International Journal on Advanced Science Engineering Information Technology

Vol.6 (2016) No. 6
ISSN: 2088-5334

Biomedical Named Entity Recognition: A Review

Basel Alshaikhdeeb[#1], Kamsuriah Ahmad[#2]

[#]*Research Center for Software Technology and Management, Faculty of Information Science and Technology,*
Universiti Kebangsaan Malaysia, 43600, Bangi, Selangor, Malaysia
E-mail: [1]shaikhdeeb@gmail.com, [2]kamsuriah@ukm.edu.my

***Abstract*— Biomedical Named Entity Recognition (BNER) is the task of identifying biomedical instances such as chemical compounds, genes, proteins, viruses, disorders, DNAs and RNAs. The key challenge behind BNER lies on the methods that would be used for extracting such entities. Most of the methods used for BNER were relying on Supervised Machine Learning (SML) techniques. In SML techniques, the features play an essential role in terms of improving the effectiveness of the recognition process. Features can be identified as a set of discriminating and distinguishing characteristics that have the ability to indicate the occurrence of an entity. In this manner, the features should be able to generalize which means to discriminate the entities correctly even on new and unseen samples. Several studies have tackled the role of features in terms of identifying named entities. However, with the surge of biomedical researches, there is a vital demand to explore biomedical features. This paper aims to accommodate a review study on the features that could be used for BNER in which various types of features will be examined including morphological features, dictionary-based features, lexical features and distance-based features.**

***Keywords*— biomedical named entity recognition; feature extraction; supervised machine learning; dictionary-based features; morphological features; POS tagging**

I. Introduction

Recently, the last six decades have witnessed a dramatic growth of biomedical data in which 3000 new articles are being published daily in various journals [1]. For instance, MEDLINE which is a large-scale resource for medical articles contains around 20 million articles. This expansion of biomedical information requires large-scale management in order to provide the knowledge in a time manner. Such surge research has caught different attentions including computer scientists and biomedical experts.

As a response to such growth of data, several tasks have been posed in order to analyse these biomedical data. These tasks such as Biomedical Question Answering which aims to identify a precise answer for a biomedical question [2], and Indexing Biomedical Documents which aims to classify the biomedical documents into its category [3]. In order to perform these tasks, it is necessary to accommodate a prior process of recognition for the biomedical entities.

Biomedical Named Entity Recognition (BNER) is the task of identifying chemical compounds, genes, proteins, viruses, disorders, DNAs and RNAs [4]. With the tremendous expansion of biomedical data such as scientific papers, books and other publications produced annually, there is an arising demand for extracting the biomedical entities. MEDLINE is one of the large-scale resources for the biomedical domain in which millions of articles are being stored in a database [5]. The earliest research efforts that intended to extract biomedical entities were relying on handcrafted rule-based approaches [6]. However, the manual building of the rules seems to be time-consuming. The surge of using machine-learning techniques and text mining has offered a great opportunity for identifying biomedical entities automatically. For instance, the Supervised Machine Learning (SML) techniques, which depend on a predefined example set of data, have been examined by many researchers and shown promising performance [7]-[9]. Since several annotated and labeled corpus have been proposed such as SCAI [10] and GENIA [11], SML techniques have become more broadly used [12].

Nonetheless, biomedical entities tend to be more complicated compared to the traditional named entities extraction [13], [14]. Campos et al. [5] have listed multiple reasons behind the complexity of BNEs. First, the biomedical entities tend to be descriptive such as “normal thymic epithelial cells”. Second, the biomedical entities could be formed differently such as “N-acetylcysteine” which can be formed as “N-acetylcysteine” or “NAcetylCysteine”. Thirdly, the biomedical abbreviations may refer different entities such as the abbreviation of ‘TCF’ may refer “T-Cell Factor” or “Tissue Culture Fluid”. Fourth, biomedical entities contain complex morphology such as numbers and punctuations (e.g. 94-KDA).

In this manner, the task of selecting appropriate features for BNER is considered to be a challenging task in which the features should have the ability to generalize and discriminate the occurrence of biomedical entities. This because the feature has a significant impact on the effectiveness of the classification process in which, some features have low performance and others have good performance. Therefore, this paper aims to extensively survey the features that could be used for BNER.

The paper is organized as follows: Section II provides the main categories of the biomedical features. First category, which is Morphological features which contain Boolean, numeric and nominal features. The second category, which is Lexical features contains only nominal features. The third category, which is Dictionary-based contains only Boolean features. The fourth category, which is Distance-based contains only numeric features. Finally, Section III provides a discussion of the four main features by analysing each feature independently.

II. Materials and Methods

Supervised machine learning techniques aim at enabling computers to predict the state of a particular instance using a predefined set that contains examples [15]. The key characteristic behind these techniques lies in the appropriate representation of the instances in which the instances can be described as a vector space. This vector space composed of the features of the instances. In this vein, features can be defined as a set of distinguishing and discriminative characteristics that have the ability to describe the instances distinctly [16]. With the release of benchmark datasets for biomedical instances that contains predefined examples and annotated entities, the rely on the supervised machine learning techniques in terms of identifying biomedical named entities has remarkably increased. In this manner, it is necessary to concentrate on the features used for identifying biomedical entities in which the strength and weakness aspects of each feature can be tackled in details.

To the best of our knowledge, there is no available survey that discusses the features used for biomedical named entity recognition. However, Nadeau et al. [17] have provided an extensive survey for general named entity recognition. In their survey, a comprehensive discussion has been given to illustrate the features of the traditionally named entities such as person's name, organization's name, location's name and dates. Unlike the traditionally named entities, the biomedical named entities such as protein, gene, DNA, chemical compounds and others tend to be more complicated in terms of the morphology. The morphological complexity behind these entities lies on the unusual characters that could be used to describe these entities such as Greek letters, digits, and special characters. According to Hashim & Omar [18], some biomedical entities consists of multi-word that are separated via punctuation such as ‘Hydro-Oxide’. Apparently, this complexity would significantly hinder the process of detecting these entities. Hence, applying the features that have been used for the traditionally named entities to identifying biomedical instances would be insufficient. Therefore, there is an essential demand to investigate the features that have been used for the biomedical entities detection in order to determine the best feature representation.

There is a wide range of features that could be used for BNER. Nadeau et al. [17] have classified the features that could be used for traditionally named entities as Numeric, Boolean, and Nominal features. Numeric features are the features that can be represented numerically such as the word's length and the number of occurrences. Boolean features are the features that can be described binary using 1 or 0 in which 1 refers the presence of such feature and 0 refers to the absence of the feature. The popular example of such feature is the capitalization in which the word is being checked in terms of ‘Is-Capitalized’ condition. Finally, nominal features are the features that could be described using symbols. The famous example of this kind of features is the Part-Of-Speech (POS) tagging which aims to identify the syntactical tag of words such as verb, noun, and adjective. In this manner, the representation of this feature would be performed using symbols such as ‘VB' for the verb, or ‘NN' for the noun.

Based on the latter taxonomy, this paper categorizes the features that would be used for BNER as Morphological Features, Lexical Features, Dictionary-based features and Distance-based features as shown in Fig. 1. The proposed taxonomy of BNER features will be discussed extensively in the following sections.

A. Morphological Features

This kind of features aims to analyse the morphology of the word in which the analysis is being performed on the word-level. As shown in Fig. 1, the morphological features consist of numeric, Boolean, and nominal features. These categories are being illustrated in the following sub-sections.

1) Boolean Morphological Features

As mentioned earlier, this kind of feature aims to check the morphology of a word in accordance with a specific condition. The condition used is a feature that is included in the word. In this manner, there are three features or conditions could be used binary Is-Capitalized, Contains-Digits and Contains-Punctuations. *The is capitalized* feature aims to address the case of a given word such as upper-case or lower-case. Since biomedical entities are named entities thus, identifying the case of the word seems to be a good indicator. This is due to most of the names are capitalized.

Contains-Digits feature aims to examine the word in terms of digit inclusion. Many biomedical entities especially chemical compounds are being formed with digits such as ‘NH2’. Therefore, identifying whether the word contains the digit or not, could be a good indicator for BNER.

Contains-Punctuations feature aims to address the word in terms of punctuation inclusion (e.g. dash and underscore). Similar to the digits, many biomedical entities contain punctuation or special characters such as the protein ’94-KDA’.

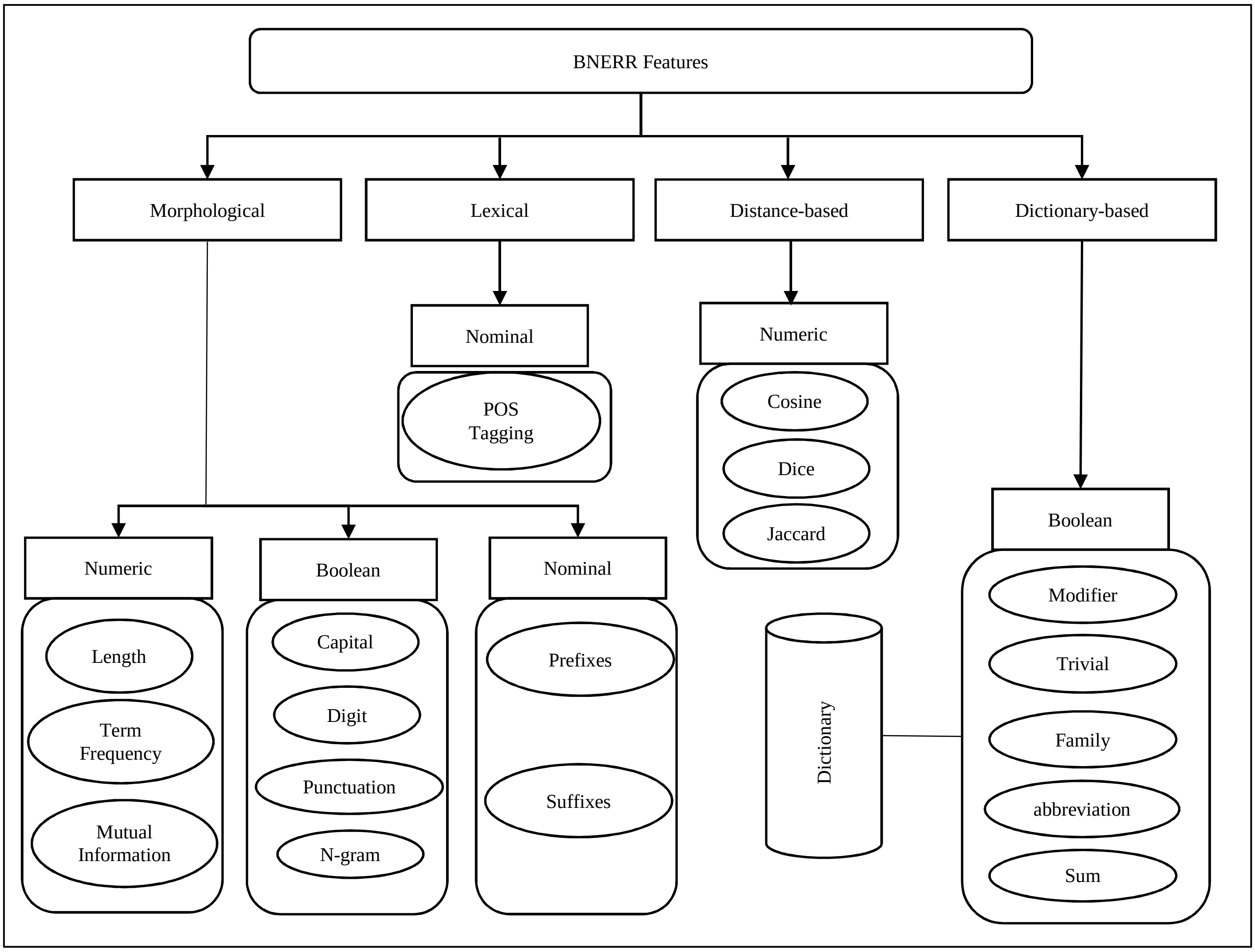

Fig. 1 Taxonomy of BNER features

Hence, examining the inclusion of special character for a given word would be a good indicator in accordance to be BNE. Table 1 shows an example of representing the Boolean features.

TABLE I
EXAMPLE OF BOOLEAN REPRESENTATION

Tokens	Is-Capital	Contain-digit	Contain-punctuation
Hydroxypyridinones	1	0	0
and	0	0	0
their	0	0	0
complexes	0	0	0
of	0	0	0
1-octanol	0	1	1

Friedrich et al. [19] have used the three mentioned features in order to identify chemical compounds. The authors have expanded these features to include more cases. For instance, the capitalization has been divided into three cases including Init-Capital which refer that the word is beginning with a capital letter (e.g. Nitrogen), All-Capital which refer that the word is uppercase (e.g. ACID), and Camel-Case which refer that the word contains multiple cases (e.g. HydroOxide). In the same manner, punctuation feature has been divided into two features Has-Dash (e.g. 'beta-cyclodextrin') and Has-Underscore (e.g. 'N_methyl'). Finally, the digit feature has been divided into three features Is-Greek (e.g. Gamma), Is-Roman (e.g. XII) and Contains-Digit (e.g. CD28).

Apart from these features, there is another Boolean feature, which called *n-gram* feature. This feature aims to use the unique terms in the corpus as features in which the representation would be as term occurrence. The presence of a term is represented as 1 and the absence is represented as 0. In addition, Batista-Navarro et al. [20] have examined the morphological Boolean features including capitalization, containing words special characters, and n-gram. The authors have used a sentence splitting approach in order to turn the text into multiple sentences. Consequentially, a tokenization task has been performed in order to turn the sentences into series of tokens (i.e. words). Hence, every token has been analyzed in terms of the mentioned features in a binary manner. Similarly, Klinger et al. [21] have also examined the n-gram feature or so-called bag-of-words in which every unique term is represented as a feature. The authors additionally used the punctuation features such as slash (e.g. / \), dash (e.g. -) and quote (e.g. ‘’ “”). Finally, Bhasuran et al. [22] have addressed the Boolean morphological features with ensemble classifier in order to recognize diseases.

2) *Nominal Morphological Features*

This kind of features aims to address the specific morphology aspect in the word nominally. The famous types of such feature are the prefixes and the suffixes. Prefixes are the initial letters that commonly occur in multiple words such as 'acylglycerol' and 'acyclonucleosides' where the initial letters of 'acy' have been commonly occurred in both entities. Suffixes are the ending letters that commonly occur in multiple words such as 'trifluoromethyl' and 'chloroethyl' where the letters 'thyl' have been commonly occurred in both entities. Alharbi & Tiun [23] have examined the role of prefixes and suffixes in terms of identifying chemical compounds, Table 2 shows some examples of utilized prefixes and suffixes by the authors.

TABLE II
SAMPLE OF PREFIXES AND SUFFIXES

Tokens	Prefix	Suffix
Acetazolamide	Ac	-
Acetaminophen	Ac	-
Aceon	Ac	-
Acetadote	Ac	-
Acebutolol	Ac	-
Acarbose	Ac	-
Aminoeth**yl**	-	yl
Aminometh**yl**	-	yl
Azidovin**yl**	-	yl
Avandar**yl**	-	yl
Azirin**yl**	-	yl
Benzothiazol**yl**	-	yl

As shown in Table 2, the representation of prefixes and suffixes features have been performed nominally using symbols. These symbols refer to the prefix or suffix that could be contained in the token.

Similarly, De Matos et al. [24] have used the suffixes and prefixes in order to identify chemical named entities.

Similarly, Hashim & Omar [18] have used the prefixes and suffixes as encoded features fed to a back-propagation neural network that adopted to classify biomedical named entities using two benchmark datasets including SCAI and GENIA.

3) *Numeric Morphological Features*

This kind of features aims to represent a specific feature numerically. Multiple features could be represented using integers or real values. The first feature is the *word's length* in which the word is being analysed in terms of its length. This feature has been inspired by the fact that most BNEs tend to be long words [25]. In this manner, the number of characters contained in each token will be analysed and represented numerically.

The second feature is the *term frequency* in which the tokens are being analysed in terms of the number of occurrences. In fact, the frequent occurrence of a term is usually indicating valuable patterns such as the occurrence of the term 'anti', which located in many drugs' names such as 'anti-tumor' and 'anti-bacterial'.

Another morphological numeric feature is the *mutual information.* This feature aims to identify the co-occurrence among two tokens. Usually, such feature is being used to extract multi-word compounds [26]. In biomedical, there are plenty of terms that occurred frequently with each other, such as 'Hydro' and 'Oxide'. To identify the co-occurrence among the terms, Mutual Information could be the appropriate method. Mutual Information aims to examine the strengthen between two words by addressing the co-occurrence among the two words (i.e. number of times the words occurred together), and the independence occurrence of the two words (i.e. the occurrence of each word separately). Let a and b are two terms, in order to compute the mutual information between these terms, the following equation is being used to [27]:

$$Mutual\ information\ (a\ ,b) = \log\frac{P(a\ ,b)}{P(a)*P(b)} \qquad (1)$$

, where *P (a)* is the occurrence number of the term *a*, *P (b)* the occurrence number of the term *b* and *P (a , b)* is the occurrence number of both terms.

Usié et al. [25] have examined the role of word's length feature in terms of identifying chemical compounds. On the other hand, Rocktäschel et al. [28] have used the term frequency feature in order to address the occurrence number of specific affixes.

B. Lexical Features

This kind of feature aims to identify the syntactic of the words in which the grammatical aspect is taking into the consideration. The popular example of this kind is the POS tagging which aims to provide the syntactical tag for each word such as verb, noun, adjective or adverb [29]. POS tagging is a word sense disambiguation method that aims to eliminate the confusion resulted from multiple meanings with a single term. For example, the word 'treat' has two meanings, which are 'pleasure' and 'cure'. The key distinguishing behind the two mentioned terms lies on their syntactical tags in which the verb of *treat* refers to *cure*, and the noun of *treat* refers to *pleasure*.

Therefore, POS tagging has been examined in many text-mining applications. However, in terms of BNER, POS tagging provides a clue for the biomedical entities. For example, most of the chemical compounds with an 'ic' suffixes such as 'Anthelmintic' could be tagged as adjectives. In this manner, knowing that a given word has an adjective syntactic class would increase the probability of being chemical compounds. POS tagging feature is being represented as nominal using specific symbols that indicate the syntactic class. Table 3 shows a sample of these symbols.

TABLE III
EXAMPLE OF POS TAGGING REPRESENTATION

Tokens	POS tag	Description
The	DT	Determiner
Advantage	NN	Noun
for	DT	Determiner
this	DT	Determiner
type	NN	Noun
of	DT	Determiner
hydroxypyridinone	JJ	Adjective
lies	VB	Verb
on	DT	Determiner
the	DT	Determiner
distribution	NN	Noun

As shown in Table 3, POS tagging feature has been represented nominally. This has been performed using specific symbols such as 'NN' which indicates noun or 'DT' which indicates determiner.

Several researchers have used this feature such as Rocktäschel et al. [28], Friedrich et al. [19], Corbett & Copestake [30] and Alharbi & Tiun [23]. However, besides using POS tagging, Batista-Navarro et al. [20] have used additional syntactic tool which is called chunk which aims to identify the noun phrases and verb phrases. Generally, most of the approaches for identifying biomedical named entities that used the POS tagging feature were not being mainly depending on such features, instead, it was used increase the probability of instances in terms of being biomedical or not.

C. Dictionary-Based Features

These features are mainly depending on a predefined list or dictionary that contains a large number of specific instances. It aims at assigning biomedical instances into their corresponding category relying on inclusion in particular list or dictionary. For example, the biomedical entity of 'Lysine' could be stored in an Amino Acids list thus, once this entity is being countered in the dataset, it could be identified using the Amino Acids list. There is four dictionary-based feature that has been widely used in the literature, such features are illustrated as follows [31]:

- *Modifier*: is a set of words that commonly followed by biomedical entities. In other meaning, this feature contains the keywords that probably occurred before the biomedical entities.

- *Trivial (company-code)*: is the trivial name of chemical compounds such as "ethyl" instead of "ethanol". This feature is associated with the alternatives names or synonyms that could be used for the biomedical entity.

- *Family*: is the family of a biomedical entity such as "Alcohol" is the family of "Ethanol". This feature indicates the 'part-of' relations among the biomedical entities.

- *Abbreviation*: is the abbreviation of chemical entities such as "cl" for Calcium. This feature tends to be hyponyms of the biomedical entities.

- *Sum (molecular formula)*: is the compound of multiple chemical entities such as "Carbohydrate" which consists of "Carbon" and "Hydrogen".

Table 4 depicts the representation of dictionary-based features.

TABLE IV
REPRESENTATION OF DICTIONARY-BASED FEATURES

Tokens	Modifier	Trivial	Family	Abb	Sum
Iron (iii)	0	1	0	0	0
Complexes	1	0	0	0	0
Acids	0	0	1	0	0
NPS	0	0	0	1	0
CH3CN	0	0	0	0	1

as shown in Table 4, the first instance which is '*Iron (iii)*' is considered to be a trivial name or a code company for the oxide, therefore, it was assigned as '1' to indicates its existence in the trivial dictionary. One the other hand, the second instance which is '*complexes*' has been assigned as a modifier where it usually occurred after biomedical entities as '*Iron (iii) complexes*', in this vein, it is considered to be a keyword therefore, it has been assigned as '1' to indicates its existence in the modifier dictionary. In addition, the third instance which is 'Acids' has been assigned to a family in which the Acids family contains sub-chemical instances thus, it has been assigned as '1' to indicates its existence in the family dictionary. Furthermore, the fourth instance which is 'NPS' has been assigned as abbreviation where it indicates the drug of 'New Psychoactive Substances' therefore, it has been assigned as '1' to indicates its existence in the abbreviation dictionary. Finally, the fifth instance which is '*CH3CN*', this instance is indicating the chemical compound of *Acetonitrile*. This chemical compound is a combination of Methene (CH3) and Cyanide (CN) therefore, it has been assigned as '1' to indicates its existence in the sum or molecular formula dictionary.

Friedrich et al. [19] have performed a study emphasizing the advantage of using dictionary-based features. In their study, the authors have demonstrated the improvement resulted from the use of dictionary-based feature in terms of identifying chemical compounds. However, the authors have recommended that these features should be used as supplementary features. In another word, these features should be used with other kinds of features in which it can facilitate the process of recognition. Corbett & Copestake [30] and Degtyarenko et al. [32] have used synonyms approach using dictionaries in order to identify matches of entities. De Matos et al. [24] have used an extension of dictionary features where more entities have been identified from lists.

Rocktäschel et al. [28] have used the dictionary-based features including family, company code and molecular formula for extracting chemical compounds. Similarly, Usie et al. [25] have used the dictionary-based features including family, molecular formula and company code with a conditional random field (CRF) classifier in order to recognize chemical compounds. Lamurias et al. [33] have used ontology-based features in order to retrieve semantic correspondences for the biomedical entities. Finally, Zhang et al. [34] have addressed the dictionary features using an unsupervised approach in order to identify semantic correspondences for the chemical entities.

D. Similarity or Distance-Based Features

These kinds of features rely on similarity or distance measures such as Cosine, Dice, and Jaccard. These measure aims to identify the similarity between two words numerically. Let w_1 and w_2 are two words, if we want to address the similarity between these two words, it should be converted into vectors as $\overrightarrow{w_1}$ and $\overrightarrow{w_2}$. In this manner, the Cosine similarity can be applied as follows [35]:

$$Cosine\ (\overrightarrow{w_1}, \overrightarrow{w_2}) = \frac{\overrightarrow{w_1}.\overrightarrow{w_2}}{|\overrightarrow{w_1}|.|\overrightarrow{w_2}|} \qquad (2)$$

Whereas Dice can be applied as:

$$Dice\ (\overrightarrow{w_1}, \overrightarrow{w_2}) = \frac{2\times|\overrightarrow{w_1} \cap \overrightarrow{w_2}|}{|\overrightarrow{w_1}|+|\overrightarrow{w_2}|} \quad (3)$$

Eventually, Jaccard can be applied as:

$$Jaccard\ (\overrightarrow{w_1}, \overrightarrow{w_2}) = \frac{|\overrightarrow{w_1} \cap \overrightarrow{w_2}|}{|\overrightarrow{w_1} \cup \overrightarrow{w_2}|} \quad (4)$$

Alharbi & Tiun [23] have used three similarity or distance-based features including Cosine, Dice, and Jaccard to identify the chemical compounds. The authors have intended to measure the similarity between the chemical compounds based on the morphological similarity using prefixes and suffixes.

III. Results and Discussion

This study attempts to provide an analysis for each feature. First, the Boolean morphological features have demonstrated fair performance in terms of recognizing biomedical entities [19]-[21]. This can be represented by the indication that would be provided by these features which facilitate training the classifier by giving example cases for instances that contain whether punctuation, digits or capital letters. However, the main drawback of these features lies on the normalization tasks that would be performed as preprocessing. Such normalization tasks aim at tokenizing the terms such as lowering the letters' cases or removing numbers, which apparently makes the use of these features non-sense.

Second, nominal morphological features have demonstrated superior performance [18], [24]. This is because most of the biomedical entities contain whether prefixes or suffixes. In the same manner, lexical feature, which is also nominal, shows low performance [19], [25], [28]. This is due to the POS tagging is giving different tags for the biomedical entities. Therefore, it can be used as an indicator rather than an independent feature.

Third, the numeric features either morphological or distance-based are considered insufficient for the supervised machine learning. It rather considered being suitable for unsupervised learning techniques regarding the real values produced by such features [23].

Fourth, dictionary-based features have shown a good performance in terms of identifying biomedical entities [19], [24], [25], [28], [30], [32], [33]. However, it has the main drawback which lies on the exact match between the instances from the corpus and the instances from the dictionary, which means that even a slight change (e.g. lowercase and uppercase) would lead to mismatch the instances [19]. Note that, it is difficult to contain all the biomedical entities in lists due to the continuous inventions of drugs which lead to new and unseen instances. Moreover, using the dictionary could counter some obstacles such as the ambiguity that lie on some abbreviations, for instance, the abbreviation *TCF* could be used to refer “*T cell factor*” or “*Tissue Culture Fluid*” [5]. Therefore, dictionary-based features can be used as supplementary features in which the detection of biomedical entities would not mainly rely on such features, but rather it could use them as indicators.

Table 5 shows the state of art classified based on the taxonomy of this paper.

Table V
Taxonomy of the Features Used by the State of the Art

Author	Morphological			Lexical	Distance-based	Dictionary-based
	Boolean	Nominal	Numeric			
Friedrich et al. [19]	√			√		√
Degtyarenko et al. [32]						√
Corbett & Copestake [30]				√		√
Klinger et al. [21]	√					
De Matos et al. [24]		√				√
Rocktaschel et al. [28]			√	√		√
Lamurias et al. [33]						√
Alharbi & Tiun [23]		√		√	√	
Batista-Navarro et al. [20]	√			√		
Usié et al. [25]			√			√
Zhang et al. [34]						√
Hashim & Omar [18]		√				
Bhasuran et al. [22]	√					

IV. Conclusions

This paper has provided an extensive review of the features of BNER in which a taxonomy has been identified based on the representation (i.e. nominal, numeric and Boolean). In addition, a discussion has been performed in order to provide a critical analysis for each feature. Morphological Boolean features shown superiority compared to other features. Establishing a comparative study using these features would be a great opportunity for future researches in terms of identifying their performances for extracting biomedical entities.

Acknowledgement

The authors would like to thank Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia by giving the authors an opportunity to conduct this research. This research is funded by Universiti Kebangsaan Malaysia under Fundamental Research Grant Scheme FRGS/1/2014/ICT07/UKM/02/3 and DPP-2015-019.

REFERENCES

[1] G. Tsatsaronis, M. Schroeder, G. Paliouras, Y. Almirantis, I. Androutsopoulos, E. Gaussier, P. Gallinari, T. Artieres, M. R. Alvers, and M. Zschunke, "BioASQ: A Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering," in *AAAI fall symposium: Information retrieval and knowledge discovery in biomedical text*, 2012.

[2] G. Balikas, A. Krithara, I. Partalas, and G. Paliouras, "BioASQ: a challenge on large-scale biomedical semantic indexing and question answering," in *Multimodal Retrieval in the Medical Domain*, ed: Springer, 2015, pp. 26-39.

[3] M. Sarrouti and S. O. El Alaoui, "A Generic Document Retrieval Framework Based on UMLS Similarity for Biomedical Question Answering System," in *Intelligent Decision Technologies 2016: Proceedings of the 8th KES International Conference on Intelligent Decision Technologies (KES-IDT 2016) – Part II*, I. Czarnowski, M. A. Caballero, J. R. Howlett, and C. L. Jain, Eds., ed Cham: Springer International Publishing, 2016, pp. 207-216.

[4] S. Saha, A. Ekbal, and U. K. Sikdar, "Named entity recognition and classification in biomedical text using classifier ensemble," *International journal of data mining and bioinformatics,* vol. 11, pp. 365-391, 2015.

[5] D. Campos, S. Matos, and J. L. Oliveira, "Biomedical named entity recognition: a survey of machine-learning tools," *Theory and Applications for Advanced Text Mining. InTech,* pp. 175-195, 2012.

[6] R. Leaman and G. Gonzalez, "BANNER: an executable survey of advances in biomedical named entity recognition," in *Pacific symposium on biocomputing,* 2008, pp. 652-663.

[7] J. i. Kazama, T. Makino, Y. Ohta, and J. i. Tsujii, "Tuning support vector machines for biomedical named entity recognition," in *Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain-Volume 3*, 2002, pp. 1-8.

[8] B. Settles, "Biomedical named entity recognition using conditional random fields and rich feature sets," in *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications,* 2004, pp. 104-107.

[9] R. T.-H. Tsai, C.-L. Sung, H.-J. Dai, H.-C. Hung, T.-Y. Sung, and W.-L. Hsu, "NERBio: using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition," *BMC bioinformatics,* vol. 7, p. 1, 2006.

[10] Fraunhofer. (2014). *SCAI Dataset: Scientific Computing Institute for Algorithms*. [online] Available: http://www.scai.fraunhofer.de/en/business-research-areas/bioinformatics/research-development/information-extraction-semantic-text-analysis/named-entity-recognition/chem-corpora.html

[11] J.-D. Kim, T. Ohta, Y. Tateisi, and J. i. Tsujii, "GENIA corpus—a semantically annotated corpus for bio-textmining," *Bioinformatics,* vol. 19, pp. i180-i182, 2003.

[12] G. Zhou, J. Zhang, J. Su, D. Shen, and C. Tan, "Recognizing names in biomedical texts: a machine learning approach," *Bioinformatics,* vol. 20, pp. 1178-1190, 2004.

[13] H. A. Shabat and N. Omar, "Named Entity Recognition in Crime News Documents Using Classifiers Combination," *Middle-East Journal of Scientific Research,* vol. 23, pp. 1215-1221, 2015.

[14] K. R. Rahem and N. Omar, "Rule-Based Named Entity Recognition For Drug-Related Crime News Documents," *Journal of Theoretical & Applied Information Technology,* vol. 77, 2015.

[15] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques," ed, 2007.

[16] J. Turian, L. Ratinov, and Y. Bengio, "Word representations: a simple and general method for semi-supervised learning," in *Proceedings of the 48th annual meeting of the association for computational linguistics,* 2010, pp. 384-394.

[17] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes,* vol. 30, pp. 3-26, 2007.

[18] B. Hashim and N. Omar, "A Back Propagation Neural Network for Identifying Multi-Word Biomedical Named Entities," *2016,* vol. 11, 2016.

[19] C. M. Friedrich, T. Revillion, M. Hofmann, and J. Fluck, "Biomedical and chemical named entity recognition with conditional random fields: The advantage of dictionary features," in *Proceedings of the Second International Symposium on Semantic Mining in Biomedicine (SMBM 2006),* 2006, pp. 85-89.

[20] R. Batista-Navarro, R. Rak, and S. Ananiadou, "Optimising chemical named entity recognition with pre-processing analytics, knowledge-rich features and heuristics," *J Chem Inf,* vol. 7, p. S6, 2015.

[21] R. Klinger, C. Kolářik, J. Fluck, M. Hofmann-Apitius, and C. M. Friedrich, "Detection of IUPAC and IUPAC-like chemical names," *Bioinformatics,* vol. 24, pp. i268-i276, 2008.

[22] B. Bhasuran, G. Murugesan, S. Abdulkadhar, and J. Natarajan, "Stacked Ensemble Combined with Fuzzy Matching for Biomedical Named Entity Recognition of Diseases," *Journal of Biomedical Informatics,* 2016.

[23] E. Alharbi and S. Tiun, "A Hybrid Method of Linguistic Features and Clustering Approach for Identifying Biomedical Named Entities," *Asian Journal of Applied Sciences,* vol. 8, pp. 210-216, 2015.

[24] P. De Matos, R. Alcántara, A. Dekker, M. Ennis, J. Hastings, K. Haug, I. Spiteri, S. Turner, and C. Steinbeck, "Chemical entities of biological interest: an update," *Nucleic acids research,* p. gkp886, 2009.

[25] A. Usié, J. Cruz, J. Comas, F. Solson, and R. Alves, "CheNER: a tool for the identification of chemical entities and their classes in biomedical literature," *J Cheminform,* vol. 7, p. S15, 2015.

[26] S. Ab Rahman, N. Omar, and M. J. Ab Aziz, "Extraction of Compound Nouns in Malay Noun Phrases Using a Noun Phrase Frame Structure," *Asia-Pacific Journal of Information Technology and Multimedia,* vol. 3, 2013.

[27] W. Zhang, T. Yoshida, X. Tang, and T.-B. Ho, "Improving effectiveness of mutual information for substantival multiword expression extraction," *Expert Syst. Appl.,* vol. 36, pp. 10919-10930, 2009.

[28] T. Rocktäschel, M. Weidlich, and U. Leser, "ChemSpot: a hybrid system for chemical named entity recognition," *Bioinformatics,* vol. 28, pp. 1633-1640, 2012.

[29] R. Navigli, "Word sense disambiguation: A survey," *ACM Comput. Surv.,* vol. 41, pp. 1-69, 2009.

[30] P. Corbett and A. Copestake, "Cascaded classifiers for confidence-based chemical named entity recognition," *BMC bioinformatics,* vol. 9, p. S4, 2008.

[31] C. Kolárik, R. Klinger, C. M. Friedrich, M. Hofmann-Apitius, and J. Fluck, "Chemical names: terminological resources and corpora annotation," in *Workshop on Building and evaluating resources for biomedical text mining (6th edition of the Language Resources and Evaluation Conference),* 2008.

[32] K. Degtyarenko, P. De Matos, M. Ennis, J. Hastings, M. Zbinden, A. Mcnaught, R. Alcántara, M. Darsow, M. Guedj, and M. Ashburner, "ChEBI: a database and ontology for chemical entities of biological interest," *Nucleic acids research,* vol. 36, pp. D344-D350, 2008.

[33] A. Lamurias, T. Grego, and F. M. Couto, "Chemical compound and drug name recognition using CRFs and semantic similarity based on ChEBI," in *BioCreative Challenge Evaluation Workshop*, 2013, p. 75.

[34] Y. Zhang, J. Xu, H. Chen, J. Wang, Y. Wu, M. Prakasam, and H. Xu, "Chemical named entity recognition in patents by domain knowledge and unsupervised feature learning," *Database,* vol. 2016, p. baw049, 2016.

[35] V. U. Thompson, C. Panchev, and M. Oakes, "Performance evaluation of similarity measures on similar and dissimilar text retrieval," in *Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), 2015 7th International Joint Conference on,* 2015, pp. 577-584.