A Biomedical Named Entity Recognition Using Machine Learning Classifiers and Rich Feature Set

Ahmed Sultan Al-Hegami
University of Sana'a ,Yemen

Ameen Mohammed Farea Othman
The Arab Academy for Banking and Financial Sciences

Fuad Tarbosh Bagash
European Academy for higher studies ,Yemen

Summary
As the wealth of biomedical knowledge in the form of literature increases, there is a rising need for effective natural language processing tools to assist in organizing, curating, and retrieving this information. The task of named entity recognition becomes more difficult from specific domain since entities are more exact to that particular domain. To that end, named entity recognition (the task of identifying words and phrases in free text that belong to certain classes of interest) is an important first step for many of these larger information management goals. In recent years, much attention has been focused on the problem of recognizing gene and protein and other biomedical entities mentions in biomedical abstracts. Thus, this study aims to design and develop a biomedical named entity recognition model. A machine learning classification framework is proposed based on Naïve Bayes, K-Nearest Neighbour and decision tree classifiers. we have performed several experiments to empirically compare different subsets of features and three classification approach Naïve Bayes, K-Nearest Neighbour and decision tree for biomedical named entity recognition. The aim is to efficiently integrate different feature sets and classification algorithms to synthesize a more accurate classification procedure. Results prove that the K-Nearest Neighbour trained with suitable features is more suitable to recognize named entities of biomedical texts than other models.
Key words:
Named entity recognition (NER), learning, classification, framework, decision tree, recognizing gene, Naïve Bayes, K-Nearest Neighbour.

1. Introduction

Named entity recognition (NER) is one of the important tasks in information extraction, which involves the identification and classification of words or sequences of words denoting a concept or entity. Examples of named entities in general text are names of persons, locations, or organizations. Domain-specific named entities are those terms or phrases that denote concepts relevant to one particular domain. For example, protein and gene names are named entities which are of interest to the domain of molecular biology and medicine. The massive growth of textual information available in the literature and on the Web necessitates the automation of identification and management of named entities in text [1]. Named entity recognition is a crucial component of biomedical natural language processing, enabling information extraction and ultimately reasoning over and knowledge discovery from text. Much progress has been made in the design of rule-based and supervised tools, but they are often genre and task dependent. As such, adapting them to different genres of text or identifying new types of entities requires major effort in re-annotation or rule development [2]. The core techniques and approaches to NER may be classified into three classes, which are rule-based approach, machine learning approach and hybrid based approach. Rule-based approaches mainly aim to extract names with the use of a set of human made rules. In general, these models include of a number of different patterns that use grammar based (such as part of speech (POS)), syntactic based (such as word precedence) and orthographic based features (such as capitalization) with the use of dictionaries. One the other hand, the rule-based models so not have the capability of being portable, dynamic and robust, and also the large costs of maintaining the rules rises when the data is changed a small amount.
Many researchers are currently making use of the available machine learning techniques and approaches for biomedical NER, because they are easy to train, and they are cheaper to maintain. The machine learning approaches and techniques may be classified into the following classes: unsupervised techniques, semi-supervised techniques and supervised techniques. Several of the supervised based machine learning techniques that are used in NER are Support Vector Machines (SVM)and naïve Bayes.
Other than the previously mentioned studies, there are a great deal of related studies as well. Most of the domains included are social media, news, and medical domains. On the other hand, the studies associated with biomedical NER remain at an early stage. The biomedical domain is chosen for the initial experiments due to its importance and inherent challenges.

2. Motivation

In view of weakness inherent in manual searching of text, it has become imperative to seek other efficient ways to carry out text mining. The massive volume of bio-medical information stored in soft documents copies form, which obviously could be due to a substantial increase in scientific research over the years has necessitated the use of text mining technology. Searching and processing information from documented data is time-consuming in many areas for example bio-medical literature and is becoming not practical and easy to achieve without computer support. Thus, today the need for intelligent text handle applications that can replace or support human information exploration in bio-medical text documents is strong.

It has become extremely difficult for biologists to keep up with the relevant journals in their own discipline, let alone publications in other, related disciplines [3]. Bio-medical literature considered as a source of authentic medical knowledge which is critical for e-health applications. These kinds of e-health applications have a huge commercial prospect. According to the US National Center for Health Statistics, 51% of USA adults people had used the surfing of internet for health information in 2009 [3]. This potential commercial prospect has led to the launch of freely provided sources and others that require fees for access [3]. Many software hosted on the internet has provided incredible assistance to patients to identify symptoms of diseases and even adverse drugs reactions early enough to take first aid before experts are consulted. Biological researchers are very considerable on the reality of use the knowledge that is founded inside bio-medical literature. For instance, there are above twenty-two million abstracts the domains of medicine, bio-medical sciences, laboratory sciences, etc. in Medline alone.

The field of Natural Language Processing is an emerging field in Text Mining, which aims to automate the process of locating and classifying important information from large unstructured text base. This gives the data some form of shape and structure for ease human use. The task obviously requires at least a limited considerate of the text itself and the introduction of new compound patterns that simulate human information search, which makes text-mining tasks more complex and challenging than traditional keyword-lookup based information retrieval tasks.

3. Related Work

Most of the work on named entity recognition has initially focused on news domain. However, the features, pre-processing and post-processing used in these work are not equally effective on biomedical text, unless domain specific knowledge and techniques are incorporated. Biomedical texts are substantially different from other genres of text (such as newspaper articles). Ranging from the terminology and sentence construction to the valence and semantics of names are created continuously. Besides, authors of biomedical texts often do not follow proposed standardized names or formats and prefer to use abbreviations or other forms depending on personal inclination [4] [5]. Because of their limited length, such abbreviations/acronyms are sometimes identical to other words or symbols which increases the ambiguity. For instance, it was reported that 80% of the abbreviations listed in the machine learning have ambiguous representation in MEDLINE [6]. Sometimes the same name is shared by different types of bio-entity types. For example, “C1R” is a cell line, but there exists a gene (SwissProt P00736) that has the same name. Usage of digits and other non-alphabetic characters inside bio-entity names is also common. Compound names further complicate the situation. Locating the beginning and ending of such names within a sentence is not so straightforward since verbs and adjectives are often embedded in such names. Due to these complexities, named entity recognition attracted a huge amount of research interests. A number of shared tasks/challenges such as BioNLP/NLPBA 2004, BioCreative, CALBC, etc. provided benchmarks to compare and showcase the advancement in this field.

[7] Pose the classifier ensemble problem under single and multi-objective optimization frameworks, and evaluate it for Named Entity Recognition (named entity recognition), an important step in almost all Natural Language Processing (NLP) application areas. We propose the solutions to two different versions of the ensemble problem for each of the optimization frameworks. [7] Hypothesize that the reliability of predictions of each classifier differs among the various output classes. Thus, in an ensemble system it is necessary to find out either the eligible classes for which a classifier is most suitable to vote (i.e., binary vote based ensemble) or to quantify the amount of voting for each class in a particular classifier (i.e., real vote based ensemble). They use seven diverse classifiers, namely Naive Bayes, Decision Tree (DT), Memory Based Learner (MBL), Hidden Markov Model (HMM), Maximum Entropy (ME), Conditional Random Field (CRF) and Support Vector Machine (SVM) to build a number of models depending upon the various representations of the available features that are identified and selected mostly without using any domain knowledge and/or language specific resources. Results for all the languages show that the proposed classifier combination with real voting attains the performance level which is superior to all the individual classifiers, three baseline ensembles and the corresponding single objective based ensemble.

[8] Propose a single objective optimization based classifier ensemble technique using the search capability of genetic algorithm GA for named entity recognition C in biomedical texts. Here, GA is used to quantify the amount of voting for each class in each classifier. They use diverse

classification methods like Conditional Random Field and Support Vector Machine to build a number of models depending upon the various representations of the set of features and/or feature templates.

[9] Present a semi-supervised learning method that efficiently exploits unlabeled data in order to incorporate domain knowledge into a named entity recognition model and to leverage system performance. The proposed method includes Natural Language Processing (NLP) tasks for text pre-processing, learning word representation features from a large amount of text data for feature extraction, and conditional random fields for token classification. Other than the free text in the domain, the proposed method does not rely on any lexicon nor any dictionary in order to keep the system applicable to other named entity recognition tasks in bio-text data. Results: We extended named entity recognition, a biomedical named entity recognition system, with the proposed method. This yields an integrated system that can be applied to chemical and drug named entity recognition or biomedical named entity recognition.

[10] Present ChemSpot, a named entity recognition (named entity recognition) tool for identifying mentions of chemicals in natural language texts, including trivial names, drugs, abbreviations, molecular formulas and International Union of Pure and Applied Chemistry entities. Since the different classes of relevant entities have rather different naming characteristics, ChemSpot uses a hybrid approach combining a Conditional Random Field with a dictionary. It achieves an F1 measure of 68.1% on the SCAI corpus, outperforming the only other freely available chemical named entity recognition tool, OSCAR4, by 10.8 percentage points.

[11] Present classifiers ensemble approaches for biomedical named entity recognition. Generalized Winnow, Conditional Random Fields, Support Vector Machine, and Maximum Entropy are combined through three different strategies. We demonstrate the effectiveness of classifiers ensemble strategies and compare its performances with standalone classifier systems. In the experiments on the JNLPBA 2004 evaluation data, our best system achieves an F-score of 77.57%, which is better than most state of the art systems. The experiment show that our proposed classifiers ensemble method especially the stacking method can lead to significant improvement in performances of biomedical named entity recognition.

State-of-the-art named entity recognition approaches use various machine learning algorithms. These include hidden Markov model (HMM), support vector machine (SVM), maximum entropy Markov model, conditional random fields (CRFs), Among these algorithms, CRFs appear to be the most popular choice.

One common characteristic in many of these systems is the combination of results from multiple classifiers (e.g. see [12]). Apart from that, there is a substantial agreement among the feature sets used by these systems, most of which are actually various orthographic features.

Most of the work to date on named entity recognition is focused on genes/proteins. The state-of-the-art gene/protein mention recognition systems achieve F-scores around 88%, which is quite high. These systems often use either gene/protein specific features (e.g. Greek alphabet matching) or post-processing rules (e.g. extension of the identified mention boundaries to the left when a single letter with a hyphen precedes them [12] which might not be equally effective for other bio-entity type identification.

More efforts should be devoted to take advantage of contextual clues and features. In the last few years, some disease annotated corpora have been released. However, they have been annotated primarily to serve the purpose of relation extraction and, for different reasons, most of them are not suitable for the development of machine learning based disease mention recognition systems [13]. For example, the BioText [14] corpus has no specific annotation guideline and contains several inconsistencies, while the PennBioIE [15] is very specific to a particular sub-domain of diseases. Among other disease annotated corpora, the EBI disease corpus [16] is not annotated with disease mention boundaries which makes it unsuitable for named entity recognition evaluation for diseases. Recently, an annotated corpus, named Arizona Disease Corpus (AZDC) [13], has been released which has adequate and suitable annotation of disease mentions by following specific annotation guidelines.

There has been some work on identifying diseases in clinical texts, especially in the context of CMC medical NLP challenge and i2b2 challenge.

However, as noted by [17], there are a number of reasons that make clinical texts different from texts of biomedical literature, e.g. composition of short, telegraphic phrases, use of implicit templates and pseudo-tables, ... Hence, the strategies adopted for named entity recognition on clinical texts.

As discussed above, systems that achieve high accuracy in recognizing general names in the newswires have not performed as well in the biomedical named entity recognition with an accuracy of 20 or 30 points difference in their F-score measure. There is a need to develop a biomedical name entity recognition system.

In addition, literature shows that classifiers ensemble (combination) approaches is always superior to all the individual classifiers and leads to significant improvement in performances of named entity recognition. So, in this work, we propose biomedical name entity recognition model based on classifiers combination.

4. The Biomedical Named Entity Recognition

Constructing a biomedical named entity recognition solution using a machine learning approach (classifiers combination using the vote based ensemble approach) requires many computational steps including data planning, pre-processing, feature selection and optimization, classification, and evaluation. The specific components included in a given solution vary but they may be viewed as making part of the following groups summarized in Figure 1.

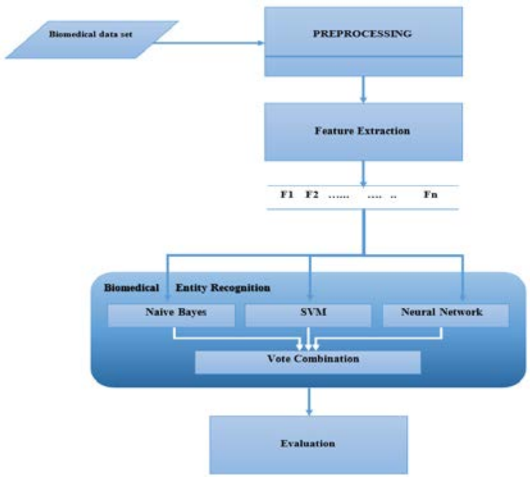

Fig. 1 The Proposed biomedical named entity recognition Architecture

4.1 Preprocessing phase

Using a supervised machine learning technique relies on the existence of annotated training data. Such data is usually created manually by humans or experts in the relevant field. The training data needs to be put in a format that is suitable to the solution of choice. New data to be classified also requires the same formatting. Depending on the needs of the solution, the textual data may need to be tokenized, normalized, scaled, mapped to numeric classes, prior to being fed to a feature extraction module. To reduce the training time with large training data, some techniques such as chunking or instance pruning (filtering) may need to be applied.

4.2 Feature Extraction

In the phase of feature extraction, test data and training is created by one or more components in order to retrieve the important information about it. The selection of feature extraction components involves the extraction of morphological and orthographic based features, text based information, linguistic based information such as POS, and domain-dependent knowledge including specialized gazetteers or dictionaries.

In the phase of feature extraction, test data and training is performed by several components in order to retrieve the important information about it. in order to extract morphological and contextual features that do not use language-specific knowledge such as part-of-speech or noun phrase tagging. The generated feature space is very large, including about a million different features. The features extracted are described below. Since words appearing separately or within windows of other words each constitutes a feature in the lexicon, the potential number of possibilities is very high. Including character n-grams describing prefixes, infixes, and suffixes would further increase the number of features in the lexicon. The feature extraction process is intentionally designed that way in order to test the scalability of the approach used and to allow the experiments to proceed in a language-independent and domain-independent fashion. All features are binary, i.e., each feature denotes whether the current token possesses this feature (one) or not (zero). Character n-grams were not included in the baseline experiment data due to memory limitations encountered during the feature extraction process.

The morphological features extracted are:

- Capitalization: token begins with a capital letter.
- Numeric: token is a numeric value.
- Punctuation: token is a punctuation.
- Uppercase: token is all in uppercase.
- Lowercase: token is all in lowercase.
- Single character: token length is equal to one.
- Symbol: token is a special character.
- Includes hyphen: one of the characters is a hyphen.
- Includes slash: one of the characters is a slash.
- Letters and Digits: token is alphanumeric.
- Capitals and digits: token contains caps and digits.
- Includes caps: some characters are in uppercase.

4.3 Machine Learning and Classification

Almost all of the machine learning based techniques and approaches have two phases, where the training is performed initially to produce a trained machine, and then a classification step is performed. In this study, the following machine learning approaches are evaluated:

4.3.1 Support vector machine (SVM)

A support vector machine (SVM) is a relatively new machine learning technique that has been proposed by Cortes & Vapnik (1995). SVM is generally a popular technique for NER, which is used in the machine learning area. SVM is considered one of the classification techniques with a very high efficiency. Based on the idea of structural-risk minimization, from the computational-learning theory, SVM tries a decision surface, in order to separate the training data nodes into two main classes, and makes decisions based on the existing support vectors, which are selected as the only components that are efficient in the training set.

$$\vec{\alpha} = \text{argmin}\left\{-\sum_{i=1}^{n}\alpha_i + \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j(\vec{x_i},\vec{x_j})\right\}$$

$$\text{Subjectto}: \sum_{i=1}^{n}\alpha_i y_i = 0; \quad 0 \le \alpha_i \le C \qquad (1)$$

4.3.2 Naïve Bayes

The naive Bayes technique is exhaustively used for NER. Given a table of feature vectors, the technique decides the rear possibility, where the term is related to multiple named entity classes, and assigns it to the category with the maximum rear possibility. There are two used approaches: multi-nominal models and multi-variate Bernoulli models. Naïve Bayes is a stochastic model of generating documents makes use of Bayes' rule. To classify as the best named entity class n* for a new term w, it computes:

$$p(c_j|w_i) = \frac{p(c_j)p(c_j|w_i)}{p(w_i)} \qquad (2)$$

4.3.3 Artificial Neural Network

A neural network is a mutual band of artificial neurons, which utilizes a computational model to process data, depending on a connectionist method. Sets of input attribute and preferred results are entered to the learning program. This is aimed at using the input characteristics to segregate the training conditions into non-overlapping models, related to the preferred results. Input layer comprises of a collection of units, identical to the number of tags, in the tag set.

The neural networks we have used is an acyclic directed graph of sigmoid units based on back propagation algorithm. The sigmoid units are like perceptrons, but they are based on a smoothed, differentiable threshold function. A sigmoid unit first computes a linear combination of its input, then applies a threshold to result, where the threshold is a continuous function of its input. The sigmoid unit computes its output o as follows:

$$o = \sigma(\vec{w} * \vec{x}) \qquad (3)$$

where

$$\sigma(y) = \frac{1}{1+e^{-y}} \qquad (4)$$

Here is called the sigmoid function. Its output ranges between 0 and 1, increasing monotonically with its input.

4.4 Performance Measures

The performance measures used to evaluate the named entity recognition systems participating in the CoNLL-02, CoNLL-03 and JNLPBA-04 challenge tasks are precision, recall, and the weighted mean Fβ=1-score. Precision is the percentage of named entities found by the learning system that are correct. Recall is the percentage of named entities present in the corpus that are found by the system. A named entity is correct only if it is an exact match of the corresponding entity in the data file, i.e., the complete named entity is correctly identified. Definitions of the performance measures used are summarized below. The same performance measures are used to evaluate the results of the baseline experiments.

5. Experimental Results

We have conducted several experiments. First, we have performed several experiments to empirically compare different subsets of features and three classification approach (Naïve Bayes, K-Nearest Neighbor and decision tree for biomedical named entity recognition. The aim is to efficiently integrate different feature sets and classification algorithms to synthesize a more accurate classification procedure.

Each subset of features is applied with almost of other features with one of the three classification approaches in each main experiment. All of the algorithms are evaluated by using ten-fold cross-validation. The results in terms of the macro-averaged F-measure are the averaged values calculated across all ten-fold cross-validation experiments.

In this section, will describe several experiments to empirically compare 10 different features and three classification approach (Naïve Bayes, K-Nearest Neighbor and decision tree for biomedical named entity recognition.

We have two primary goals with our experiments in biomedical named entity recognition. The first is to define a better classification approach that will use in the model to classify the dataset. The second is to evaluate the features described in the previous chapter to their usefulness for this task and the better classification model for biomedical named entity recognition.

Table 1 show a sample of the used dataset for the experiments

Word	F1	F2	F3	F4	F5	F6	F7	F9	Label
IL-2	_	_	I	IL	IL-	2	-2	L-2	B-DNA
gene	IL-2	B-DNA	g	ge	gen	e	ne	ene	I-DNA
expression	gene	I-DNA	e	ex	exp	n	on	ion	O
and	expression	O	a	an	and	d	nd	and	O
NF-kappa	and	O	N	NF	NF-	a	pa	ppa	B-protein
B	NF-kappa	B-protein	B	B	B	B	B	B	I-protein
activation	B	I-protein	a	ac	act	n	on	ion	O
through	activation	O	t	th	thr	h	gh	ugh	O
CD28	through	O	C	CD	CD2	8	28	D28	B-protein
requires	CD28	B-protein	r	re	req	s	es	res	O
reactive	requires	O	r	re	rea	e	ve	ive	O
oxygen	reactive	O	o	ox	oxy	n	en	gen	O
production	oxygen	O	p	pr	pro	n	on	ion	O
by	production	O	b	by	by	y	by	by	O
5-lipoxygenase	by	O	5	5-	5-l	e	se	ase	B-protein
.	5-lipoxygenase	B-protein	.	.	.	.	.	.	O
Activation	.	O	A	Ac	Act	n	on	ion	O
of	Activation	O	o	of	of	f	of	of	O
the	of	O	t	th	the	e	he	the	O
CD28	the	O	C	CD	CD2	8	28	D28	B-protein
surface	CD28	B-protein	s	su	sur	e	ce	ace	I-protein
receptor	surface	I-protein	r	re	rec	r	or	tor	I-protein
provides	receptor	I-protein	p	pr	pro	s	es	des	O
a	provides	O	a	a	a	a	a	a	O
major	a	O	m	ma	maj	r	or	jor	O
costimulatory	major	O	c	co	cos	y	ry	ory	O
signal	costimulatory	O	s	si	sig	l	al	nal	O
for	signal	O	f	fo	for	r	or	for	O
T	for	O	T	T	T	T	T	T	O
cell	T	O	c	ce	cel	l	ll	ell	O
activation	cell	O	a	ac	act	n	on	ion	O
resulting	activation	O	r	re	res	g	ng	ing	O
in	resulting	O	i	in	in	n	in	in	O
enhanced	in	O	e	en	enh	d	ed	ced	O
production	enhanced	O	p	pr	pro	n	on	ion	O
of	production	O	o	of	of	f	of	of	O
interleukin-2	of	O	i	in	int	2	-2	n-2	B-protein

In the first experiment, the KNN Classifier is applied on testing set using 10-fold cross-validation. As shown in

Table, there are 9 features which means 512 different experiments can be performed. However, the results here are obtained for the best 10 experiments from these 512 experiments. The idea is to show the best results obtained when the KNN is applied. Table 2 shows the performance in terms of the precision, recall, F-measure of the biomedical named entity recognition by applying the KNN Classifier with different set of features. As shown Table 2, the use of features sets has an obvious effect on the quality of biomedical named entity recognition for KNN Classifier classification model in general.

Table 2 shows the performance in terms of the precision, recall, F-measure of the biomedical named entity recognition by applying the KNN Classifier

KNN											
	previous word	Previous Tag	Prefix 1	Prefix 2	Prefix 3	Suffix 1	Suffix 2	Suffix 3	PRECISION	RECALL	FMEASURE
1	O	O							97.61904	99.12281	98.33801
2	O	O	O	O					97.5	99.09091	98.25923
3					O	O			91.66666	87.41685	89.01163
4			O	O	O	O			81.74603	78.98551	80.10796
5							O	O	77.81609	75	74.35082
6	O	O					O	O	95.38721	96.29121	95.81846
7		O	O	O	O				96.05533	96.05533	96.05533
8	O		O	O	O				95.65218	98.18182	96.80135
9	O			O	O	O			98.07692	99.07408	98.55232
10		O		O	O	O			96.68033	96.68033	96.68033
11			O	O	O				93.05556	85.29411	87.64797
12	O	O	O	O	O	O			97.16783	97.16783	97.16783

In the second experiment, the NB Classifier is applied on testing set using 10-fold cross-validation. The results are obtained for the best 9 experiments from these 512 experiments. The idea is to show the best results obtained when the NB is applied. Table 3 shows the performance in terms of the precision, recall, F-measure of the biomedical named entity recognition by applying the NB Classifier with different set of features. As shown Table 3, the use of features sets has an obvious effect on the quality of biomedical named entity recognition for NB Classifier classification model in general. However, the results obtained using NB classifier is less than that obtained using KNN. It means that effect of the feature sets on the performance of the NB classifier is lower than their effect on KNN Classifier.

Table 3 shows the performance in terms of the precision, recall, F-measure of the biomedical named entity recognition by applying the NB Classifier

Naive Bayes											
	previous word	Previous Tag	Prefix 1	Prefix 2	Prefix 3	Suffix 1	Suffix 2	Suffix 3	PRECISION	RECALL	FMEASURE
1	O	O							0.553	0.649	0.566
2	O	O	O	O					0.667	0.739	0.674
3					O	O			0.475	0.506	0.472
4			O	O	O	O			0.525	0.488	0.478
5							O	O	0.442	0.478	0.44
6	O	O					O	O	0.66	0.762	0.672
7		O	O	O	O				0.706	0.692	0.678
8	O		O	O	O				0.556	0.539	0.52
9	O			O	O	O			0.56	0.544	0.522
10		O		O	O	O			0.709	0.712	0.691
11			O	O	O				0.518	0.446	0.457
12	O	O	O	O	O	O			0.664	0.718	0.662

In the third experiment, the decision tree Classifier is applied on testing set using 10-fold cross-validation. The results are obtained for the best 9 experiments from these 512 experiments. The idea is to show the best results obtained when the decision tree is applied. Table 4 shows the performance in terms of the precision, recall, F-measure of the biomedical named entity recognition by applying the decision tree Classifier with different set of features. As shown Table 4, the use of features sets has an obvious effect on the quality of biomedical named entity recognition for decision tree Classifier classification model in general. However, the results obtained using decision tree classifier is less than that obtained using KNN. It means that effect of the feature sets on the performance of the decision tree classifier is lower than their effect on KNN Classifier.

Table 4 shows the performance in terms of the precision, recall, F-measure of the biomedical named entity recognition by applying the decision tree Classifier

Decision tree id3											
	previous word	Previous Tag	Prefix 1	Prefix 2	Prefix 3	Suffix 1	Suffix 2	Suffix 3	PRECISION	RECALL	FMEASURE
1	O	O							60.47168	67.98289	61.17422
2	O	O	O	O					74.47727	77.89099	73.93359
3					O	O			54.85217	62.07214	56.07804
4			O	O	O	O			54.85217	62.07214	56.07804
5							O	O	43.48093	56.90242	47.00568
6	O	O					O	O	75.86456	78.40318	75.11893
7		O	O	O	O				72.92546	80.08942	75.17076
8	O		O	O	O				65.9024	68.84908	63.74495
9	O			O	O	O			67.50764	70.97287	65.68054
10		O		O	O	O			74.00443	80.57849	75.93639
11			O	O	O				49.2048	52.84519	49.49225
12	O	O	O	O	O	O			78.57255	80.74608	77.79234

6. Conclusion

The core objective of this work is to design and implement a new model for biomedical named recognition. A new model is produced based on support vector machine (SVM), Naïve Bayes (NB), and Artificial Neural Network. The machine learning techniques have been used for building and developing biomedical named recognition which requires several steps, including data pre-processing, feature selection and extraction, machine learning models, and classification. The reported results analysis shows that the proposed model is satisfactory and effective for biomedical named recognition

References

[1] Habib, M. S. Biomedical Named Entity Recognition Using Support Vector Machines: Performance vs. Scalability Issues.

[2] Zhang, S., & Elhadad, N. (2013). Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. Journal of biomedical informatics, 46(6),1088-1098.

[3] Chowdhury, M., & Mahbub, F. (2013). Improving the Effectiveness of Information Extraction from Biomedical Text. University of Trento.

[4] Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research, 32(suppl 1), D267-D270.

[5] Dai, H.-J., Chang, Y.-C., Tsai, R. T.-H., & Hsu, W.-L. (2010). New challenges for biological text-mining in the next decade. Journal of computer science and technology, 25(1), 169-179.

[6] Liu, H., Aronson, A. R., & Friedman, C. (2002). A study of abbreviations in MEDLINE abstracts. Paper presented at the Proceedings of the AMIA Symposium.

[7] Saha, S. and A. Ekbal (2013). "Combining multiple classifiers using vote based classifier ensemble technique for named entity recognition." Data & Knowledge Engineering 85: 15-39.

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

[9] Munkhdalai, T., Li, M., Batsuren, K., Park, H., Choi, N., & Ryu, K. H. (2015). Incorporating domain knowledge in chemical and biomedical named entity recognition with word representations. J. Cheminformatics, 7(S-1), S9.

[10] Rocktäschel, T., Weidlich, M., & Leser, U. (2012). ChemSpot: a hybrid system for chemical named entity recognition. Bioinformatics, 28(12), 1633-1640.

[11] Wang, H. (2008). "Biomedical Named Entity Recognition Based on Classifiers Ensemble." International Journal of Computer Science and Applications (IJCSA).

[12] Torii, S., Saito, N., Kawano, A., Hou, N., Ueki, K., Kulkarni, R. N., & Takeuchi, T. (2009). Gene silencing of phogrin unveils its essential role in glucose-responsive pancreatic β-cell growth. Diabetes, 58(3), 682-692.

[13] Leaman, R., Miller, C., & Gonzalez, G. (2009). Enabling recognition of diseases in biomedical text with machine learning: corpus and benchmark. Paper presented at the Proceedings of the 2009 Symposium on Languages in Biology and Medicine.

[14] Rosario, B., & Hearst, M. A. (2004). Classifying semantic relations in bioscience texts. Paper presented at the Proceedings of the 42nd annual meeting on association for computational linguistics.

[15] Kulick, S., Bies, A., Liberman, M., Mandel, M., McDonald, R., Palmer, M., . . . White, P. (2004). Integrated annotation for biomedical information extraction. Paper presented at the Proc. of the Human Language Technology Conference and the Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL).

[16] Jimeno, A., Jimenez-Ruiz, E., Lee, V., Gaudan, S., Berlanga, R., & Rebholz-Schuhmann, D. (2008). Assessment of disease named entity recognition on a corpus of annotated sentences. BMC Bioinformatics, 9(Suppl 3), S3.

[17] Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C., & Hurdle, J. F. (2008). Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform, 35, 128-144.

Ahmed Sultan Al-Hegami received his B.Sc degree in Computer Science from King Abdul Aziz University, Saudi Arabia, MCA (Master of Computer Application) from Jawaharlal Nehru University, New Delhi, India; and Ph.D. degree in Data Mining from University of Delhi, Delhi, India. He is Associate professor of Artificial Intelligence and Intelligent Information Systems at the Faculty of Computers and Information Technology, Sana'a University, Yemen. His research interest includes artificial intelligence, machine learning, temporal databases, real time systems, data mining, and knowledge discovery in databases.

Ameen Othman received the B.E. degrees in computer science from Sana'a University, Yemen in 2003. During 2003-2007, he worked a programmer in Compuware for IT, Yemen. He is now working in Fisheries Information Center (FIC), Ministry of Fisheries of Yemen

Fuad Tarbosh Bagash B.E degree in Production Engineering from Technology University , Iraq , MSC.IT (Master of Information Technology) India.. Ph.D. Scholar at European Academy for the higher studies,Yemen.. His research interest includes Web mining,web technology, cloud computing, web programming