur, 2012). The implementation of Natural Language Processing (NLP) algorithms into enhancing the processing of text chunks is normally affected by the domain of the studies (Don, 2010). Different language requires the proficiency of varied language in order to identify the named entity. Up until now, there exist several credible NER systems for major language typing in the world including English, Arabic, German, Chinese, Hindu, Indonesian, and so on. However, there is no credible system that would improve the NER detection of nouns in Malay language (Abu Bakar et. al., 2013; Rayner et. al., 2014; Rayner, Mujat & Orbit; 2013). A research that came close to this purpose is done by Rayner, Mujat & Orbit (2013) and Rayner et. al. (2014) that detect the words, punctuation, and numbering schemas in Malay using rule-based NER algorithm. A few researches had indicated the proficiency of applying ontology to improve the performance of document clustering as expert systems is the conceptualization for a domain into an individually-identifiable format, along with the visualization of machine-readable format containing entities, attributes, relationships, and axioms (Abdul-Hamid & Darwish, 2010; Elsebai, 2009).

SCOPE OF THE PROBLEM

After the identification of the problems endured in major linguistic NER is highlighted, the solution related to the language category which is desired to be classified could be done more appropriately. The nature of NER research generally carries out named entity sorting into the major entity constitution: *person, location,* and *organization* (Elsayed & Elghazaly, 2015; Gifu & Vasilache, 2014). Therefore, particular stages of development could be devised in order to delegate relative named entities according to their own clusters and categories.

Firstly, NER research attempts to obtain the surface form or meaning in the particular context as the primary classification criterion (Ismail, 2013). Varied definition of existing words affects its level of redundancy in a particular corpus domain (Ananiadou et. al., 2010; Liao, 2011; Patrick & Nguyen, 2011; Turian, Ratinov & Bengio, 2010). As the degree of general context for its normal usage in the language itself possesses a high value, the word would have many clues and relations to the other word entities for that domain. This aspect is shared among the major entity categorization. For example, "Sarawak" could be

applied in a geographical context, and it could also pinpoint to the government of Sarawak, which is placed under the *organization* entity categorization. In the effort to minimize the ambiguity, NER research attempts to harvest as much as possible the surface form of the word as the primary clue (Nanda, 2014; Turian, Ratinov & Bengio, 2010). The logic behind this approach is that most of the current NE taggers did not use semantic information, and disambiguation at the named entity tagger level is almost impossible (Turian, Ratinov & Bengio; 2010). There also existed ambiguity definition for each Named Entity categories, making the effort of categorizing words with similar definition context into a precise cluster group difficult (Turian, Ratinov & Bengio; 2010).

It is almost impossible to proclaim the similarity of a single word compared to others, even if it is under the same linguistic category of sharing the similar definition or syntax structure with others (Turian, Ratinov & Bengio; 2010). As such, every word structure has its own rule to set up a sentence that would deliver the appropriate hierarchy of definition and appropriate contextual usage. Named entity in general illustrates the proper naming and numeric expression of words, in the sense of capitalized form or the number of thing respectively (Zamin et. al., 2012). This context, however, changes when another word category is added. For example, is the name of cars such as *"Iris"* or *"Axia"* categorized as a proper noun? For the NER research, it is an important point to include them as named entity to assist natural language processing (NLP) applications such as *Information Extraction (IE)* and *Information Retrieval (IR).* However, upon careful analysis, *"Iris"* and *"Axia"* are considered names of classes and not the names of specific cars. This is different from the problem that more than one person could have the same name, as the nature of the name "Iris" relates to some concept, "class". A particular class' categorization is usually decided upon the traits that it possesses to make up the distinctive naming scheme, for example colour, type, and material name (Gifu & Vasilache, 2014; Turian, Ratinov & Bengio, 2010). However, ordinary words such as *"air mata"*, *"tempayan"*, and *"emosi"* were not included as named entities. Some arbitrary decision is necessary to determine the classification of these words, as there is no particular rule that binds their characteristics to any available classes expressions (Soo-Fong et. al., 2011). Another NER trait that dampens the appropriate division of hierarchy for the major NE types lies in the arbitrary degree of fineness of each level of datasets (Ismail, 2013; Kanagavalli & Raja, 2010). Flexibility is considered to enable the named entities adaptability to as many possible definitions as possible. For example, the "ORGANIZATION" class could be delegated via different criteria, such as occupations, number of workers, job range, and so on. NER research neglects the need of dividing a class into more diverse branches should the division depend primarily on the context rather than the definition in the name itself (Patrick & Nguyen, 2011; Zamin et. al., 2012). The following table illustrates the analysis of the problems that arise for Malay corpus NER.

Table 1 Main factors surrounding the lack of Malay corpus resources and named entity recognition

Problems	Description
The lack of abundance of annotated corpus to be implemented into the existence of a Malay Named Entity Recognition (NER) system.	Domain specific NER application is only reserved for that particular language in context, and mostly is not applicable towards other languages with versified morphological and syntax constitutions (Rayner et. al. (2014). For example, biology-related NER system such as AbGene and AbNER would not perform well in processing military articles due to its nature that is reserved for different domains. This fact is contributed by the nature of most languages that differ morphologically.
The implication of supervised or unsupervised learning process to influence the efficiency of a search algorithm to be applied into the data clustering procedures for Malay proper nouns.	Pre-processing involves the access of a collection of annotated corpus, memorizing lists of entities, and division of distinctive disambiguation rule based on specific features (Gifu & Vasilache, 2014). However, it requires a proper training session for the user and system to remember and recall data. The improvisation of semi-supervised learning involves the re-construction of statistical inferences from training data (Turian, Ratinov & Bengio, 2010). Semi-supervised learning is a different approach from supervised learning, as it involves only a minute degree of supervision for the learning process to be initiated (Turian, Ratinov & Bengio, 2010), in which the latter needs a start-to-finish attention.
Suitable programming paradigm that could be applied for the creation of a better dictionary list/ model for Malay-related categories of Named Entity Recognition.	Programming paradigm is an element incorporated in the programming language in order to initialize the workflow of a coding sequence. System performance is shown to be comparable to that of a simple supervised learning-based system (Turian, Ratinov & Bengio, 2010). Selection of appropriate programming paradigm that uses information retrieval measures could assist in manifesting the most optimum quality of named entity recognition.

RELATED WORK

Even though there had been several rule-based NER systems for Malay language, the lack of Malay resources is still at large. Among the identified linguistic NER research that had the closest similarity to Malay NER is Arabic (Naji & Omar, 2012; (Oudah & Shaalan, 2012) and Chinese NER (Zamin et. al., 2012; Liao, 2011). Both of these languages have their words composed in unique characters rather than the conventional alphabetic schematic ruling that had been encased in a majority of the world's languages.

Abdul-Hamid & Darwish (2010) had conducted a research where they placed aside the common necessity of morphological or syntactic analysis or gazetteers in an effort to identify characters written in Arabic. Oudah & Shaalan also have devised a rule-based approach in order to develop a Named Entity Recognition System for Arabic language, abbreviated as NERA. Benajiba, Rosso & Ruiz (2007) had successfully presented an NER system dedicated for Arabic text structures based on *Maximum Entropy* (ME) implemented on the construction of their own training and test corpora (ANERcorp) and gazetteers (ANERgazet) for the purpose of evaluating and presenting their own system. Kanagavalli & Raja (2010) had proposed the idea of using neuro-fuzzy reinforcement learning for classifying the spatial descriptors and expressions from free text documents. Cao, Tang & Chau (2012) provided fuzzy named entity-based document clustering based on conventional keyword-based document. An innovated named entity annotation via the accumulation of sentences from the web containing gazetteer entities, thus producing a 1.8 million word Korean corpus that gave similar results to manually annotated data (Nothman et. al., 2013). The approach of using semi-supervised NER classification is indicated in a research done where NER is performed on the MUC-7 corpus accompanied with supervision at a minimal rate. This study embeds a short list of names for each NE type, performing 16% lower than other systems that embed state-of-the-art approach in the MUC-7 evaluation.

Zamin et. al. (2012) proposed an unsupervised technique using bit text mapping as a tool to tag Malay text. The approach includes the translation of texts into a resource-rich language along with dictionary lookup. One point to be taken note of for the research inducted for Malay noun is its lack of resources, especially in terms of annotated data (Abu Bakar et. al., 2013; Aboaoga & Aziz, 2013). Implementation of semi-supervised technique shows an optimum performance should the sentence pair exist at an appropriate quantity; in other words, when there is no data sparsity problem. Zamin et. al. (2012) indicated a suggestion of improvement in her research where the consideration of tag sets need to be issued so as to increase the performance and consistency of languages targeted for that particular domain.

Approaches in NER

Prior research indicated that learning methods in NER field could be categorized into four different systems: supervised systems, semi-supervised systems, bootstrapping, along with unsupervised system (Liao, 2011; Naji &

Omar, 2012). The field of NER diverges the learning approach of categorized dataset into the automated *supervised learning* process, besides leaning towards user participation type of *semi-supervised* approach (Liao, 2011; Naji & Omar, 2012; Turian, Ratinov & Bengio, 2010). For expert system that involves more human participation in data classification, supervised technique is placed into priority (Abu Bakar et. al., 2013; Ananiadou et. al., 2010; Zhan & Sun, 2011). However, contemporary rule-based expert system practices semi-supervised approach, where the machine's effort to assimilate the identification of new, unsorted dataset is placed under joint venture with human's involvement in the classification of named entity (Cao, Tang & Chau, 2012).

A huge advantage of this process over supervised classification lies in its ability to compensate the lack of information from labeled examples by information extracted from a large set of unlabeled examples (Abu Bakar et. al., 2013; Ananiadou et. al., 2010; Rayner, Mujat & Orbit, 2013). In Malay language context, studies conducted by Mohamed, Omar & Aziz (2011) and Don (2010) had become the pioneer of word classification. Mohamed, Omar & Aziz (2011) had implemented a modified Hidden Markov Model (HMM) approach with Malay morphological information as an emphasis for the learning of word classification. This study utilized a number of 18,135 corpus data which contain a total of 1381 ambiguous words. It was proven that the classification accuracy had achieved up to 67.9% for a total of 21 tags on appropriate datasets. Zamin et. al. (2012) proposed an unsupervised technique using bit text mapping as a tool to tag Malay text. Results from the study indicated a 76% percent rate in precision and 67% in recall.

Supervised Learning

Supervised training is deduced on the ability to investigate the features of positive and negative examples for named entity over a huge accumulation of annotated documents and design rules which captures the instances of a given type (Kanagavalli & Raja, 2010). Due to the incorporation of human and computer interaction in the aspects of dataset collection and analysis, supervised learning approach includes the necessity of the presence for a large annotated corpus (Senthil, Thangmani, & Zubair; 2014). The absence or lack of availability for said resources, along with the restrictions of cost in creating them, leads to the proposal of an alternative solution to the fundamentals in NER research: semi-supervised learning and unsupervised learning (Rayner et. al., 2014). For supervised systems, mostly machine learning approach was used, thus reducing the participation of humans in the matter of data clustering and categorization (Turian, Ratinov & Bengio, 2010). The data as submitted to the machine learning algorithm are fully labeled, either manually tagged or distinctively annotated by a lexicon. Among the approaches available to automatically classify words that could be used include rule-based methods, probability-based, and transformation-based (Abu Bakar et. al., 2013; Ananiadou et. al., 2010; Rayner, Mujat & Orbit, 2013).

Semi-Supervised Learning

The goal of resolving problems arising in linguistic domains for the field of research in NER includes the necessities of overcoming costly corpus annotation (Rayner et. al., 2014). This includes the automatic creation of a standardized corpora and semi-supervised methods (Elsayed & Elghazaly, 2015). Semi-supervised learning technique involves the incorporation of already existing data to be used as the rule to annotate raw, uncategorized dataset from a given domain (Don, 2010; Nanda, 2014; Soo-Fong et. al., 2011). The main method utilized for the semi-supervised technique is known as "bootstrapping" and inquires the minute degree of supervision to initiate the learning process, for example the usage of seed data (Turian, Ratinov & Bengio, 2010; Wang, Liu & Sun, 2012). The most frequent contexts found are used as a set of contextual rule. This particular rule could be utilized to discover other relevant rules within a similar domain (Turian, Ratinov & Bengio; 2010). Urbansky had also devised a system to learn NER from fragmentary training instances on the web (Nothman et. al., 2013). Their evaluation on English CoNLL-03 data had illustrated an F-score 27% lower via automatic training than the similar system trained on CoNLL training data. Among the features that were managed to be identified by the incorporation of knowledge from unlabeled text include the highly-predictive criterion from related tasks [20], output selection for a supervised system (Naji & Omar, 2012), jointly modeling labeled and unlabeled, partially-labeled (Zhan & Sun, 2011) language, and induced word class features (Nothman et. al., 2013).

Rule-based

Algorithms for NER systems could be divided into a few categories in order of automated procedures to data harvesting that involves less participation from user: rule-based, machine learning, and hybrid (Abu Bakar et. al., 2013; Ananiadou et. al., 2010). Rule-based algorithms were induced as a replacement for the lack of legitimate corpus resources for the particular language (Aboaoga & Aziz, 2013; Ananiadou et. al., 2010). Rule-based NER algorithm performs detection of named entity via the manual implementation of a set of rules aside from data mining tools such as dictionary and gazetteers that are predefined by humans (Ananiadou et. al., 2010; Rayner, Mujat & Orbit, 2013). Raw, un-annotated datasets are applied a set of rules so that a distinctive pattern could be extracted (Rayner, Mujat & Orbit, 2013). These rules are derived based on the key named entities element to establish a legitimate pattern base, for example names for location, organization, and person. The formation of word patterns were retrieved based on the morphology and syntax for a particular linguistic structure, which includes grammatical, syntactic, and orthographic features (Gifu & Vasilache, 2014). Dictionaries and gazetteers were molded from fundamental features contained in word groups, such as common first names and titles, list of countries, major cities, and companies (Nand, 2014).

The reason behind the selection of rule-based NER approach depends on the factor of abundance for the existing language to be used as training data (Powers, 2011). Rule-based measure is used to complement the corpus deficiency for language classes that is still lacking in tagged resources. Among the researches that are seen to be evident to the tagging of lexical features in language elements include Rayner (2013) for Malay NER, Srinivasa for Indonesian text collection (Abu Bakar et. al., 2013), Zamin et. al. (2012) with bitext mapping for Malay word tagging, Soo-Fong et. al. (2011) who devised a named entity recognition system for Iban language, and Juhaida for Malay and Jawi word assignment to related classes (Abu Bakar et. al., 2013). Specific rule would be applied toward the current word to determine whether it is an entity or not. These rules were designed to indicate the three major types of named entities available, including *person, organization,* and *location* (Naji & Omar, 2012).

NER Evaluation

This section addresses briefly the fundamental formula used to evaluate and analyze the outputs produced from NER research, reviewed from various existing related researches. Almost all NER research involves the evaluation of datasets via three determining factors: *precision rate*, *recall rate*, and *F-measure* (Ismail, 2013; Kral, 2014). Fundamentally, there should be a minimum of two cluster problems emulated by every dataset. Via these clusters, the approximate value of learning rate would be evaluated and analyzed. Along with the two correct situations, where both class and cluster labels were either A or B, four possibilities could be devised (Powers, 2011). As the most common clustering and classification problems include two classes, they are usually called positive and negative (Cao, Tang & Chau, 2012; Ismail, 2013). The cluster-to-cluster evaluation assigns to each document one of the including values:

True positive (TP)	:	values categorized as actual positive, and predicted as positive
False positive (FP)	:	actual negative values, but predicted as positive
True negative (TN)	:	values indicated as negative, and predicted a negative
False negative (FN)	:	actual positive values, but predicted as negative

The most convenient approach of indicating the error structure in an NER cluster is to include the number of documents falling in each of the categories for the value above into a matrix, known as the contingency table (also known as the *confusion matrix*).

Table 2 Confusion Matrix between cluster labels TP (*true positive*), FP (*false positive*), TN (*true negative*), and FN (*false negative*)

Actual (Classes)	Predicted (Clusters)	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

Precision and Recall

Manning and Schutze in Ismail (2013) defined the Precision value in Information Retrieval field to be *"a measure of selected items that the system got right"*, whilst the Recall value is derived as *"the proportion of the target items that the system selected"*. The Precision value, *p* could be determined where *t*p and *f*p represent the value for true positive and false positive respectively (Kral, 2014; Nothman et. al., 2013). The Recall value, *r* can be measured as shown in Figure 1.

$$\text{Precision, } p = \frac{tp}{tp+fp} \qquad \text{Recall, } r = \frac{tp}{tp+fn}$$

Figure 1 The Precision value (*p*) and the Recall value (*r*)

From the results obtained, a *contingency matrix* representing the concepts used during the evaluation of the precision and the recall rate could be devised (Liao, 2011; Nothman et. al., 2013). A confusion matrix, as shown in Figure 2, could be constructed with more than two classes and clusters by using more rows for the classes and more columns for the clusters.

System	Actual	
	Target	Target
Selected	***t*p**	***f*p**
Selected	***f*n**	***t*n**

Where *f*n represents the number of cases the system failed to take the target item into account, which is also identified as the *false negative*.

Figure 2 A Confusion Matrix (from Manning and Schutze (2002) in Ismail (2013)

F-Measure

The F_1 score, or relatively the F-measure, is the evaluation method that combines precision and recall values for overall performance. This measure is applied to resolve the dominant trade-off issue in between precision and recall values (Nothman et. al., 2013). This is due to the fact that the values of both precision and recall rate were not concluded as a fixed variable; however, it might be possible to change them according to the output. For example, all items in the dataset could be selected so that 100% recall rate would be obtained, although causing a drop in precision value. The F_1 score is determined as shown in Figure 3:

$$F_1 = \frac{1}{\alpha \frac{1}{p} + (1 - \alpha) \frac{1}{r}}$$

Where α is a factor that decides the weighting for precision and recall. The value for α is set to 0.5 to obtain equal weighting of p and r. For α value range placed at 0.5, the measure is simplified as follows.

$$F_1 = \frac{2pr}{(p+r)}$$

Figure 3 The F_1 score

When expanded, the calculation of both values of p and r could be expanded into the formula as shown in Figure 4.

$$F(i,j) = \frac{2P(i,j)R(i,j)}{P(i,j) + R(i,j)}$$

Figure 4 Expanding both p and r values

The scores obtained via F_1 possess a similarity to accuracy measures, as they are both prone to certain cases encountered by the system. In the Information Retrieval field of research, the F_1 measure is sensitive to the numbers of correct cases whilst the accuracy is only prone to sensitivity induced by the number of errors. Gaussier in Liao (2011) insisted that the rank of the candidates is an essential trait. Sorting and ranking of the candidate group according to their respective scores are vital if any of the evaluation measures were to be performed.

FUTURE DIRECTION AND CONCLUSION

The paper attempts to provide a general review on the approaches already conducted for Named Entity Recognition that could be improvised for a language group that has not received much attention and is still lacking in availability, such as Malay. The utilization of Malay language that is used by a vast community in the Southeast Asian territories had placed in priority the importance of the language to be properly identified in the usage of Information Retrieval (IR) and Information Extraction (IE) tasks. The lack of reference corpus specific in Malay had addressed studies to be done so as to resolve ambiguities of un-annotated Malay text. In the NER approach targeted for linguistic categories such Malay that is still least present in the mainstream, the application of appropriate NER algorithm in terms of precision and recall rate for entity categorization is important so as to assist the proper utilization of the word structure in near future. For future works, appropriate algorithms encasing other linguistic NER systems would be identified in order to be incorporated into the creation of a better Malay noun NER system.

ACKNOWLEDGEMENTS

The authors would like to thank Universiti Pendidikan Sultan Idris for the support of conducting the studies under the Research Grant 2014-0123-109-72, to the research group that contributed precious insights, and to all the lecturers that were involved, be it directly or otherwise.

REFERENCES

Abdul-Hamid, A. & Darwish, K. (2010). Simplified feature set for arabic named entity recognition. Published in *Proceedings of the 2010 Named Entities Workshop*, ACL 2010. Association for Computational Linguistics: Sweden, pp. 110-115.

Aboaoga, M. & Aziz, M.J. (2013). Arabic person names recognition by using a rule-based approach. Published in *Journal of Computer Science* 9 (7). Universiti Kebangsaan Malaysia: Bangi, pp. 922-927.

Abu Bakar, J., Omar, K., Nasrudin, M.F. & Murah, M.Z. (2013). Part-of-speech for old Malay manuscript corpus: A Review. Published in *Second International Multi-Conference on Artificial Intelligence Technology*, pp. 53-66.

Althobaiti, M., Kruschwitz, U., & Poesio, M. (2014). Automatic creation of Arabic named entity annotated corpus using Wikipedia. Published in *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Gothenburg: Sweden, pp. 106-115.

Ananiadou, S., Pyysalo, S., Tsuji, J., & Kell, D.B. (2010). Event extraction for systems biology by text mining the literature. Published in *Journal of Trends in Biotechnology*, vol 28, Issue 7. Elsevier: United Kingdom, pp. 381-390.

Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., & Etzioni, O. (2015). Open information extraction from the web. Published under the Grant of University of Washington. Seattle: United States of America.

Benajiba, Y., Rosso, P., & Ruiz, J.M. (2007). ANERsys: An Arabic named entity recognition system based on maximum entropy. Published in *Lecture Notes in*

Computer Science: *Computational Linguistics and Intelligent Text Processing*. Volume 4394, pp. 143-153.

Cao, T.H., Tang, T.M., & Chau, C.K. (2012). Text clustering with named entities: A model, experimentation and realization. Published in *Data Mining: Foundations and Intelligent Paradigms*, Vol. 23 of Intelligent Systems Reference Library. Springer-Verlag. Berlin, Hiedelberg.

Carvalho, J.P., Batista, F., & Coheur, L. (2012). A critical survey on the use of fuzzy sets in speech and natural language processing. Journal of *WCCI 2012 IEEE World Congress on Computational Intelligence*, Australia.

Derczynski, L., Maynard, D., Giuseppe Rizzo, Van Erp, M., Gorrell, G., Troncy, R., Petrak, J., & Bontcheva, K. (2014). Analysis of named entity recognition and linking for tweets. Published in *Journal of Information Processing & Management*, Volume 51, Issue 2. University of Sheffield: United Kingdom, pp. 32-49.

Don, Z.M. (2010). Processing natural Malay texts: A data-driven approach. Trames. Published under *Journal of the Humanities and Social Sciences* 14(1), pp. 90-103.

Elsayed, H. & Elghazaly, T. (2015). A rule-based entities recognition system for modern standard Arabic. Published in *IJCSI International Journal of Computer Science Issues*, Volume 12, Issue 1, No. 2. Cairo University: Egypt.

Elsebai, A. (2009). A rule-based system for named entity recognition in modern standard Arabic. Submitted as PhD Thesis, University of Sanford: United Kingdom.

Gifu, D. & Vasilache, G. (2014). A language independent named entity recognition system. Published in the *Proceedings of the 10th International Conference "Linguistic Resources and Tools for Processing the Romanian Language"*. University of Craiova: Rome.

Ismail, A. (2013). Minimally supervised techniques for bilingual lexicon extraction. Submitted as PhD Thesis, University of York: United Kingdom.

Kanagavalli, V.R. & Raja, K. (2010). Detecting and resolving spatial ambiguity in text using named entity extraction and self-learning fuzzy logic techniques. Published in *National Conference on Recent Trends in Data Mining and Distributed Systems*. Sathyabama University: Chennai.

Kral, P. (2014). Named entities as new features for Czech document classification. Published in *Journal of Computational Linguistics and Intelligent Text Processing*. University of West Bohemia: Czech Republic.

Liao, J.C. (2011). A method of combining ontology and closed frequent item sets for hierarchical document Clustering. Submitted as Master Thesis. National Taiwan University of Science & Technology: Taiwan.

Mohamed, H., Omar, N., & Aziz, M.J.A (2011). Statistical Malay part-of-speech (POS) Tagger using hidden Markov Approach. Published in *2011 International Conference on Semantic Technology and Information Retrieval*, IEEE.

Montalvo, S., Martinez, R., Casillas, A., & Fresno, V. (2007). Bilingual news clustering using named entities and fuzzy similarity. Published in *Proceedings of the 10th International Conference on Text, Speech and Dialogue*. Springer-Verlag. Berlin, Heidelberg.

Naji F. Mohammed & Omar, N. (2012). Arabic named entity recognition using artificial neural network. *Journal of Computer Science*. Vol. 8 (8), pp.1285-1293.

Nanda, M. (2014). The named entity recognizer framework. Published in *International Journal of Innovative Research in Advanced Engineering (IJIRAE)*. Madhav Institute of Technology & Science: India.

Nothman, J., Ringland, N., Radford, W., Murphy, T., & Curran, J.R. (2013). Learning multilingual named entity recognition from Wikipedia. Published in *Journal of Artificial Intelligence*. University of Sydney: Australia.

Oudah, M.M. & Shaalan, K. (2012). A pipeline Arabic named recognition using a hybrid approach. Published in *Proceedings of COLING 2012*: Technical Papers. British University: Mumbai, pp. 2159-2176.

Patrick, J. & Nguyen, D. (2011). Automated proof reading of clinical notes. Published in 25th Pacific Asia Conference on Language, Information and Computation. PACLIC: Australia, pp. 303-312.

Powers, D.M. (2011). Evaluation: From precision, recall and F-Measure to ROC, informedness, markedness & correlation. Published in *Journal of Machine Learning Technologies*, 2(1). Flinders University: Australia, pp. 37-63.

Rayner, A. Mujat, & J.H. Orbit. (2013). A rule-based part of speech (RPOS) tagger for Malay text articles. Proceedings from the *5th Asian Conference on Intelligent Information and Database System (ACIIDS)*, vol. 2, Springer-Verlag Berlin Heidelberg, pp. 50-59.

Rayner, A., Chin Leong, L., Kim On, C., & Anthony, P. (2014). Malay named entity recognition based on rule-based approach. *International Journal of Machine Learning and Computing*, Vol. 4, No. 3, June 2014.

Senthil, K., Thangmani, M., & Zubair, R. (2014). Bio-inspired fuzzy expert system for mining big data. Published in *Mathematical and Computational Methods in Science and Engineering.* Nehru Institute of Information Technology & Management: India.

Sinoara, R., Sundermann, C.V., Marcacini, R.M., Domingues, M.A., & Rezende, S.O. (2014). Named entities as privileged information for hierarchical text clustering. Published in *International Database Engineering & Applications Symposium*. Portugal.

Soo-Fong, Y., Ranaivo-Malacon, B. & Alvin Yeo Wee. (2011). The named entity recognition system for Iban language. *25th Pacific Asia Conference on Language, Information & Computation*. Published in *PACLIC*, pp. 549-558.

Turian, J., Ratinov, L., & Bengio, Y. (2010). Word representations: A simple and general method for semi-supervised learning. Published in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics: Sweden, pp. 384-394.

Wang C., Liu, Y., Sun, M. (2012). Minimum error rate training for bilingual news alignment. Published in *Lecture Notes in Computer Science, Chinese Lexical Semantics*. Tsinghua University: Beijing.

Yu, S., Eunji, Y., Eunju, K., & Gary, G.L. (2004). POSTBIOTM-NER: A machine learning approach for bio-named entity recognition. Published in *Proceedings of the EMBO Workshop on Critical Assessment of Text Mining Methods in Molecular Biology*.

Zamin, N., Oxley, A., Bakar Z.A., & Farhan, S.A. (2012). A statistical dictionary-based word alignment algorithm: An unsupervised approach. Published in *2012 International Conference on Computer & Information Science (ICCIS)*, Kuala Lumpur, pp. 396-402.

Zhan, Z. & Sun, L. (2011). Improving word sense induction by exploiting semantic relevance. Published in the *Proceedings of the 5th International Joint Conference on Natural Language Processing*. AFNLP: Thailand, pp. 1387-1391.