Original Article
International Journal of Fuzzy Logic and Intelligent Systems
Vol. 15, No. 2, June 2015, pp. 111-120
http://dx.doi.org/10.5391/IJFIS.2015.15.2.111

ISSN(Print) 1598-2645
ISSN(Online) 2093-744X

Protein Named Entity Identification Based on Probabilistic Features Derived from GENIA Corpus and Medical Text on the Web

Sagara Sumathipala , Koichi Yamada, Muneyuki Unehara and Izumi Suzuki
Graduate School of Engineering, Nagaoka University of Technologyy, 1603-1, Kamitomioka-machi, Nagaoka, Niigata 940-2188, Japan

Abstract

Protein named entity identification is one of the most essential and fundamental predecessor for extracting information about protein-protein interactions from biomedical literature. In this paper, we explore the use of abstracts of biomedical literature in MEDLINE for protein name identification and present the results of the conducted experiments. We present a robust and effective approach to classify biomedical named entities into protein and non-protein classes, based on a rich set of features: orthographic, keyword, morphological and newly introduced Protein-Score features. Our procedure shows significant performance in the experiments on GENIA corpus using Random Forest, achieving the highest values of precision 92.7%, recall 91.7%, and F-measure 92.2% for protein identification, while reducing the training and testing time significantly.

Keywords: biomedical text mining, named entity recognition, protein named entity, random forest

1. Introduction

The evolution of biomedical text produces a strong demand for automated text mining techniques that can facilitate biomedical researchers to gather and make use of the knowledge in biomedical literature. Figure 1 shows the rapid increase of the number of publications in the MEDLINE [1] database from 1940 until 2014. Protein Named Entity identification is one of the most essential and fundamental predecessor for identification of protein-protein relationships [2, 3], keep protein databases such as UniProtKB [4] up-to-date and many more. However, manual approaches targeting this task are extremely time-consuming, expensive and laborious work, which has led an increasing attention towards automated approaches to help ease these tasks.

Biomedical Named Entity Recognition (BNER) focuses on extraction of words or phrases referring to Biomedical Named Entities (BNEs) in biomedical text and classifying them into appropriate biomedical entity classes such as proteins, genes, DNAs and drug names. However Named Entity Recognition (NER) approaches for biomedical literature do not perform well than those focusing on general text such as newswire domain [5, 6]. BNER is a difficult task because: 1) BNEs contain highly complex vocabulary and are rapidly evolving, 2) most of the BNEs are compound terms and may or may not possess a suffix or a prefix, 3) combination of BNEs may form another BNE, 4) they may have many abbreviations, 5) different aliases can

Received: Apr. 9, 2015
Revised : Jun. 24, 2015
Accepted: Jun. 25, 2015

Correspondence to: Izumi Suzuki
(suzuki@kjs.nagaokaut.ac.jp)

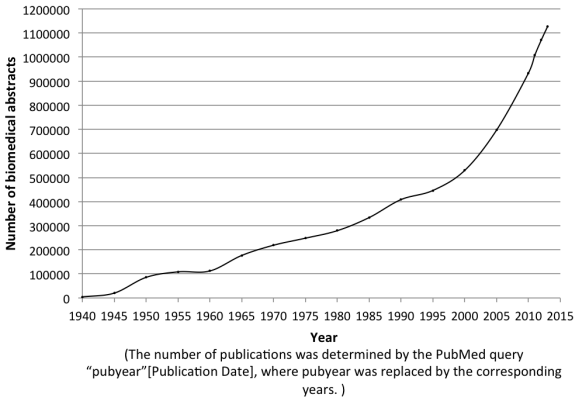

Figure 1. Yearly number of MEDLINE publications from 1940 to 2014.

be used to refer the same BNE where the type depends on the context where they exists, etc.

For example *"TNF alpha"* can refers to a protein entity as well as a DNA. Indeed, even humans agree only about 77.0% in the case of protein, gene and RNA classification [7]. In addition, BNEs are written using a mix of letters, numbers, Greek letters and symbols, which make it further complicated for computers to identify those with automated approaches.

Much work has been done to develop robust and effective protein identification approaches. They could be classified into dictionary based, heuristic rule based, Machine Learning (ML) based and hybrid approaches. Dictionary based approaches use existing protein dictionaries and/or protein databases to extract and identify proteins in biomedical literature [8]. Due to the different naming schemes referring to the same protein, dictionary based approaches are not effective in identifying them. Even though dictionary based approaches achieve high classification performance, predefined biomedical dictionaries are not able to identify newly introduced protein names.

Rule based approaches use generated rules to identify protein names in biomedical literature [9]. These approaches need domain experts' knowledge to derive these rules and often time-consuming. ML based BNER approaches mostly use supervised learning algorithms such as Hidden Markov Model (HMM), Support Vector Machines (SVM), Maximum Entropy Model (ME) and Conditional Random Fields (CRFs) [10–16]. However, unsupervised learning models are also proposed [17]. State of the art ML based BNER approaches are dominating the other approaches and can be improved further.

There were several two-step BNER approaches proposed, where extraction of BNEs in text and classification are done in two separate stages. This helps to reduce the training time and also to select more relevant features for each stage [14, 15]. Even though BNEs in text are identified successfully, classifying them into relevant biomedical entity classes still remains as a challenging task. Therefore, identifying protein names is still an open and important task. This paper presents a new method with the highest protein identification performance among BNEs, while significantly reducing the training and testing time.

This paper proposes a statistical feature called Protein-Score, which could be understood as the likelihood that the term "`Protein`" appears in a MEDLINE abstract with the given BNE. Then, BNEs are denoted by a new set of features including orthographic, keyword, morphological as well as the newly proposed Protein-Score features, and are classified into protein and non-protein entities using Random Forest (RF) [18], a ML technique getting attention lately. A series of experiments is carried out to compare the results with those of the other state of the art approaches. Our Protein Named Entity Recognition (PNER) model based on RF achieved the best performance via experiments on GENIA corpus while significantly reducing the training and testing time.

We discussed the possibility of using the unithood and MEDLINE statistics for protein identification in Sumathipala et al. (2014) [19], where we showed the probabilistic capability of using sub-terms of BNEs to predict whether or not the type of the BNE is a protein. In this study we extend the previous work by further modifying the Protein-Score feature using likelihood measure and introducing new additional features. Here we use the RF machine learning algorithm and show the F-measure value over existing solutions.

The rest of this paper is organized as follows: the next section reviews some related works. In Section 3, we discuss the features used in the study and propose a set of features including the new web-based feature called Protein-Score. In Section 4, we evaluate the effect of the proposed features in identification of Protein Named Entities (PNEs) using RF. We conclude the paper by a summary and directions for future work.

2. Related Work

Krauthammer et al. (2000) described a dictionary based system which can automatically identifies gene and protein names in journal articles [8]. Their system was based on BLAST, a popular DNA and protein sequence comparison tool, and on a database of gene and protein names. They achieved a recall

of 78.8% and precision of 71.7% for gene and protein name matching.

Seki et al. (2005) proposed a rule based approach for identifying PNEs in biomedical literature with an emphasis on protein boundary expansion [9]. Their method used surface clues to detect potential protein name fragments. They achieved F-measure of 63.7% for exact protein matches while achieving F-measure around 81.6% for partial matches. Kuo et al. (2014) proposed a protein name identification model from biological literature and achieved F-measure of 80.6% on GENIA corpus [20]. They used N-gram language model to determine the protein name boundaries and some rules were used to improve the performance. In addition, a dictionary was used to strengthen recognition of abbreviations.

Tater et al. (2009) proposed the use of two different machine learning techniques for protein name extraction [21]. First, they used Bigram language model to extract protein names and then, an automatic rule learning method was used which can identify protein names located in the biological texts. They achieved an F-measure of 66.8% on the GENIA corpus. Zhou et al. (2005) proposed an ensemble of classifiers for protein and gene identification in text, based on a SVM and two discriminative HMMs, which were combined using simple majority voting strategy [10]. They achieved the best F-measure of 82.6% for protein and gene name recognition task.

Finkel et al. (2005) presented a maximum-entropy based approach for identify gene and protein names in biomedical abstracts [11]. They used diverse set of features including "*word features, bigrams, abbreviations, word shape features*", etc. and achieved F-measure of 83.6%. Mitsumori et al. (2005) proposed Gene and protein name recognition system based on SVM [12]. They used a feature set of the word, part-of-speech, orthography, prefix, suffix, preceding class and dictionary matching features. They achieved F-measure of 79.2% for Gene/Protein name recognition. Ju et al. 2011 used SVM to recognize BNEs in biomedical literature [13]. They used two kinds of features which are orthographic and part-of-speech features to identify biomedical and non-biomedical named entities using SVM. They evaluated the method on GENIA corpus and achieved precision and recall, 84.2% and 80.8% respectively.

Yang et al. 2013 presented a two-phase approach for BNER based on semi-Markov Conditional Fields [14]. They used a rich set of features including orthographic, morphological, part-of-speech, features etc. Their experiments based on JNLPBA04 dataset showed 77.7% F-measure for protein name identification. Li et al. 2009 proposed a two-phase BNER approach based on CRFs [15]. First, they identified each BNE with CRFs without considering its biomedical class type and at the second stage they used another CRFs model to determine the relevant class type for each identified BNE. Their experiments achieved overall F-measure of 76.0% for PNE identification on JNLPBA04 corpus. Lin et al. 2004 proposed a hybrid method that uses ME as the ML method incorporated with rule based and dictionary based approaches for post processing [16]. They used orthographic, head noun, morphological and part-of speech features and achieved overall F-measure of 78.5% for protein name identification on GENIA corpus.

Zhange et al. 2013 have proposed an unsupervised learning technique to identify BNEs including proteins [17]. They have achieved F-measure of 67.2% for protein identification on GENIA corpus.

3. Protein Name Identification Features

Feature selection is crucial to the success of ML based PNER systems. In this section, we describe the features used in our system. We utilize orthographic, keyword, morphological features as well as Protein-Score feature based on citations for biomedical abstracts from MEDLINE cited in PubMed. In this paper, we refer both biomedical and non-biomedical named entities as "**terms**". A term is composed with one or more "**words**"such as a single letter, a series of letters, a digit or a series of digits. "**word**" is called a sub-term if it is not a single letter word or a single digit word. If a term is composed of multiple sub-terms, they are separated by a space or a hyphen.

3.1 Orthographic Features

Orthographic features are used to capture knowledge about word orientation such as capitalization, digitalization and other word formation information. These features have been widely used in both biomedical domain [9, 13, 22, 23] and non-biomedical domain [24, 25]. Orthographic features used in this study are: 1) whether or not the term contains at least a capital letter, 2) whether or not the term is composed only of capital letters or only of small letters, 3) whether or not the term contains at least a numerical digit, 4) whether or not the term contains a hyphen and 5) the number of words the term contains.

3.2 Keyword Feature

There are many words appearing frequently in BNEs, most of them are compound as mentioned above. Such words (called

Table 1. Keywords used in this study

Keywords				
activator	adhesion	aiolos	alf1	alpha
amino	antigen	beta	bsap	calcineurin
cd	cells	chain	chemokine	chimeric
ciita	coactivator	creb	cytokine	cytoplasmic
e2f	epo	er	erythropoietin	ets
factor	fas	fos	gamma	gata
gcr	gr	heterodimer	homodimer	ifn
ige	ikappab	isoform	interferon	interleukin
isoform	jak	janus	jnk	jun
kappa	kappab	kda	kinase	latent
lmp1	mab	mapk	monoclonal	mutant
necrosis	nef	nf	nfat	nuclear
phosphatase	pkc	prb	protein	rar
ras	receptor	rxr	sp1	tat
tcr	tnf	transactivator	transcription	tumor
vdr				

Table 2. Prefixes used in this study

Prefixes							
alpha	anti	beta	calcin	cytoki	gamma	glu	activ
transc	granul	inte	nuclea	protei	tumor	rec	

Table 3. Suffixes used in this study

Suffixes						
ain	alpha	and	as	ase	tk	odimer
asome	atase	ator	ax	bunit	rm	form
ceptor	chain	ctin	cule	dimer	ta	ras
dy	ectin	eta	ex	factor	tase	mokine
ferase	gamma	gen	genase	ger	tax	mutant
globin	in	inase	it	kappa	tease	nase
kine	lase	ligand	ls	ly	rin	neurin
ntigen	rase					

keywords in this paper) appears very often in a similar group of BNEs can be used to distinguish certain biomedical entity classes. For example, keywords such as '-protein-', '-alpha-', '-factor-' and '-receptor-' are frequently appeared in PNEs. We use the number of keywords in Table 1 contained in the given BNE as the keyword feature.

3.3 Morphological Features

PNEs frequently have a prefix and/or a suffix. Suffixes/prefixes provide important clues for discriminating protein and non-protein entities. For example BNEs ending with *"-ase"*, *"-globin"* are usually PNEs. We use suffixes in Table 3 and prefixes in Table 2, which are common in PNEs. They are used as two binary features by considering whether or not the given BNE starts/ends with a prefix/suffix in Table 2 and Table 3 respectively.

3.4 Protein-Score Feature

External features of a term, which are features calculated from data external to the corpus, might provide additional evidential clues for classification of the term. Presence of the term in an external database such as UniProt, external gazetteers and other resources could be an external feature that enhances the performance of BNER. We chose a number of biomedical abstracts from MEDLINE cited in PubMed as the external resource.

Suppose W represents a BNE composed of an ordered set of sub-terms $w_1, ..., w_k, ..., w_K$, where K is the number of sub-terms and $1 \leq k \leq K$. When $K = 1$, the sub-term itself is a term. The conditional probability that W appears in a MEDLINE abstract with the word "`Protein`", or equivalently the likelihood that the word "`Protein`" appears in a MEDLINE abstract with W is given by:

$$P_{ML}(W \mid W_P) = P_{ML}(w_1, ..., w_K \mid W_P), \quad (1)$$

where W_P represents the word "`Protein`". In order to simplify the equation, we ignore the order of sub-terms as well as we assume that sub-terms in W are independent of each other given W_P.

$$P_{ML}(W \mid W_P) = P_{ML}(w_1 \mid W_P) \times ... \times P_{ML}(w_K \mid W_P). \quad (2)$$

Suppose $H(w_k, W_P)$ is the number of MEDLINE abstracts containing both sub-term w_k and W_P, and that $H(W_P)$ is the number of abstracts containing W_P.

$$P_{ML}(W \mid W_P) = \frac{H(w_1, W_P)}{H(W_P)} \times ... \times \frac{H(w_K, W_P)}{H(W_P)}. \quad (3)$$

Then we define $Protein - Score(PS)$ as follows:

$$PS = K. \ln H(W_P) - \{\ln H(w_1, W_P) + ... + \ln H(w_K, W_P)\}. \quad (4)$$

PS could be understood as the likelihood that the abstract with the BNE W is one mention "`Protein`". The reason why $P(W \mid W_P)$ is used instead of $P(W_P \mid W)$ is that the number of abstracts with W is expected to be far fewer than that with W_P. Computing could also be easier with $P(W \mid W_P)$ than with $P(W_P \mid W)$ taking into account that we should calculate it for many BNEs, because we should examine only abstracts with "`Protein`".

Table 4 shows several sub-terms used in this study and corresponding $H(w_k, W_P)$ values from MEDLINE queried on the date 03/09/2014. Table 5 presents several BNEs in GENIA

Table 4. Several sub-terms used and corresponding $H(w_k, W_P)$ values. $H(W_P) = 1,846,091$

Sub-Term (w_k)	$H(w_k, W_P)$
ick	42
ie2	278
antiestrogen	827
uracil	1348
glycosylase	1356
2r	1665
u937	2930
globin	3190
cytomegalovirus	4846
1a	7693
immunodeficiency	14029
hiv	22072
cytokine	39470
il	60355
calcium	64667
virus	104733
tissue	163521
alpha	193269
mrna	193554
dna	207537
human	402981
gene	407818
cell	601447

Table 5. Several BNEs and estimated Protein-Score values

Biomedical Entity	**Type**	**PS**
Calcium	Atom	3.35
Human tissue	Tissue	3.95
HIV	Virus	4.43
IL-2 gene	DNA	4.93
Cytokine gene	DNA	5.36
U937 cell	Cell	7.57
Antiestrogen	Lipid	7.71
Immunodeficiency virus	Virus	7.75
Globin gene	DNA	7.87
IL-2R alpha mRNA	RNA	14.94
ICK-1A	Protein	16.17
Human Cytomegalovirus IE2 Protein	Protein	16.27
Uracil-DNA glycosylase	Protein	16.62

corpus and their estimated PS values.

In this section we discussed nine different features including five orthographic, a keyword, a suffix, a prefix and the PS features, which are used to encode BNEs to classify them into protein and non-protein classes using ML algorithms.

4. Implementation and Evaluation

In this section we discuss resources and methods employed in our experiments, following the workflow of ML based PNER model. We also discuss the performance of our PNER system using the proposed features and compared with some of the existing solutions.

Table 6. Feature Statistics

Feature	**'YES'%**	**'NO'%**
At least one capital letter	60.5%	39.5%
Only of capital or only of small letters	20.3%	79.7%
At least one numerical digit	27.9%	72.1%
Contains a hyphen	11.5%	88.5%
Contains a suffix	17.6%	82.4%
Contains a prefix	8.3%	91.7%

	Average	**Standard Deviation**
Number of words	2.03	1.28
Number of keywords	0.88	1.14
Protein-Score	8.80	4.91

4.1 Tools and Resources

In our experiments, Weka 3.6.12 was used to implement the ML algorithm [26]. GENIA corpus was used as the lexical resource of BNEs during training and testing. It contains MEDLINE abstracts, selected using a PubMed query for the three MeSH (Medical Subject Headings) terms *"human"*, *"blood cells"*, and *"transcription factors"*. It is the largest annotated resource for BNEs in the molecular biology domain. It also contains 36 annotated biomedical concept classes for BNER task.

PubMed was used to access the MEDLINE database for biomedical abstracts. It is a free resource which comprises more than 24 million citations for biomedical literature from MEDLINE, life science journals and online books [27]. MEDLINE covers journal citations and abstracts for biomedical literature from around the world [1].

Hardware configuration we used are as follows: Intel ®Core ™ i7 CPU, clock speed 3.50 GHz, Memory (RAM) 32 GB 1600 MHz DDR3, System type: OSX Version 10.9.5 Operating System.

4.2 Experiment and Evaluation

We carried out experiments to classify BNEs annotated in GENIA corpus into protein and non-protein classes. We have extracted 92,512 BNEs in total from GENIA including 34,221 PNEs.

Table 6 shows some statistical information of the nine features of BNEs extracted from GENIA.

BNEs in GENIA corpus with protein annotations, which are enclosed by XML semantic tags prefixed with `G#protein` are considered as PNEs and all others as non-protein named entities. These non-protein named entities belong to different biomedical entity classes such as cell, atom, DNA, Virus, RNA. Ten-fold cross validation was used to evaluate the ML algorithm. In each

iteration, nine-tenth of BNEs was used for training and the rest was used for testing.

PS feature plays a vital role in our PNE identification model. In training and testing PS was estimated using more than 24 million biomedical abstracts in MEDLINE accessed through PubMed. Total number of sub-terms taken from all the BNEs are 8,247. Average number of sub-terms in a PNE is around 2. We employed well-known RF machine learning technique with the nine proposed features. In Table 7, we summarize the results from RF algorithm. Rows of the confusion matrix in Table 7, correspond to actual classes and its columns correspond to the predicted classes. Figure 3 shows the overall architecture of our proposed approach.

Ensemble methods have become a popular and widely used tool within the past few years in the field of bioinformatics, because they are applicable in high dimensional problems with complex interactions [28–31]. In the study of [19], we conducted experiments using several ML techniques, and RF performed the best. RF is a powerful classification algorithm in the group of ensemble learning and obtains growing attention on these days. It uses an ensemble of unpruned Decision Trees called CART [32], each of which is constructed on bootstrap sampling of the training data set based on randomly selected subset of features [18]. The major parameters of RF are the number of trees (T) in the forest and number of features randomly selected at each tree (m). We consider different parameter settings for the values of $T = \{20, 40, ..., 100, 200\}$ and $m = \{3, 4, ..., 9\}$ and selected the best performing configuration; $m = 9$ and $T = 60$ by using the well-known grid search algorithm as shown in Figure 2. Since we use all the nine features for 60 trees, the RF is almost the same as another ensemble learning called Bagging with CART [33].

In order to analyze the impact of various other techniques and compare the ultimate results of our approach with other existing solutions, we use common evaluation measures: precision (P), recall (R) and F-measure (F). These measures are formulated as follows:

$$P = \frac{TP}{TP+FP}, R = \frac{TP}{TP+FN}, F = 2.\frac{PR}{P+R},$$

where TP is the number of true positives retrieved, FP is the number of false positives retrieved and FN is the number of false negatives.

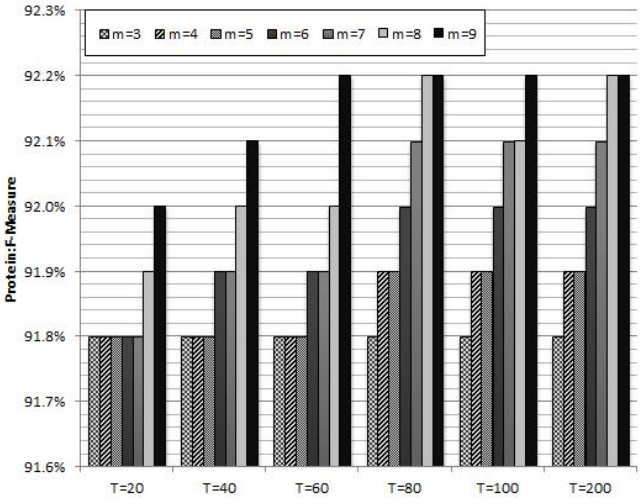

Figure 2. Protein F-measure with different parameters of RF.

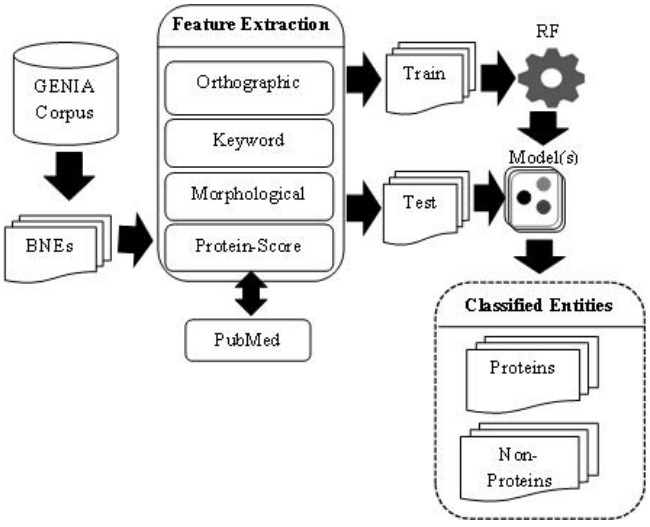

Figure 3. Overall architecture of the proposed approach

4.3 Results and Discussion

In our experiments, RF classifier achieved the best results yielding the overall precision, recall and F-measure values of 92.7%, 91.7% and 92.2% respectively, with the overall classification accuracy of 94.3%. RF has taken 65.6 seconds for training of an iteration of the ten-fold cross-validation.

In order to evaluate the contribution of each feature and subset of features, we carried out experiments on GENIA corpus using RF algorithm. Figure 4 shows the contribution of each feature to the PNER task. According to the figure we can notice that the PS feature plays a vital role in increasing all the measures. As shown in Figure 4, the highest classification performance was obtained with the proposed combination of

Table 7. Testing results on the GENIA corpus by our approaches. **a**: Protein, **b**: Non-Protein

ML Algorithm	Confusion Matrix			F-Measure		Accuracy	Training Time / Seconds	Evaluation Time / Seconds
		a	b	a	b			
Random Forest	a	31381	2840			94.3%	65.6	10.0
	b	2479	55812	0.922	0.955			

Table 8. Our Systems Compared with several existing BNER approaches on the GENIA corpus

			Kazama	Tatar	Zhang	Lee	Patrick	PowerBio	Kuo	Ours	Zhu	Ours
			2002	2009	2013	2004	2005	2004	2014	2014	2012	2015
Reference			[35]	[21]	[17]	[36]	[22]	[23]	[20]	[19]	[34]	-
Features	Orthographic		-	-	-	✓	✓	✓	✓	✓	✓	✓
	Morphological		✓	-	-	✓	✓	✓	✓	✓	✓	✓
	Linguistic		✓	-	-	✓	✓	✓	-	-	✓	-
	Contextual		✓	-	-	✓	✓	-	✓	-	-	-
	Proteinhood		-	-	-	-	-	-	-	✓	-	-
	Protein-Score		-	-	-	-	-	-	-	✓	-	✓
Model	Dictionary Based		-	-	-	-	-	-	✓	-	-	-
	Rule Based		-	✓	-	-	-	-	✓	-	-	-
	ML Based	ME	-	-	-	-	✓	-	-	-	-	-
		HMM / CRF	✓	-	-	-	-	✓	-	-	✓	-
		SVM	✓	-	-	✓	-	-	-	-	✓	-
		RF	-	-	-	-	-	-	-	✓	-	✓
Protein F-Measure %			56.5	66.8	67.2	70.7	74.1	75.8	80.6	88.5	89.0	92.2

features: orthographic, keyword, morphological and PS features. When we use PS feature individually, the m parameter of RF becomes one and it is used in all trees in the forest. Hence, the trees are all the same Decision Stumps with the feature, which means that the effect is equivalent to the Bagging with Decision Stump and PS.

PS can be considered as a dynamic feature because it is based on the number of MEDLINE abstracts cited in the PubMed database, which are gradually increasing. In Sumathipala et al. (2014) [19] we introduced a unithood measure called Proteinhood which quantified the dependency between sub-terms of biomedical term candidates by measuring the probabilistic strength of forming a PNE. Proteinhood values were estimated using the protein sub-terms in the training data set. Therefore the measure might not be effective in identifying PNEs if their sub-terms are not in the training data.

Figure 2 shows the average protein F-measure values of ten trials with different combinations of the parameters in RF: m and T. It suggests a tendency that larger the m value is higher the protein F-measure. The suggestion does not agree with the default value of Weka ($m = \sqrt{numberoffeatures}$), which might be based on the idea that m should be less enough to keep the correlation among trees low [18].

The reason of the inconsistency would be that the assumption does not hold in the problem: the contribution of the features is distributed almost equally among them. Instead, as seen in Figure 4, the only feature PS has a large contribution to the classification in the case [18].

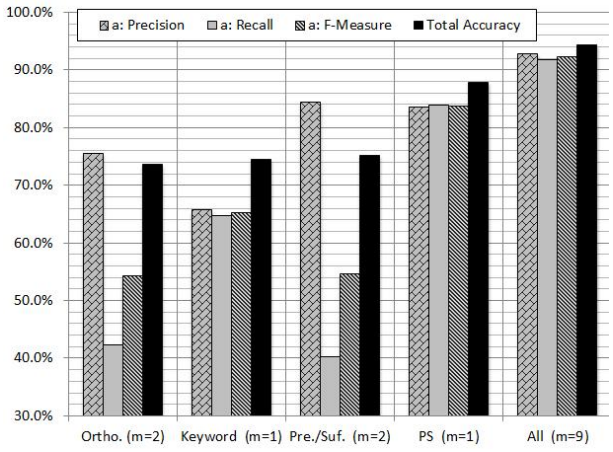

Figure 4. Feature wise contribution for the recognition of PNEs. Orthographic, Keyword, Prefix/suffix and PS represent cases where only those features are used. All shows that all features are used. "m" represents the number of features used in each tree. $T = 60$ is used in all cases.

In general, direct comparison of protein name identification methods is challenging because some classify several kinds of BNEs and the others do not as well as there is a wide variation of both entity classes and test sets used by each research group. In many studies, PNE classification was conducted as a part of large systems aiming to extract BNEs from biomedical literature

and to classify them into several biological classes.

However, the highest performance of our approach shown in Table 8, would never be underestimated, because protein identification is an important task in the molecular biology domain and the approach is itself could be incorporated in a large scale BNER system as well as the same idea of PS could be introduced into BNEs other than proteins.

Our approach outperformed all the other solutions on GENIA corpus, achieving an F-measure of 92.2% for PNER task. It presents an improvement of 3.12% over the second best system we compared, Zhu et al. (2012) which achieved the F-measure of 89.0% on GENIA corpus for PNE identification task [34].

In our approach, we used RF with a small number of features to identify proteins. The experiments show that, our approach takes short training time which shows that it is efficient, effective and economically beneficial.

5. Conclusion and Future Work

In this paper, we presents a PNER approach based on a new set of features including orthographic, morphological and PS features. Our approach outperforms the other state of the art BNER methods, in the view point of protein name identification. We achieved the best performance, which proves the importance of features we used in protein name identification task. We demonstrated the effectiveness of our approach on GENIA corpus, and make comparisons with some related tasks.

Our future work is focused on extending the classification into more biomedical classes including "DNA", "RNA", "CELL-LINE" and "Cell-TYPE".

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

References

[1] MEDLINE®/ PubMed®/ Resources Guide, "http://www.nlm.nih.gov/bsd/pmresources.html"

[2] Bui, Q. C., Katrenko, S., and Sloot, P. M. "A hybrid approach to extract protein-protein interactions." Bioinformatics 27, no. 2 (2011): 259-265.

[3] Blaschke, C., Andrade, M. A., Ouzounis, C. A., and Valencia, A. "Automatic extraction of biological information from scientific text: protein-protein interactions." In Ismb, vol. 7, pp. 60-67. 1999.

[4] UniProtKB, "http://www.uniprot.org/help/uniprotkb"

[5] Ratinov, L., and Roth, D. "Design challenges and misconceptions in named entity recognition." In Proceedings of the Thirteenth Conference on Computational Natural Language Learning, pp. 147-155. Association for Computational Linguistics, 2009.

[6] Sundheim, B. M. "Overview of results of the MUC-6 evaluation." In Proceedings of a workshop on held at Vienna, Virginia: May 6-8, 1996, pp. 423-442. Association for Computational Linguistics, 1996.

[7] Tanabe, L., Xie, N., Thom, L. H., Matten, W., and Wilbur, W. J. "GENETAG: a tagged corpus for gene/protein named entity recognition." BMC bioinformatics 6, no. Suppl 1 (2005): S3.

[8] Krauthammer, M., Rzhetsky, A., Morozov, P., and Friedman, C. "Using BLAST for identifying gene and protein names in journal articles." Gene 259, no. 1 (2000): 245-252.

[9] Seki, K., and Mostafa, J. (2005). "A hybrid approach to protein name identification in biomedical texts". Information processing and management, 41(4), 723-743.

[10] Zhou, G., Shen, D., Zhang, J., Su, J., and Tan, S. "Recognition of protein/gene names from text using an ensemble of classifiers." BMC bioinformatics 6, no. Suppl 1 (2005): S7.

[11] Finkel, J., Dingare, S., Manning, C. D., Nissim, M., Alex, B., and Grover, C. "Exploring the boundaries: gene and protein identification in biomedical text." BMC bioinformatics 6, no. Suppl 1 (2005): S5.

[12] Mitsumori, T., Fation, S., Murata, M., Doi, K., and Doi, H. "Gene/protein name recognition based on support vector machine using dictionary as features." BMC bioinformatics 6, no. Suppl 1 (2005).

[13] Ju, Z., Wang, J., and Zhu, F. (2011, May). "Named entity recognition from biomedical text using SVM". In Bioinformatics and Biomedical Engineering,(iCBBE) 2011 5th International Conference on (pp. 1-4). IEEE.

[14] Yang, Li, and Yanhong Zhou. "Exploring feature sets for two-phase biomedical named entity recognition using semi-CRFs." Knowledge and Information Systems (2013): 1-15.

[15] Li, L., Zhou, R., and Huang, D. “Two-phase biomedical named entity recognition using CRFs.” Computational biology and chemistry 33, no. 4 (2009): 334-338.

[16] Lin, Y. F., Tsai, T. H., Chou, W. C., Wu, K. P., Sung, T. Y., and Hsu, W. L. “A maximum entropy approach to biomedical named entity recognition. ” In BIOKDD, pp. 56-61. 2004.

[17] Zhang, S., and Elhadad, N. “Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts.” Journal of biomedical informatics 46, no. 6 (2013): 1088-1098.

[18] Breiman, L. “Random forests.” Machine learning,(2001), 45:5-32.

[19] Sumathipala, S., Yamada, K., and Unehara, M. “Protein Named Entity Classification with Probabilistic Features Derived from GENIA Corpus and MEDLINE”, Joint 7th International Conference on Soft Computing and Intelligent Systems and 15th International Symposium on Advanced Intelligent Systems (2014): 1257-1261, Japan

[20] Kuo, H. C., and Lin, K. I. “Extracting Protein Names from Biological Literature.” Advances in Computer Science: an International Journal 3, no. 2 (2014): 58-68.

[21] Tatar, S., and Cicekli, I. “Two learning approaches for protein name extraction.” Journal of biomedical informatics 42, no. 6 (2009): 1046-1055.

[22] Patrick, J., and Wang, Y. “Biomedical named entity recognition system.” In Proceedings of the Tenth Australasian Document Computing Symposium (ADCS 2005). 2005.

[23] Zhou, G., Zhang, J., Su, J., Shen, D., and Tan, C.“Recognizing names in biomedical texts: a machine learning approach.” Bioinformatics 20, no. 7 (2004): 1178-1190.

[24] Liu, X., Zhang, S., Wei, F., and Zhou, M. “Recognizing named entities in tweets.” In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, pp. 359-367. Association for Computational Linguistics, 2011.

[25] Chieu, H. L., and Ng, H. T.“Named entity recognition: a maximum entropy approach using global information.” In Proceedings of the 19th international conference on Computational linguistics-Volume 1, pp. 1-7. Association for Computational Linguistics, 2002.

[26] Witten, I. H., and Frank, E.“Data Mining: Practical machine learning tools and techniques.” Morgan Kaufmann, 2005.

[27] PubMed Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2005-. PubMed Help. [Updated 2014 Mar 25], “http://www.ncbi.nlm.nih.gov/books/NBK3827/”

[28] Chen, X., and Ishwaran, H. (2012). “Random forests for genomic data analysis”. Genomics, 99(6), 323-329.

[29] Boulesteix, A. L., Janitza, S., Kruppa, J., and Knig, I. R. (2012). “Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics.” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(6), 493-507.

[30] Okun, O., and Priisalu, H. (2007). “Random forest for gene expression based cancer classification: overlooked issues”. In Pattern Recognition and Image Analysis (pp. 483-490). Springer Berlin Heidelberg.

[31] Yang, P., Hwa Yang, Y., B Zhou, B., and Y Zomaya, A. (2010). “A review of ensemble methods in bioinformatics”. Current Bioinformatics, 5(4), 296-308.

[32] Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). “Classification and regression trees”. CRC press.

[33] Breiman, L. (1996). “Bagging predictors”. Machine learning, 24(2), 123-140.

[34] Zhu, F., and Shen, B.“Combined SVM-CRFs for biological named entity recognition with maximal bidirectional squeezing.” PloS one 7, no. 6 (2012): e39230.

[35] Kazama, J. I., Makino, T., Ohta, Y., and Tsujii, J. I. “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, pp. 1-8. Association for Computational Linguistics, 2002.

[36] Lee, K. J., Hwang, Y. S., Kim, S., and Rim, H. C. “Biomedical named entity recognition using two-phase model based on SVMs.” Journal of Biomedical Informatics 37, no. 6 (2004): 436-447.

[37] PubMed, http://www.ncbi.nlm.nih.gov/pubmed/

Sagara Sumathipala born in 1983, Ph.D. candidate in Information Systems and Control Engineering at Nagaoka University of Technology. He received BSc. degree in Computer Science and Technology from Sabaragamuwa University of Sri Lanka in 2009 and M.Eng. degree in Management and Information Systems Engineering from Graduate School of Engineering, Nagaoka University of Technology in 2012. His research interests include Bioinformatics and Biomedical Text Mining.

E-mail : s105089@stn.nagaokaut.ac.jp

Koichi Yamada received B.Eng. in Control Engineering in 1978, and M.Eng. and PhD in Systems Science in 1980 and 1996 respectively, all from Tokyo Institute of Technology, Japan. He worked as a research engineer for Mitsubishi Research Institute, Inc. from 1980 to 1987, Digital Equipment Corporation Japan, Inc. from 1987 to 1989, and Yamatake-Honeywell Co., Ltd from 1989 to 1996. He also worked as the leader of an intelligent human interface project in Laboratory for International Fuzzy Engineering Research developed by MITI, Japan between 1991 and 1993. In 1996 he joined Nagaoka University of Technology as an associate professor and is currently a full professor in Department of Information and Management Systems Engineering. He was a board member of Japan Society for Fuzzy Theory and Intelligent Informatics in 7th, 8th and 11th terms. His main research interests are automated reasoning, machine learning, decision making under uncertainty, human-computer interactions and affective Engineering.

E-mail : yamada@kjs.nagaokaut.ac.jp

Muneyuki Unehara received his B.E. degree from Collage of Engineering Systems, University of Tsukuba in 1999, M.S. degree in Engineering from Graduate School of Science and Engineering, University of Tsukuba in 2002. He is also received PhD in Engineering from Graduate School of Systems and Information Engineering, University of Tsukuba in 2005. He is currently an assistant professor in Department of Information and Management Systems Science, Nagaoka University of Technology. His research interests are intelligent systems especially human centered systems with Kansei (feeling), interactive systems, and generative systems with evolutionary algorithm. He is a member of Japan Society for Fuzzy Theory and Intelligent Informatics, Japan Society of Kansei Engineering, and the Japanese Society for Artificial Intelligence.

E-mail : unehara@kjs.nagaokaut.ac.jp

Izumi Suzuki received B.A. in Science (Mathematics) from Hokkaido University in 1991, M.A. in Science (Mathematics) from Niigata University in 1993, and Doctor of Engineering from Nagaoka University of Technology in 2009. From 2000 to 2007, he worked as a research associate, and since 2008, as an assistant professor, both in Nagaoka University of Technology. His main research interests are image recognition and understanding, and statistical pattern classification.

E-mail : suzuki@kjs.nagaokaut.ac.jp