ENHANCED GRAPH BASED WORD REPRESENTATION FOR BIOMEDICAL NAMED ENTITY RECOGNITION

A. A. Kalilah [1]*
F. A. Alhadsha [1]
M. Albared [1]
S. Alassali [1]

Received: 13/02/2025
Revised: 08/04/2025
Accepted: 09/04/2025

[1] Computer Science Department, Faculty of IT&CS, University of Saba Region, Marib, Yemen
*Corresponding Author's Email: eng.kalilah@gmail.com

Enhanced Graph Based Word Representation for Biomedical Named Entity Recognition

A A Kalilah
Computer Science Department, Faculty of IT&CS, University of Saba Region,
Marib, Yemen
eng.kalilah@gmail.com

F. A. Alhadsha
Computer Science Department, Faculty of IT&CS, University of Saba Region,
Marib, Yemen
alhadsha.f@gmail.com

M. Albared
Computer Science Department, Faculty of IT&CS, University of Saba Region,
Marib, Yemen

S. Alassali
Computer Science Department, Faculty of IT&CS, University of Saba Region,
Marib, Yemen

Abstract— As the biomedical literature continues to expand rapidly, the significance of extracting biomedical-named entities from this extensive body of work is steadily increasing. Bio-NER presents a greater challenge compared to general entity recognition due to the non-standard use of abbreviations, synonyms, homonyms, ambiguities, and the continual creation of new biomedical entity names. These factors combine to create a significant hurdle in the accurate identification and classification of biomedical entities. The underperformance of machine learning models in biomedical text analysis is primarily attributed to the inadequate representation of these texts through manually created features. In addressing this challenge, this study aims to create enhanced word representation methods to improve biomedical named entity recognition and are based on enhanced graph-based word representation techniques, utilizing machine learning approaches: CRF, SVM, and ensemble learning. These methods are assessed using the well-known GENIA corpus. The results show that SVM, CRF and ensemble learning with morphological, orthographic and context features achieves good results with overall F-measure of (54.6%), (81.87%) and (85.64) respectively. In addition, experimental results also show that enhanced graph-based word representation techniques achieve higher performance with overall F-measures (85.62%), (89.69%) and (91.17) respectively. Results show that proposed graph-based word representations significantly improve the overall performance of CRF, SVM, and ensemble learning over traditional feature representation techniques. In general, results show that word representation is a key factor in constructing a suitable recognition method.

Keywords— Biomedical, Named Entity Recognition, Word Representation, Supervised Machine Learning.

I. INTRODUCTION

The continuous expansion of biomedical texts has unleashed an extensive corpus of freely available, yet largely unstructured, biomedical literature. The challenge of keeping up with new discoveries has become more and more difficult as this amount of data continues to grow. Extracting valuable insights from this huge amount of data to locate relevant literature for research in the biomedical field has become a hard task for researchers. As a result, biomedical text mining and knowledge extraction tools can play a vital role in facilitating the extraction of valuable information from biomedical texts. Biomedical-named entity recognition stands as a pivotal component within biomedical mining tools. Its primary objective is the automatic identification and categorization of biomedical-named entities such as genes, proteins, SNPs, chemicals, and drug names from unstructured textual data in the biomedical domain [1]. The extraction and identification of biomedical entities present a greater level of complexity compared to traditional entities. This heightened complexity arises due to the non-standard use of abbreviations, synonyms, homonyms, ambiguities, and the continual creation of new biomedical entity names. [2].

Several approaches have been employed to tackle the challenge of biomedical named entity recognition. These approaches can be broadly classified into rule-based, dictionary-based, and supervised methods. In the early stages of Named Entity Recognition (NER), systems were built using manually crafted rules, lexicons, orthographic features, and ontologies [3, 4]. While these methods offer certain advantages, such as they do not need annotated training data, they also come with notable drawbacks. For instance, lexicons and handcrafted rules must be comprehensive, and the associated dictionaries require continuous updates by domain experts to remain relevant. In contrast, supervised learning methods rely on annotated training data, typically represented with morphological, orthographic, and contextual features that have been carefully chosen by domain experts. The primary challenge with these approaches lies in the manual feature engineering required for each specific dataset [5]. In addition, a crucial concern within these supervised techniques pertains to the selection of data representations. Creating robust word representations that are domain and datasets independent and can capture useful recognition information is a major issue to enhance the performance of these techniques.

Existing literature highlights the pivotal role of crafting effective data representation techniques when developing named entity recognition systems. Thus, Hence, the creation of enhanced graph-based word representation techniques, capable of capturing valuable information for recognition and classification while remaining independent of specific domains and datasets, is essential for substantially improving the performance of machine learning methods in biomedical named entity recognition. To tackle this issue, this study seeks to design effective dataset- and entity-independent graph-based word representation methods to enhance the performance of supervised machine learning methods for biomedical named entity recognition.

The main contributions can be summarized as follows:

- This work evaluates baseline feature extraction and representation with supervised machine learning methods namely Conditional Random Fields (CRF), Support Vector Machines (SVM), and ensemble learning for biomedical named entity recognition.

- This work introduces enhanced graph-based word representation methods for biomedical named entity recognition.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 introduces the materials and methods used in this work. The experimental setting and experimental results are discussed in Section 5. Finally, Section 6 offers some conclusions and suggestions for future work.

II. Literature Review

Several approaches have been used to address biomedical named entity recognition problems which can be categorized as Rule-based approach: In rule-based systems, a set of hand-crafted rules the experts can be built to identify and extract the named entities [6, 17, 18], namely, matches names with a strongly defined morphological and orthographic structure. Dictionary-based approaches: where string-matching methods are used to identify entities in text, are common [4, 6]. Machine learning approaches: When the annotated corpora on biomedical is available. Machine learning approaches are based on statistical models to make predictions about named entities in a given text. These models have their mathematical approaches and techniques for training the corpus, determining the probabilistic values and have their methodologies of working to get the desired result [6]. State-of-the-art Bio-NER approaches use various machine learning algorithms. Each modelling technique uses the feature matrix to create a probabilistic description of the entity name boundaries. Different supervised models have been developed on Bio-NER systems, namely Conditional Random Fields (CRFs) [7], Support Vector Machines (SVMs) [8, 9], Hidden Markov Models (HMMs) [10] and Maximum Entropy Markov Models (MEMMs) [11]. CRFs have been actively used during the last years, since they present several advantages over other methods. Firstly, CRFs avoid the label bias problem, a weakness of MEMMs. In addition, CRFs also have advantages over HMMs, a consequence of their conditional nature that results in the relaxation of the independence assumptions. Finally, although SVMs can provide comparable results, more time is required to train complex models. Semi-supervised solutions use both annotated and unannotated data, in order to solve the data sparseness problem. Thus, the main goal is to collect features of the unannotated data that are not present in the annotated data, which may contribute to a better identification of the entity name boundaries. There are various approaches to implementing semi-supervised solutions [12, 13]. Song, Yu [14] proposed a hybrid dictionary-based entity extraction technique. The proposed technique consists of 1) an approximate string matching technique, 2) a shortest path edit distance technique, and 3) context-enabled text mining techniques. Kuo and Lin [15] achieved F-measures of 80.6%, and 79.7% on the GENIA corpus and the YAPEX corpus respectively for extraction of protein names from biological literature. They used rule-based method to improve protein name prediction rate and N-gram language model to determine the boundary of protein names. In order to enhance the recognition performance of proteins, they used a dictionary to strengthen recognition of abbreviations and words beginning with uppercase. Pilehvar et al. [16] introduced ELMo which generate contextual embeddings by considering the contexts and morphological structures of individual words at each state in text. This way, the embeddings of the same word can vary depending on their syntactical contexts and morphological structures in text.

III. Research Ethodology

The section outlines the methodology for enhancing biomedical named entity recognition through enhanced graph-based word representation. To begin, this work evaluates baseline techniques that involve training and testing machine learning models with datasets structured around traditional features. This phase encompasses a feature engineering task aimed at identifying the traditional features. Subsequently dataset was prepared accordingly. Subsequently, this work explores the development and evaluation of multiple enhanced Biomedical Named Entity Recognition models utilizing the newly proposed graph-based word representation techniques.

A. Baseline Method For Biomedical Named Entity Recognition

In this section, a brief overview is provided of the state-of-the-art in Biomedical Named Entity Recognition models. These models are executed through a series of steps:

1. *Feature Extraction Step*: Each word in both the training and test datasets is transformed into a vector of values using a predefined set of manually crafted features. These features are selected based on existing literature and encompass morphological, orthographic, and contextual characteristics, as detailed in Table 1.
2. *Recognition Step*: In this phase, a range of supervised machine learning models, specifically Conditional Random Fields and Support Vector Machines, are trained and evaluated using the datasets prepared in the previous feature extraction step.

Table 1. Used Features Set

Feature set	Actual features in the feature set
Surrounding words	Four words in the surrounding context: two words before and two words after the current word
Dynamic feature	Dynamic feature denotes the output tag of previous words (tags of two words)
Orthographic features	Orthographic features: Several binary features are defined: initial capital, all capital, includes caps, has slash, has punctuation, has a digit, Start with a dash.
Word affixes	Hyphen suffix, word prefix and suffix character sequences of length up to 4.

B. Enhanced Biomedical Named Entity Recognition Methods

This section provides a detailed description of the enhanced Biomedical Named Entity Recognition models that are being

proposed. These models undergo a sequence of steps. Initially, every word in both the training and test datasets is encoded using one of the proposed graph-based representations. Subsequently, supervised machine learning models are trained and evaluated using these datasets.

C. Co-occurrence graph-based representation method

The main idea of this representation is to generate a graph representation where entities are only connected and have the same class if they co-occur with similar words. The following describe the main steps of this algorithm:

1. *Step 1*: Category Words Extraction: A set of representative words is extracted for each biomedical named entity class c from the training corpus based on the correlation between the word and its class c. The point-wise mutual information (PMI) was used by the following equation:

$$\mathrm{PMI}(\mathrm{class}, \mathrm{word}) = \log_2 \frac{\mathrm{p}(\mathrm{class}, \mathrm{word})}{\mathrm{p}(\mathrm{class}).\mathrm{p}(\mathrm{word})} \quad (1)$$

The top z words for each class are selected as category words for each biomedical entity class.

2. *Step 2*: Co-occurrence Matrix Construction: To measure the association between the word w_i and word w_j, this work uses the co-occurrence relation value by the following equation:

$$\mathrm{crv}(w_i, w_j) = \frac{p(w_i,w_j)[p(w_i)+p(w_j)]}{p(w_i).p(w_j)} \quad (2)$$

If the value returned by crv is greater than the threshold (0.20), then crv is added to the co-occurrence vector for the word w_i , otherwise, 0 will be added. After calculating the co-occurrence relation value (crv) between word w_i and all other words, the co-occurrence vector for the word w_i is added to the co-occurrence matrix. Figure 1 shows co-occurrence vector using the co-occurrence relation value (crv) between word w_i and all other words:

Representation						
words	w_1	w_2	w_3	w_4	…	w_n
w_1	$crv_{1\times1}$	$crv_{1\times2}$	$crv_{1\times3}$	$crv_{1\times4}$	…	$crv_{1\times n}$
w_2	$crv_{2\times1}$	$crv_{2\times2}$	$crv_{2\times3}$	$crv_{2\times4}$	…	$crv_{2\times n}$
:	:	:	...	:	…	:
w_m	$crv_{m\times1}$	$crv_{m\times2}$	$crv_{m\times3}$	$crv_{m\times4}$	…	$crv_{m\times n}$

Fig. 1. Co-occurrence Vector Construction

3. *Step 3*: Co-occurrence Graph Construction: in this step, **k**-nearest neighbours method is used, and a K-NN graph representation for each word in the dataset is constructed as follows:
   a) Node assignment: for each word in the dataset, a vertex is assigned.
   b) K_{cu}NN vertex calculation: to construct the graphs, the k nearest neighbors method is used. $k_{cu}NN(w_i)$ is a set of k nearest neighbours of the word w_i from class cu. A word w_j from category words of class C_u is assigned as one of the k_{cu} nearest neighbours set of word w_i if cosine similarity between their co-occurrence vectors is greater than ε. Cosine similarity was used by the following equation:

$$\mathrm{co\_sim}_{ij} = \cos(WV_i, WV_j) \frac{\sum_{k=1}^{n} wv_i^k . wv_j^k}{\sqrt{\sum_{k=1}^{n}(wv_i^k)^2} \cdot \sqrt{\sum_{k=1}^{n}(wv_j^k)^2}} \quad (3)$$

Where WV_i and WV_j are the co-occurrence vectors of a normal word w_i and class word w_j , respectively.

   c) Representation Construction: a representation vector rv for each word w_i ($rv(w_i)$) of length $|z \times C|$ is constructed for each word from both training and testing data. $|z|$ is the length of each category words set and C is the number of classes. Figure 2 shows representation vector using co-occurrence graph-based word representation:

Representation										
Words	**Category Words Of Class 1**				**Category Words Of Class 2**				…	**Category Words Of Class C**
	w_1	w_2	…	w_z	w_1	w_2	…	w_z	…	w_{zc}
w_1	$rs_{1\times1}$	$rs_{1\times2}$	…	$rs_{1\times z}$	$rs_{1\times1}$	$rs_{1\times2}$	…	$rs_{1\times z}$	…	$rs_{1\times zc}$
w_2	$rs_{2\times1}$	$rs_{2\times2}$	…	$rs_{2\times z}$	$rs_{2\times1}$	$rs_{2\times2}$	…	$rs_{2\times z}$	…	$rs_{2\times zc}$
:	:	:	...	:	:	:	...	:	…	:
w_m	$rs_{m\times1}$	$rs_{m\times2}$	…	$rs_{m\times z}$	$rs_{m\times1}$	$rs_{m\times2}$	…	$rs_{m\times z}$	…	$rs_{m\times zc}$

Fig. 2. Co-occurrence Graph-Based Word Representation by cosine similarity

D. Feature Level Graph-Based Representation Method

The following describe the main steps of this algorithm:

1. *Step 1*: Category Words extraction: as in step 1 of co-occurrence graph-based representation method.
2. *Step 2:* Feature Extraction: this work represents each word using a set of morphological features.
3. *Step 3:* Feature Graph Construction: in this step **k**-nearest neighbours method is used, a K-NN graph representation for each word in the dataset is constructed as follows:
   a) Node assignment: for each word in the dataset, a vertex is assigned.
   b) K_{cu}NN vertex calculation: to construct the graphs, the k nearest neighbors' method is used. $k_{cu}NN(w_i)$ is a set of k nearest neighbours of the word w_i from class cu. A word w_j from category words of class C_u is assigned as one of the k_{cu} nearest neighbours set of word w_i if the Jaccard index similarity between their feature vectors is greater than ε. Jaccard index similarity was used by the following equation:

$$F\_sim_{ij} = \text{Jacard}\left(fv_i, fv_j\right) = \frac{|fv_i \cap fv_j|}{|fv_i \cup fv_j|} \quad (4)$$

   Where $\boldsymbol{fv_i}$ and $\boldsymbol{fv_j}$ are the feature vectors of a normal word w_i and category word w_j, respectively.
   c) Representation Construction: a representation vector **rv** for each word $\mathbf{w_i}$ ($\mathbf{rv(w_i)}$) of length $|\mathbf{z} \times \mathbf{C}|$ is constructed for each word from both training and testing data. $|\mathbf{z}|$ is the length of each category words set and C is the number of classes.

Figure 3 shows representation vector using Feature Level graph-based word representation:

Representation										
Words	**Category Words of Class 1**				**Category Words of Class 2**				**…**	**Category Words of Class C**
	w_1	w_2	…	w_z	w_1	w_2	…	w_z	…	w_{zc}
w_1	$rs_{1\times1}$	$rs_{1\times2}$	…	$rs_{1\times z}$	$rs_{1\times1}$	$rs_{1\times2}$	…	$rs_{1\times z}$	…	$rs_{1\times zc}$
w_2	$rs_{2\times1}$	$rs_{2\times2}$	…	$rs_{2\times z}$	$rs_{2\times1}$	$rs_{2\times2}$	…	$rs_{2\times z}$	…	$rs_{2\times zc}$
:	:	:	…	:	:	:	…	:	…	:
w_m	$rs_{m\times1}$	$rs_{m\times2}$	…	$rs_{m\times z}$	$rs_{m\times1}$	$rs_{m\times2}$	…	$rs_{m\times z}$	…	$rs_{m\times zc}$

Fig. 3. Feature Level Graph-Based Representation by Jaccard index similarity

E. Graph-Based Word Representation Integration

Several proposed strategies have been considered to integrate two graph-based data representations. These strategies are:

1. *Single Representation:* each of the co-occurrence graph-based word representation and feature-based graph-based representation are used as a standalone representation. This allows us to better understand the limits of considering a single representation at a time.
2. *Concatenation*: The concatenated representation is defined as the vector concatenation of both representations in one representation. This representation shows the advantages of multi-information. Figure 4 shows the strategy of graph concatenation representation.

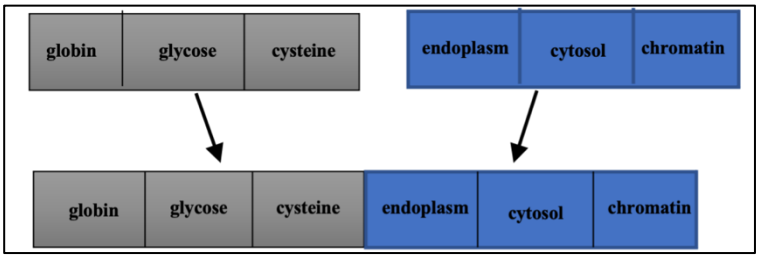

Fig. 4. The strategy of graph concatenation representation.

3. *Combination*: The representation is defined as a principled aggregation of both representations. In this scenario, equation (3) and equation (4) are integrated and combined as in the following equation:

$$\text{CF\_sim}_{ij} = \lambda_1 \cos\left(\text{WV}_i, \text{WV}_j\right) + \lambda_2 \text{ Jacard } \left(\text{WV}_i, \text{WV}_j\right) \quad (5)$$

F. Machine Learning Models

This section briefly overviews the supervised machine-learning techniques used: Conditional Random Fields (CRF) and Support Vector Machines (SVM), both known for their reliability and high performance in biomedical named entity recognition. The following describes these classifiers:

1. *Support vectors Machine (SVM):* Support Vector Machine (SVM) is a supervised machine learning model that was initially conceived for binary classification tasks. However, their utility has been extended to accommodate multi-class classification as well as regression challenges. SVMs have

garnered a solid reputation as an efficient classifier, standing out as one of the top choices across diverse data mining and machine learning applications. SVM achieves this by constructing a decision hyperplane that effectively partitions the training data into two primary classes. Given training data consists of n k-dimensional real vectors X, and labels Y, where y_i is either -1 or 1. The label y_i is +1 or -1 for the vector x_i. The training phase of the SVM aims to plot data vectors in a k-dimensional hyperspace and to draw a hyperplane, which as possible separates points from both classes.

$$\vec{\alpha} = \text{argmin}\left\{-\sum_{i=1}^{n}\alpha_i + \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \langle \vec{x_i}, \vec{x_j}\rangle\right\} \quad (6)$$

$$\text{Subjectto: } \sum_{i=1}^{n}\alpha_i y_i = 0;\ 0 \le \alpha_i \le C \quad (7)$$

2. *Conditional Random Fields (CRFs):* Conditional Random Fields (CRF) is a statistical sequence-labeling framework initially introduced by Lafferty, McCallum, and Pereira in 2001. CRF is a statistical model structured as an undirected graphical model, commonly employed for tasks such as pattern recognition and sequence labeling, including applications in Biomedical Named Entity Recognition. Formally, given a sentence $X = (x_1, x_2, \dots, x_n)$ to represent an input sentence, where x_i is the input vector of the i-th word, and $Y = (y_1, y_2, \dots, y_n)$ represents a sequence of predicted biomedical labels for input sentenceX. All classes or labels $y_i$of Y are restricted sets of labels over a set L(X). The global feature of CRF, F(y, x), is the summation of CRF's local feature vector f(y, x, i) for input sequence x and label sequence y, where i ranges over input positions. The probabilistic model for the CRF calculates the conditional probability of all possible sequences of labels y, given x using the following equation:

$$p(y|x) = \frac{1}{z(X)}\prod_{t=1}^{T}\exp\left\{\sum_{k=1}^{k}\omega_k f_k(y_t, y_{t-1}, x_t)\right\} \quad (8)$$

$$z(x) = \sum_y \exp\left(\sum_k \omega_k f_k(y, x)\right) \quad (9)$$

Where $Z(x)$is a normalization factor. where $\omega_k$is a parameter to be learned and estimated during training and shows the informativeness of the particular feature.

G. Experimental Setting

The primary objective of this study is to assess the effectiveness of the improved models proposed for biomedical named entity recognition. To achieve this, we employed the widely accepted and frequently used GENIA corpus as our dataset, which is a standard benchmark dataset for biomedical named entity recognition. The GENIA corpus is a comprehensive compilation of biomedical literature created as part of the GENIA project. While the original GENIA corpus contains a diverse array of 36 entity classes, a more commonly adopted version for tasks like BioNLP/NLPBA groups these entities into five major categories: protein, DNA, RNA, cell line, and cell type. To gain a more in-depth understanding of the models' performance, this work uses the confusion matrix to calculate recall and precision metrics. The confusion matrix as in Table 2 helps evaluate the classifier's effectiveness for each of these specific classes.

Table 2. Confusion Matrix

	actual (yes)	**actual (no)**
predicated (yes)	True Positive (TP)	False Positive (FP)
predicated (no)	False Negative (FN)	True Negative (TN)

In this work, three evaluation metrics are used as performance metrics namely, precision, recall, and F-measure.

1. *Precision*: Precision is the proportion of true positive predictions out of all positive predictions made by the model. It measures the model's ability to correctly identify named entities without making many false positive errors.

$$P_i = \frac{TP_i}{TP_i + FP_i} = \frac{Relevant\ Entites\ Recognized}{Total\ Entites\ Recognized} \quad (10)$$

2. *Recall*: Recall is the proportion of true positive predictions out of all actual positive instances in the dataset. It assesses the model's ability to identify all relevant named entities without missing many.

$$R_i = \frac{TP_i}{TP_i + FN_i} = \frac{\text{Relevant Entites Recognized}}{\text{Relevant Entities In Corpus}} \quad (11)$$

3. *F1-Score:* The F1-Score is the harmonic mean of precision and recall, providing a balanced measure of a model's overall performance

$$F1_i = \frac{2(P_i * R_i)}{P_i + R_i} \quad (12)$$

A series of experiments were conducted to assess the performance of both baseline feature representation techniques and enhanced graph-based word representation methods for biomedical named entity recognition. Initially, the focus was on evaluating the baseline recognition models, which were constructed using traditional feature extraction and representation techniques. As well as two supervised models, Conditional Random Fields (CRF) and Support Vector Machines (SVM) were used.

In these baseline experiments, words were represented as vectors, with values assigned to various morphological, orthographical, and contextual features, as detailed in Table 3. The primary objective was to investigate how supervised machine learning models could leverage these features and understand their impact on the model's performance.

The results presented in Table 3 display the outcomes achieved when using Conditional Random Fields (CRF) and Support Vector Machines (SVM) in conjunction with the morphological, orthographical, and contextual features. It is worth noting that the results clearly demonstrate that Conditional Random Fields (CRF) outperform Support Vector Machines (SVM) in the domain of biomedical named entity recognition. This is because CRF excels in handling sequential data, which is essential in the biomedical field,

where term order and context are critical for accurate recognition.

Table 3. Performance of Conditional random fields (CRF) and support vector machines (SVM) models on GENIA Dataset with Traditional Feature Representation

Training size	Precision	Recall	F-Measure
SVM	84.2	40.4	54.6
CRF	83.5	80.3	81.87

In the second part of our study, we conducted multiple experiments to assess the effectiveness of ennhanced graph-based representation methods in combination with conditional random fields (CRF) and support vector machines (SVM) for biomedical named entity recognition. We developed several enhanced learning models for this purpose. We evaluated four graph-based methods, which can be categorized into two types: standalone and integrated representations. The standalone methods included co-occurrence graph-based representation (CGBR) and feature graph-based representation (FGBR), while the integrated methods were represented by concatenated graph-based representation (ICCGBR) and combined graph-based representation (ICMGBR).

Figure 5 illustrates the results achieved using these proposed graph-based representations in conjunction with the two supervised models. The findings demonstrate that these novel representations significantly enhance the performance of the conditional random fields (CRF) and support vector machines (SVM) models compared to traditional feature representations.

When considering the word representations, all four proposed graph-based methods substantially improve the performance of the two supervised machine-learning methods compared to the baseline feature representation methods.

From the perspective of standalone graph-based representations, Figure 5 reveals that the results obtained with the co-occurrence graph-based representation surpass those achieved with the feature graph-based representation.

Regarding integrated graph-based representations, Figure 5 demonstrates that the results of the two supervised models with the concatenated graph-based representation significantly outperform their counterparts with the combined graph-based representation.

When we consider both standalone and integrated graph-based representations, Figure 5 emphasizes that the results obtained with the concatenated graph-based representation stand out as the most effective among all representation methods.

In terms of supervised machine learning methods, Table 3 and Figure 5 consistently show that conditional random fields (CRF) consistently outperform support vector machines (SVM), regardless of the word representation method used.

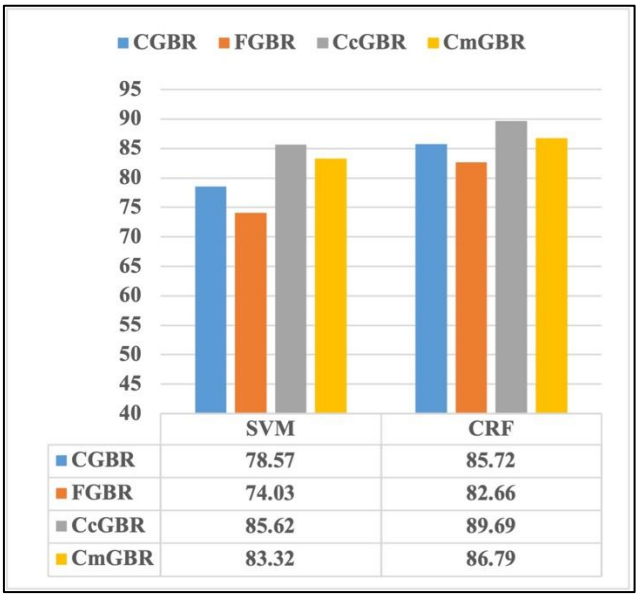

Fig. 5. Performance of the proposed (four) graph-based representation with conditional random fields (CRF) and support vector machines (SVM) on GENIA Dataset

H. Baseline Feature Extraction and Representation Methods for Biomedical Named Entity Recognition

The underperformance of traditional machine learning models namly (CRF, SVM) in biomedical text analysis is primarily attributed to the inadequate traditional feature representation techniques of biomedical literature through manually created features based on existing literature and encompass morphological, orthographic, and contextual characteristics. Subsequently, these techniques aren't capable of capturing valuable information for recognition and classification while remaining dependent of specific domains and datasets.

I. Enhanced Graph-Based Word Representation Methods for Biomedical Named Entity Recognition

Enhanced graph based word representation methods improving the performance of machine learning methods namely (CRF, SVM) in biomedical named entity recognition through:

- The correlation (using PIM) between the word and its class as feature set of class c (eg. Category Words of Class c).
- The co-occurrence relation value (crv) to measure the association between the wi and word wj as word embedding (eg. co-occurrence vectors all of words).
- K-NN graph representation is used for similarity between co-occurrence vectors all of words with co-occurrence vectors of category words of class c (using cosine similarity and Jaccard index similarity) as more feature set of class c (eg. More Category Words of class c).

Subsequently, these methods are capable of capturing valuable information for recognition and classification while remaining independent of specific domains and datasets.

IV. CONCLUSIONS

The quality of traditional machine learning methods highly depends on data and words representation. However, a critical issue of such techniques is the choice of the data and word representation. many machine learning for NER that are trained using data represented using based feature-engineering that have been selected using domain experts. These approaches are the need of manually engineering features for each specific dataset [4, 17, 18]. This means they are dataset and domain dependent.

This research study introduces new enhanced graph-based word representation methods for biomedical named entity recognition, specifically co-occurrence graph-based representation and feature graph-based representation methods that are not datasets dependent. The results show that the proposed graph-based representation methods significantly enhance the performance of both conditional random fields (CRF), and support vector machines (SVM) compared to the baseline representation.

Furthermore, the findings show that incorporating concatenated graph-based representations significantly improves the overall quality of biomedical named entity recognition across all machine learning models. This highlights graph-based representations as a more effective approach than traditional methods for this task.

V. FUTURE WORKS

Future research study is required for better development and contribution in this research area. The following are some suggestions for future work:

- The proposed graph-based word representation method should be tested on other datasets. To verify the effectiveness of the proposed model, experiments should be conducted on several datasets that belong to the biomedical domain.
- Future work should build other ensemble graph-based word representation methods.
- Future work should evaluate with proposed graph-based word representation method with advanced deep-learning models.
- Future works are encouraged to engage other integration algorithms to combine graph-based and advanced representation methods and evaluate them with recognition methods.

REFERENCES

[1] Á. A. Casero, *Named entity recognition and normalization in biomedical literature: a practical case in SARS-CoV-2 literature*, ETSI Informatica, 2021.

[2] Z. Chai *et al.*, "Hierarchical shared transfer learning for biomedical named entity recognition," *BMC Bioinformatics*, vol. 23, no. 1, pp. 1–14, 2022.

[3] J. Lee *et al.*, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2019.

[4] R. Ramachandran and K. Arutchelvan, "ArRaNER: A novel named entity recognition model for biomedical literature documents," *J. Supercomput.*, pp. 1–14, 2022.

[5] R. M. Rivera-Zavala and P. Martínez, "Analyzing transfer learning impact in biomedical cross-lingual named entity recognition and normalization," *BMC Bioinformatics*, vol. 22, no. 1, pp. 1–23, 2021.

[6] B. Song *et al.*, "Deep learning methods for biomedical named entity recognition: a survey and qualitative comparison," *Brief. Bioinform.*, vol. 22, no. 6, Article ID bbab282, 2021.

[7] P. Malik *et al.*, "NLP techniques, tools, and algorithms for data science," in *Artificial Intelligence for Signal Processing and Wireless Communication*, vol. 11, pp. 123, 2022.

[8] I. Lauriola *et al.*, "Learning adaptive representations for entity recognition in the biomedical domain," *J. Biomed. Semant.*, vol. 12, no. 1, pp. 1–13, 2021.

[9] U. Kanimozhi and D. Manjula, "A systematic review on biomedical named entity recognition," in *Proc. Int. Conf. Data Sci. Analytics and Applications (DaSAA)*, Chennai, India, Jan. 2017, pp. 1–12. Revised Selected Papers. Springer, 2018.

[10] A. Alves-Pinto *et al.*, "Iterative named entity recognition with conditional random fields," *Appl. Sci.*, vol. 12, no. 1, Art. no. 330, 2021.

[11] L. A. Mady, Y. M. Afify, and N. L. Badr, "Biomedical named entity recognition using structured support vector machine," *Int. J. Comput. Appl.*, vol. 28, no. 4, 2021.

[12] R. Ramachandran and K. Arutchelvan, "Optimized version of tree-based support vector machine for named entity recognition in medical literature," in *Proc. 3rd Int. Conf. Intelligent Sustainable Systems (ICISS)*, 2020, pp. 1124–1129.

[13] Y. Li, L. Song, and C. Zhang, "Sparse Conditional Hidden Markov Model for Weakly Supervised Named Entity Recognition," *arXiv preprint*, arXiv:2205.14228, 2022.

[14] M. T. Abd and M. Mohd, "A comparative study of word representation methods with conditional random fields and maximum entropy Markov for bio-named entity recognition," *Malays. J. Comput. Sci.*, pp. 15–30, 2018.

[15] H. Chen, S. Yuan, and X. Zhang, "ROSE-NER: Robust Semi-supervised Named Entity Recognition on Insufficient Labeled Data," in *Proc. 10th Int. Joint Conf. Knowledge Graphs*, 2021.

[16] L. Liu *et al.*, "A semi-supervised approach for extracting TCM clinical terms based on feature words," *BMC Med. Inform. Decis. Mak.*, vol. 20, no. 3, pp. 1–7, 2020.

[17] M. Song, H. Yu, and W.-S. Han, "Developing a hybrid dictionary-based bio-entity recognition technique," *BMC Med. Inform. Decis. Mak.*, vol. 15, no. S1, p. S9, 2015.

[18] H.-C. Kuo and K.-I. Lin, "Extracting protein names from biological literature," *Adv. Comput. Sci.: Int. J.*, vol. 3, no. 2, pp. 58–68, 2014.

[19] M. T. Pilehvar and J. Camacho-Collados, "WiC: the word-in-context dataset for evaluating context-sensitive meaning representations," *arXiv preprint*, arXiv:1808.09121, 2018.