Journal of Computers Vol. 31 No. 4, 2020, pp. 157-173
doi:10.3966/199115992020083104012

Research-based-named Entity Recognition Learning Text Biomedical Extraction by Adoption of Training Bidirectional Language Model (BiLM)

Alshreef Abed*, Yuan Jingling, Lin Li

Department of Computer and Science, Technology, Wuhan University of Technology, Hubei 430000, Wuhan, China
alshreefabed8@outlook.com, yjl@whut.edu.cn, cathylilin@whut.edu.cn

Received 30 January 2019; Revised 9 May 2019; Accepted 31 May 2019

Abstract. Deep learning entities are a fundamental task biomedical and text extraction; it represents an amazing research scope where End-to-End configuration can be adopted without any requirement for hand-engineered function. However, most of the methods currently used only focus on High-Quality labeled configuration, which is an expensive option for users. To satisfy this inconvenient, combining ext mining for discovering andbiomedical knowledge extraction can be seen as the best appropriate scope in the context of computer sciences applications. In this article, we combine Biomedical named entity recognition (NER) with learning entity to increase the labeled data extracted. By adopting bidirectional language model (BiLM) in NER environment, two stages are defined; first, we evaluate BiLM- NER F1 score in the context of training on unlabeled data and transfer. The evaluation is set up on four benchmarkdatasets in the purpose to show leads and F1 scores. The NER-BiLM results obtained show a high performance of F1 score. However, a high-level challenge concerning the time factors cost still required. To fix this issue, in the second step of our interpretation, a comparative study between BiLM- NER with CCA and Canonical Correlation two current performances are investigated; the results show that, compared with a baseline having a 70.09% F1 score, BiLM-NER F1 score show 72.82%, which represent a gap of 0.3% compared with CCA and Canonical Correlation. This new performances confirm the highest-level of our proposed-NER-BiLM approaches. This work can be considered as a new contribution to data mining and biomedical research approaches.

Keywords: big data processing, biomedical text data, computer, CNN, data mining, engineering sciences

1 Introduction

Today, there is a high demand for accessible extraction works utilizing biomedical and content information writing contains. The current number of written articles available online are over a billion (abstracts and full articles) [1-2]. Projects in unsupervised information process handling strategies have made biomedical extraction works conceivably to be more adapted with the use of modern annotations such as statistical language approaches. These new projects require the high implementation of vector strategies in the context of the space representations, which will include query extension, text classification, and the use of named entity recognition. Biomedical articles consolidate a noteworthy number of figures with graphs which ordinarily show test results, describing investigate models used, and giving serial cases of biomedical objects (the cases of colors and tissue). The data extracted from that information processing play a vital role in biomedical information, and today it is given much consideration inside the sphere of biomedical research community [3-4].

The information extracted in biomedical and mining systems is used in different sectors such as the

* Corresponding Author

medical field, security or commercial entity. Therefore, in that process, “abstract” is usually considered as one of the type models that give the structure and contain the whole powerful meaning of a document. Sometimes, other information such as author names, topic title or scope area of publication can be also associated in the process [5]. It is shown in several types of research [6] that expressed text mining has gotten to be one of the trendy fields that have been joined in data mining; this category includes also the computational linguistics and information retrieval (IR) area. Text mining field is very complex from data mining [7]. And in this context, data mining area works on finding out targets from colossal database lexical or literature [8]. Data retrieval strategies such as text indexing methods have been released and figure out to meet new unstructured information requirements. In routine inquiries about all current implementation process, it is assumed that any user who is searching information on-line focus especially in some terms relative to his/her research target, which have been already utilized or released by other researchers. However, sometimes the most issue encountered is that the target results are not pertinent or not totally satisfy the user’s demand. In this situation, an arrangement can consist to use data mining to find all relevant data that are not totally extracted. A specific document can be exploited through text mining process and by exploring its characteristics and sets; it’ll be pre-processed by other characteristics enrolled. Data contained in that document will be analyzed through different steps. Diverse text analysis that includes semantic methods can, for this reason, be used to perform highly the level information obtained. In a few cases, the methods used to extract data can reuse several times until data is extracted. The results can be put away in a monitoring data framework that gives a good performance on the quality of data for the user of the system.

NER has been widely a considered research priority in the scope of learning processing language, and a number of works have applied machine learning approaches inside NER from diverse biomedical domains [9-10]. Due to high linguistic variation in data, building NER systems with high precision and high recall for the medical domain constitute them a quite challenging task.

First, a dictionary-based approach doing pattern matching will fail to correctly tag ambiguous abbreviations that can belong to different entity types. For example, the term CAT can refer to several phrases such as “*Computerized Axial Tomography*” (Stevenson et al., 2010).

Second, as the vocabulary of biomedical entities such as proteins is evolving, it makes the task of entity identification even more challenging and error-prone more difficult to create labeled training; for examples, we can cite the case concerning having a wide coverage. Also, in contrast to natural text, entities in the medical domain can have short or long names that can lead a NER tagger to incorrectly predict the tags.

Lastly, the most recent stage in the development of machine learning methods for NER task relies on high-quality labeled data, which is expensive to procure and is therefore available only in limited quantity. Therefore, there is a need in the implementation of new approaches to using unlabeled data to improve the performance of NER systems.

NER can be devised as a supervised machine learning task in which the training data consists of labels or tags, with each token in the text [11]. A typical approach for NER task consists to extract word-level features by following the training results of the linear model for tag classification.

According to all listed challenge detailed above, to extract all features and perform the biomedical environment, our work introduces a called “*transfer learning method*”, to evaluate unlabeled information by using pre-training-weights in named entity recognition. The proposed approach “*pre-train*” uses words embeddings, high learned character, and BiLSTM [12]. The characters or expression implanted are analyzed from a high series of PubMed lexical abstracts and improved on NER *datasets-F1 score*. Finally, to confirm the obtained performances, a random comparison around a group of the word-vectors is initialized. BiLSTM, when applied in combination with character features gives a map with similar terms like “*lymphoblastic leukemia*,” “*null-cell leukemia*,” which vary in term of latent space that captures the semantic meaning in the phrases. This powerful representation of terms in a latent semantic space can also help in the correct classification of unused entities. All these entities with similar contexts are mapped closer together.

Following this introduction, the remaining parts of this paper are designed as below: From Section 2, we explain the current challenges and main innovation aspects developed during the last five years. Section 3 introduces our proposed model approach, and provides concise details concerning the main contribution of our paper. Section 4 lists the main experiment work obtained. Finally, Section 5 concludes the paper.

2 Challenges

2.1 The Meaning of Named Entity Recognition (NER)

NER is a very essential function in biomedical document processing, and it includes varieties of applications. NER can be defined as an expression, character or group of words used to identify a target expression from a series of related items [13]. Inbiomedical domain, basically, some specific words such as people, tissue, disease color and others constitute its main lexical. NER discovers and classifies target expressions into groups.

The first research on NER task was organized by Grishman et al. (1996) in the Sixth Message Understanding Conference. Since then, there have been numerous NER tasks research introduced: (1)Tjong Kim et al (2003), (2) Tjong et al (2002), (3) Piskorski et al. 2017, (4) Segura Bedmar et al. (5) 2013, Bossy et al., 2013, (6) Uzuner et al., 2011), (7); all these different methods-based NER describes concepts and application and efficiency in NER environment. There are also approaches implied in events [14], in ACE [15], IREX, and TREC lexical discovery research areas. The Fig. 1 illustrates the basic model of NER architecture where conditional random fields (CRF) are deployed. And from Fig. 2, the Entity extraction using Deep Learning is drawn.

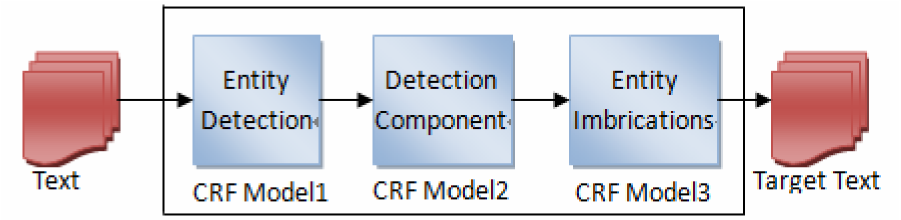


Fig. 1. Architecture model of the NER system

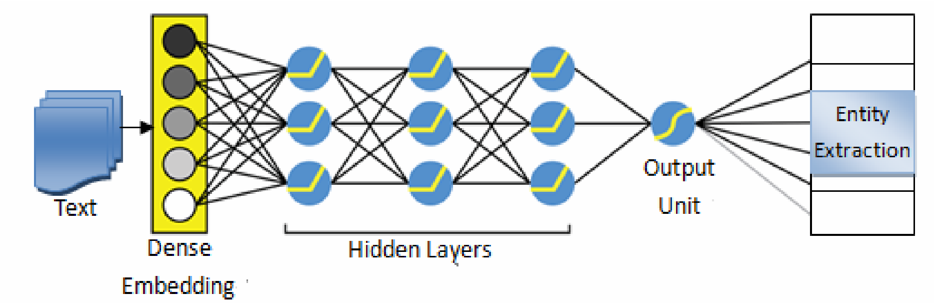


Fig. 2. Model based deep learning text

Formally, giving a group of tokens $\beta = \delta_1 + \delta_2, ...\delta_\alpha$ named entity can allow to create a list of subsets (A_m, A_n, A_o), every sub-set will represent NER in β. Here, $A_m \in [1, N]$ and $A_n \in [1, N]$ are the beginning and the achievement list factors of named substance mentioned; t represents the class chosen from a specific category. Fig. 1 gives an illustration of the NER in a recognition process established on 3 specific items [16].

2.2 Challenges in Text Mining and Data Mining

By observing the challenges in the context of mining data application, new processes on extraction represent high demand claimed by many users to provide feedbacks concerning targeted researches. This category of research occupies a major position and many scholars already introduced variety of review on this concern [17]. In this challenges, NER and association extraction are the main used mining tasks. NER will focus on fixing specific references, and the association extraction entity will provide listings (lexical of words and expressions). All these information will be evaluated and classified into different content according users demands. This scenario represents the fundamental construction process of extraction data in biomedical concepts. New methods can be also inserted to solve the problem of matching or statistical with the adoption of strategic dictionary that content most of learning based systems lexical, and with this new adoption, no limitations can be possible to extend or provide new entities and classes. However, new efforts need to be deployed to satisfy and improve some lakes observed in training engine systems. These new efforts are the core of main problem of efficiency issue that requires achievement, especially in scientists' annotation functions.

We can then, formulate that labeling particular targets in NER is a very crucial step towards the genuine objective to extract exact biomedical information. All content data are extracted obeying to some specific rules to make able all connections from biomedical datasets. All system designs and programs also obey to these requirements. Current systems are also based on rule file-based, and present high demand on processed information that are transmitted from one user framework system to another one.

1. High dimensional and high-speed data streams;
2. Data and time series in Mining sequence;
3. Network setting in data mining field;
4. Complex knowledge analysis id data mining;
5. Multi-agent and distributed data mining;
6. Related issues in Data Mining process;
7. Data integrity and Security;
8. Non-static and cost-sensitive data operations;

All these challenges listed above represent today the major large part of works reports developed in the field of biomedical domain data mining scope. In the context of this paper, the challenge in Data mining and the network setting context represent one of the most critical areas since it concerns the implementing platform where all data go to be designed; Scaling, sequence data and time, the Distributed, Security, privacy and data integrity which include Dealing, all those challenge improvements depend on the capacity of integral platform.

By suggesting a new platform with BiLM capable to improve based on NER and combining pretraining sequences, we show that such pretraining of weights can highly help to strengthen new performancesby substantially increase the A1 score on 4 dataset classes for biomedical named entity recognition by comparison with the scientific field achievements.

Let establish in Table 1 the level of NER development for dataset platform configuration. As we can see from Table 1, after 2014, there is large volume of datasets developed and implemented with text sources, which data are found in documents discovered in Wikipedia and YouTube; the nature of that content is mainly text.

Table 1. Evolution of NER dataset platform from last 10 Years

Year	Corpus	Platform	Target	Web Address
2005	BBN	Wall Street Journal system	6	https://catalog.ldc.upenn.edu/ldc2005t33
2007	NYT	The N.Y Times documents	5	https://catalog.ldc.upenn.edu/LDC2007T19
2009	WikiGold	Wikipedia text	5	https://figshare.com/articles/Learning_multilingual_named
2012	WiNER	Wikipedia text	5	http://rali.iro.umontreal.ca/rali/en/winer-wikipedia-for-ner
2012	WikiFiger	Wikipediatext	110	https://github.com/xiaoling/figer
2014	N^3	Text and videos	4	http://aksw.org/Projects/N3NERNEDNIF.html
2014	GENIA	Biology and clinical reports	37	http://www.geniaproject.org/home
2014	NCBI-Disease	PubMed text	788	https://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/DISEASE/
2015	GENETAG	MEDLINE text	4	https://sourceforge.net/projects/bioc/files/
2015	BC5CDR	PubMed content	5	http://bioc.sourceforge.net/
2018	DFKI	Text and video	5	https://dfki-lt-re-group.bitbucket.io/product-corpus/

2.3 Printing Area

If we take a look by analysis this concept mathematically, precision on the evaluation factor is less certain. Therefore, the condition "*true*" must be verified on both boundaries and type match as detailed in [18-19]. Several elements play essential role to determine this condition: the false negatives (FN), the false positives (FP), and the true positives (TP).

These factors play also the role of benchmark and are very essential in the definition of match evaluation:

(1) TP represents the entities that are discovered by NER to match ground truth;

(2) FP plays the role of the entities that are discovered by NER, but in this case, there is no any obligation to match with ground truth;

(3) FN is the entities conferred to the ground truth that is not recognized by NER.

In this process, we choose to use the Precision to evaluate the capacity of NER platform in the purpose to prepare reliable entities, and make a review on all NER measures and capacities. The calculation of precision, recall and score can be obtained as below:

$$Precision\ value = \frac{TP}{TP + FP} \tag{1}$$

The recall value is a dependent factor to the precision:

$$Recall\ value = \frac{TP}{TP + FN} \tag{2}$$

With knowing these two factors, *F-score* is then evaluated as the concept of "*precision*" and "*recall*" are obtained proportionally to "*F-score*" value:

$$F - Score = 2x\frac{Precision\ x\ Recall}{Precision\ x\ Recall} \tag{3}$$

The named entity recognition can enroll several categories of datasets, and in this kind of scenario, to avoid lost and mixing, we can frequently required to evaluate the execution over all substance classes, and Macro-averaged *F-score* and micro-averaged *F-score* represent here the two main values that require special attention during all implementation process:

Macro-averaged *F-score* will compute the *F-score* without separately from each entity, at this stage, consequently starts the process of equality in all categories of dataset discovered. The factor "*F-score*" will interfere to establish aggregation and contributes evaluates effectively all classes [20-21].

3 Proposed Method

3.1 Motivations & Contributions

3.1.1 Improving Listed Challenges

According to the challenges dressed above in chapter 2, the first contribution in this paper stands on the evaluation of the process in the current implementation of data mining system; we implement the convolutional neural network (CNN). The main mining system building blocks of the neural network implemented based NER model are character-level CNN layer, word embedding layer, word-level BiLSTM layer, decoder layer, and sentence-level label prediction layer [22-24].

3.1.2 Lack on Use of Neural Networks Methods Labeled Data

Our second contribution in this work concerns our improvement on the manual Task-specific Feature Engineering framework; currently, researchers have worked on carefully designing hand-engineered features to represent a word such as the use of parts-of-speech (POS) tags, capitalization information, use of rules such as regular expressions to identify numbers, use of gazetteers, etc. A combination of supervised classifiers using such features was used to achieve the best performance on *CoNLL*-2003 *benchmark* NER dataset (Florien et al., 2013). Lafferty et al. (2001) popularized the adoption of graphic models; the linear-chain conditional random fields (CRF) is one of them that model largely used in NER tasks.

However, it is observed in these previous contributions that the current methods involved in neural networks are only trained End-to-End and only use the available labeled data without the need of manual task-specific feature engineering. In their seminal work, Collobert et al. (2011) trained window-based and sentence-based models for several NLP tasks and demonstrated competitive performance. For NER task on newswire texts, Huang et al. (2015) uses word embeddings, spelling, and contextual features that are fed to a BiLSTM-CRF model. To incorporate character features, Lample et al. (2016) applies BiLSTM while Chiu et al. (2016); Ma et al. (2016) applies CNNs on character embeddings respectively. For biomedical NER task, Wei et al. (2016) combines the output of BiLSTM and traditional CRF-based model using an SVM classifier. Zeng et al. (2017) realized several experiments with high impact evaluation in words and expressions based-BiLSTM in the context of NER tsks. Habibi et al. (2017) investigates the effect of pretraining word embeddings on several biomedical NER datasets.

3.2 Description of based NER-BiLM and CNN

The main building blocks used to design our neural network system is implemented on NER model, and its main characteristics include: character-level CNN layer, word embedding layer, word-level BiLSTM layer, decoder layer, and sentence-level label prediction layer. During model training, the entire layers are jointly trained. Before set up training scenario, we first pre-train CNN parameters, word embedding, and BiLSTM layers in the NER model by using the learned parameters from a language model that has the same architecture. Specifically, we perform BiLM to pretraining the weights of both the forward and backward LSTMs in the NER model.

Now, nextstage consists to describe in details these different layers.

CNN presents high characteristic levels; this system is widely used in computer vision tasks for visual feature extraction (Krizhevsky et al., 2012). In NLP, where the data is highly sequential, successful applications of CNNs include tasks such as content classification (Kim, 2014) and arrangement labeling (Clobber et al., 2011).

In this paper, we use CNNs to extract features from characters (Kim et al., 2016) as they can encode morphological and lexical patterns observed in languages. For its evaluation, it is represented by an embedding vector. These character embeddings are stored in a lookup table defined as shown in equation below:

$$Z_D \in R^{S_D * E_D}, \tag{4}$$

Where S_D is the character vocabulary, E_D is the dimensionality of character embeddings. To compute character-level features, we perform one dimensional convolution along the temporal dimension, and the global mathematical representation can be design as:

$$K_Y[A] = f(Z_K * X[:, I + B - 1] + b_K), \tag{5}$$

Where* stands for the dot product operator, b_k is the bias, $X \in R^{w_l * E_D}$ is the character-basedembedding representation of a word, w_l stands for the length of a word, Z_k represent filter weights, *B*is the convolution stride, *f* can be any nonlinear function and play major role in modifying the linearunits *(f(x) = max(0, x)).*To capture the important features of a word, multiple filters with different strides are used. Finally, the maximum value is computed over the time dimensionalso called max-pooling to get a single feature for every filter weight.

Concerning the word-Level Bidirectional LSTM value, recurrent neural network (Werbos, 1988) such as LSTM (Hochreiter et al., 1997), it is widely used in NLP. This kind of level can model the long-range dependencies in languagestructure with their memory cells and explicit obtain mechanism. The dynamics of anLSTM cell is controlled by an *input vector (xt)*, which include the *forget gate (ft)* value, the *input gate (it)*value, the*output gate (og)*, the*cell state (ct)*, and a *hidden state (ht)*, which are computed as:

$$i_t = \delta(Z_i * [G_{t-1}, X_t] + b_i), \tag{6}$$

$$f_t = \delta(Z_i * [G_{t-1}, X_t] + b_f), \tag{7}$$

$$p_t = \delta(Z_p * [G_{t-1}, X_t] + b_p), \tag{8}$$

$$j_t = than(Z_j * [G_{t-1}, X_t] + b_j), \tag{9}$$

$$c_t = f_t * c_{t-1} + i_t * j_j, \tag{10}$$

Where c_{t-1} and G_{t-1} stand for the cell state and hidden state respectively from previous timestep, δ stands for the sigmoid function which is equal to $(\frac{1}{1-e^{-x}})$, than is the hyperbolic tangent function which can also be calculated as $(\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}})$.

3.3 Evaluation of Training BiLM and Transfer to "Pre-training"

For every word, hidden state representations from BiLSTM are concatenated by $[\vec{l}_t, \overleftarrow{l}_t)$ and are fed to the decoder layer. The decoder layer computes an affine transformation of the hidden states:

$$d_t = Z_d l_t + b, \tag{11}$$

Where *l* is the dimensionality of the BiLSTM hidden states, *t* is the total number of *tags*, $Z_d \in R^{t^{*}H}$, and *b* represents the learnable parameters. Decoder outputs are referred to as logics in the subsequent discussions. To compute the probability of a tag $(\hat{Y}_t)$ for a word, softmax function is used:

$$p(\hat{Y}_t = M \mid z_t) = soft\max(d_t), \tag{12}$$

We define $Y = \{y_a, y_b, ..., y_z\}$ denote the sequence of tags in the training corpus, then the cross-entropy loss is calculated as:

$$CE_{NER}(Y, \hat{Y}_t) = -\sum_{t=1}^{N}\sum_{j=0}^{t} L(y_t = \hat{y}_{t,j}) \log \hat{y}_{t,j}. \tag{13}$$

To study the design parameters and cross-entropy loss value, it is important to minimize backpropagation through time (BPTT).

For the language modeling configuration, we provide a short description for the main parameters that are used to initialize the NER model. In language modeling, the main objective consists to train a model that maximizes the most near value of a given sequence of words. At each step, a language model computes the probability of sequential expressions or words in the sequence discovered throughdatasets. If the sequence of words is w_1, w_2... w_n, its likelihood is given as:

(14)

Where *wn+1* represents a specific symbol used to achieve the sequence. LSTM is also in the same time adopted to predictthe probability of the next word. It can also give the current word position and the previous sequence position words (Graves, 2013). This process is done by applying an affine transformation to the hidden states of LSTM at every time step to obtain the logics for all words in the vocabulary. This scenario is used in our work toforward language to the model LM_f which represent one of four benchmarks we selected in our implementation process.

4 Experiment Results

In this Section, our article gives the biomedical text mining effect in the adoption of bidirectional language technique approach; the biomedical context focused concerns only the study in entity recognition pattern. To achieve this goal, in this paper:

First, we figure out the proposed BiLM-NER concept on a database with four biomedical text mining samples in the purpose to establish a comparative study with current challenging approach.

Second, we estimate NER approach, and its several samples on 3 different techniques.

Finally, the Bio medical text mining interpretation system is covered by NCBI with new interpretation based disease entity by the adoption of BiLSTM-CRF design.

4.1 Experiment Setup

4.1.1 Dataset Preparation and Evaluation

Our approach is evaluated on four datasets: NCBI-disease, with the adoption of Disease Relation to realize the extraction process of BC5CDR task, BC2GM is also adapted to calculate the score. For each category of dataset selected in our work, training phase, development phase, and test set are split according to the statistics obtained. An overall summary of these datasets such as the number of sentences, words, and entities is presented in Table 2. For each dataset, we use its training and development splits as unlabeled data for language modeling task.

Table 2. Datasets statistics of NCBI-disease, BC5CDR, BC2GM, NLPBA

Properties	JNLPBA	BC2GM	BC5CDR	NCBI-disease
Category	5 NEs	Gene/Protein	Disease, Chemical	Disease
# Entity mentions	51,438	24,583	4,500	3,890
# Sentences	24,926	19,000	9,909	5,290
# Words	568,791	839,824	360,315	354,552
# Train documents	1,810	—	200	200
# Dev documents	250	—	400	68
# Test documents	505	—	300	52

4.1.2 Set Up of the NER Model

Due to the fact that language strategy weight is set up to fix our NER model, the integral modelwill be identical after configurations, excepting for the top decoder layer. Dimensions of character discovered and wordsanalyzed are set up to *40* and*200* respectively. CNN filters have *widths (w)* in the range from *1* to *7*. The number of filters are computed as a functionof filter width as *min (200, 40*z)*.

4.2 Results and Discussion

In this section, we first confront our NER-BiLM approach efficiency in the general context by confronting his F1 performance with traditional based NER works.

In the second series of our simulation results, we confront our obtained NER-BiLM F1 score performances with Global vector technique, and Canonical Correlation Analysis (CCA).

4.2.1 NER-BiLM F1 Score Evaluation

Let evaluate the proposed BiLM in NER datasets. According to the current level of research, the main models dressed in literature review of Basel Alshikhdeeb et al. (2016) the impact of features eligible for named entities identification bring the more essential value in the biomedical concept. However, according to the current level of idea, the biomedical features require more experiments and BEER type variety to improve the quality of text mining based lexical features.

If we compare BiLM by focusing only on NER as presented by Hye Song et al. (2016), the paper approach observes an essential high-level challenge concerning the time factors cost required. This is due to the implementation of the words embedding constructed through the unsupervised texts mining learning.

This two work performances described above present good quality on dataset result if the main experiment concept lies on the adoption of multi-task learning such as to evaluate the neural network scenario with high effect confirmed by Grichton et al. (2017); there are good implementation results with a better average observed on NER dataset. In the same way, Wang et al. (2018) used a similar interpretation on multi-task learning framework for BioNER approach with different category of evaluation. Giorgi et al. (2018) in learning methods confirmed this performance on biomedical named entity recognition concept by focusing on neural networks; this new solution can highly outperform named learning into an especial consistent targeting score with a minimized volume of labels.

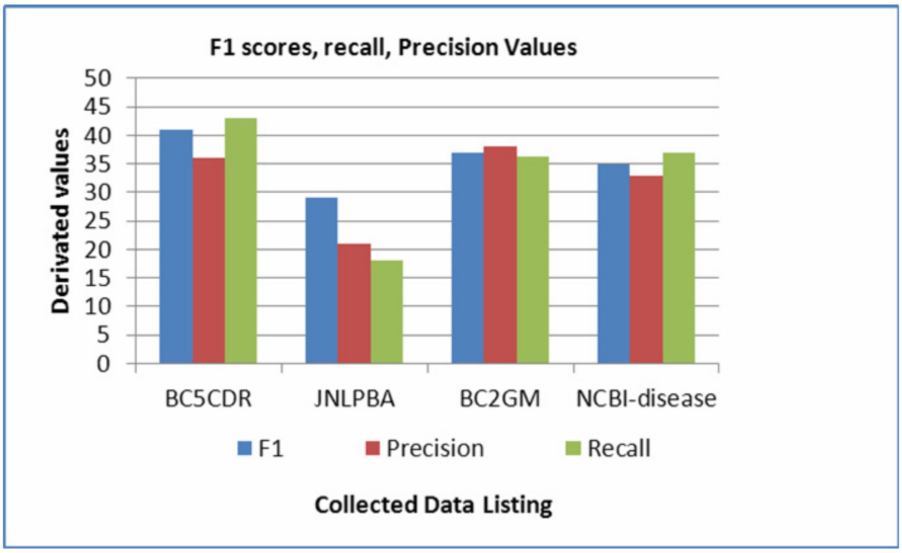


Fig. 3. Different Variations of Proposed BiLM

However, all these performance evaluations can be more efficient if we can integer BiLM-NER concept high profit substantially from biomedical transfer learning. And this new approach is actually welcome in German where BiLSTMs is mainly recommended on multi-task with NER recognition platform.

Now, let show in Table 3 the resume precision, which include the *F1scores* evaluation and the recall-precision. The main consequence of this comparative work presents the proposed BiLM with the Maximum *F1 score* in all variation data results. The behavior of BiLM shows a *Maximum F1 score* result. The NCBI-disease dataset stands at *87.35%* which represents a plus of *1.19%* performance comparing to current multi-task learning approaches.

Table 3. Values of a1scores, recall, precision of previous various of proposed model

Collected Data Listing	Metric	No Derivative	LM_f Derivation	LM_b Derivation	BiLM Derivation
BC5CDR	F1	88.80	88.98	88.77	89.29
	Precision	88.96	88.68	88.13	88.11
	Recall	88.65	89.29	89.42	90.61
JNLPBA	F1	73.79	73.67	73.88	75.03
	Precision	71.24	70.52	71.01	71.39
	Recall	76.53	77.12	76.99	79.06
BC2GM	F1	80.62	81.28	80.59	81.70
	Precision	81.41	82.01	81.05	81.82
	Recall	79.90	80.57	80.13	81.58
NCBI-disease	F1	85.36	86.23	86.35	87.35
	Precision	84.39	84.63	84.76	86.42
	Recall	87.38	87.90	88.10	88.32

4.2.2 Implementation of Variation Weights

From result of BiLM variations, the behavior improvement of our proposed approach BiLM-NER pretraining shows its positive impact on the biomedical learning model.

But this result can only be usefully for a particular named biomedical lexical predefined.

To get a generalized impact compatible to all kind of biomedical text based mining extraction, in this section we design the pretraining of the model weights to improve their generalization ability on the test set. Another positive aspect concerns the confirmation of good performance observed on *F1 scores*.

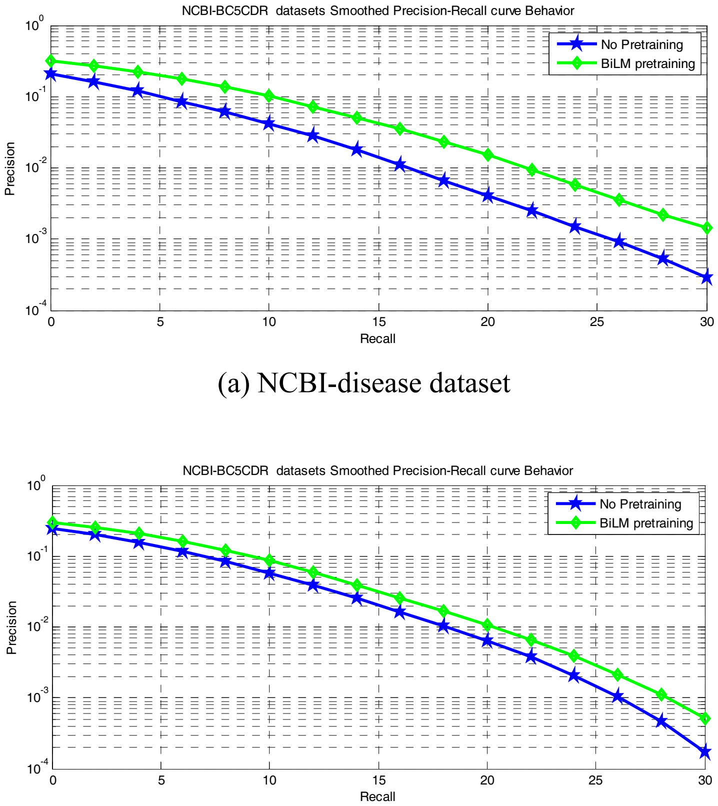
Now, let define *4dataset classes* in our biomedical text mining NER interpretation as: the *no Derivative*, the LM_F *Derivative*, the LM_B *Derivative*, and the *BiLM*:

(1) The No-derivative stands for the initialization of the NER model, which is made randomly on all text including expression, based biomedical text mining.

(2) The LMF-derivative consists to initialize NER parameters with the adoption of model weights. In this process, we also initialize randomly LSTM and CRF.

(3) The LMB-derivative is used to initialize NER parameters; LSTM and CRF are also randomly initialized.

(4) BiLM derivative is finally initialized on NER forward expression model based-weights. The resume of result is shown in Fig. 4.2, where the behavior of variation model expressed is obtained in the same implementation context of data listed in Table 3. From Fig. 3, *F1 scores* on NCBI-disease confirm performance improvement at *2.1%* and *0.6%.*



(a) NCBI-disease dataset

(b) BC5CDR dataset

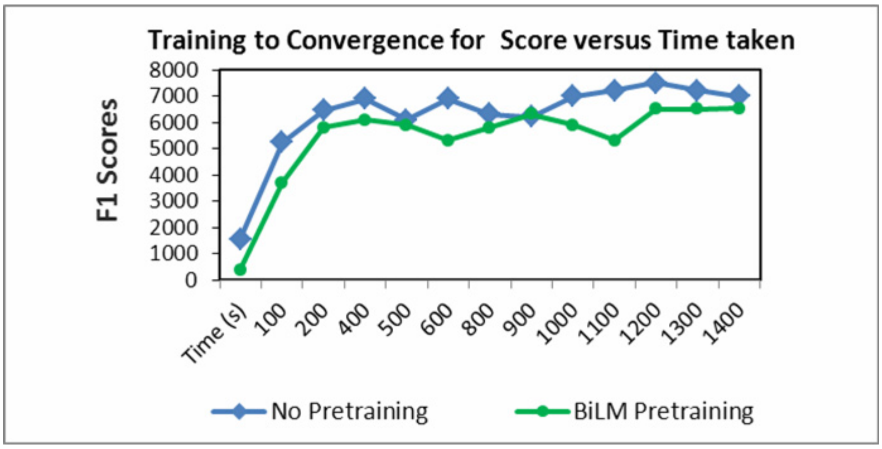
Fig. 4. Precision-recall Model for BiLM pretraining and no pretraining

In Fig. 4, after implementing NCBI dataset and BC5CDR, we show the result of pre-precision and precision of BiLM approach; the smoothed precision-recall curve for NCBI-disease andBC5CDR datasets presents an optimization as we can on Fig. 4(a) and Fig. 4(b). The precision values for BiM are respectively up to *0.92* and *0.94*, and the factors of its corresponding dataset implemented to evaluate *recall value* are also up to *0.850* and *0.875*. This result describes also the behavior of LM_F and LM_B which confirm a pretraining improvement. The BiLM precision achieves better *F1 score*.

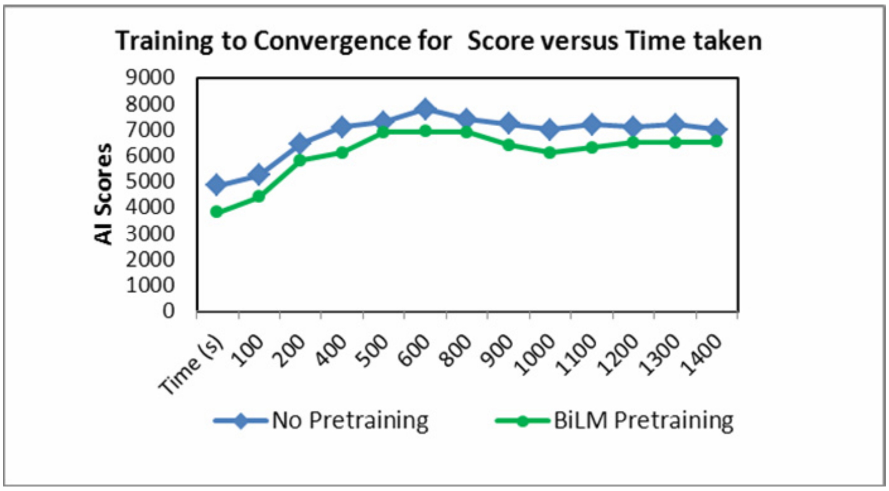
4.2.3 Implementation of Rate Convergence

The implementation of BiLM variations in the insertion of model variation weights only testify the positive impact of named biomedical lexical. However, today the time evaluation factor is considered as a big challenge in most of the system implemented; the most objective with time estimation is to obtain fast processing.

To set up time factor and its value, let in this section use *rate convergence*; we adopt "*Clock time*" and "*time taken*" two elements to design our convergence model. We choose to use the same training process as outlined in *Section 4.2* where NCBI-disease and BC5CDR dataset results were performed as shown respectively in Fig. 5(a) and Fig. 5(b). The result of BiLM pretraining show that there is convergence near to *600*, which is better compared with the no-pretraining result (*results near to 700*). The similar observation is also made in BC5CDR dataset where the results are approaching *700s* where the no-pretraining is close to *700*. There is fast convergence observed around *30-40%* compared with *F1* and random score initialized above.



(a) Results with NCBI-disease dataset



(b) Results with BC5CDR dataset

Fig. 5. Results of score versus time taken during training to converge using BiLM Pretraining/No Pretraining

This result confirms the behavior of Entity Extraction and Network Model as shown in Fig. 6, where it extracts meaningful information from biomedical lexical text by using mining system hidden networks within NCBI and BC5CDR data set. A *high level* of *F1 score* is obtained with the model pretrainning based BiLM.

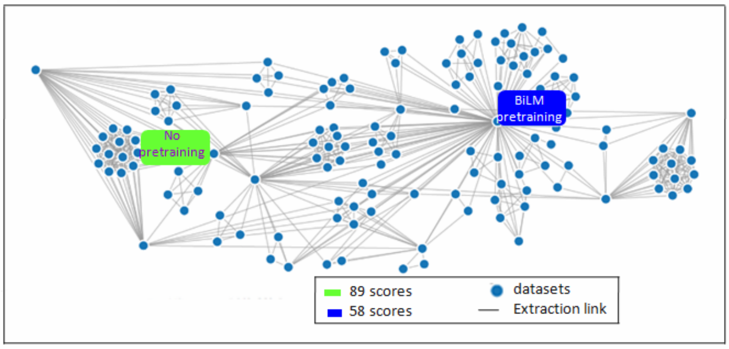
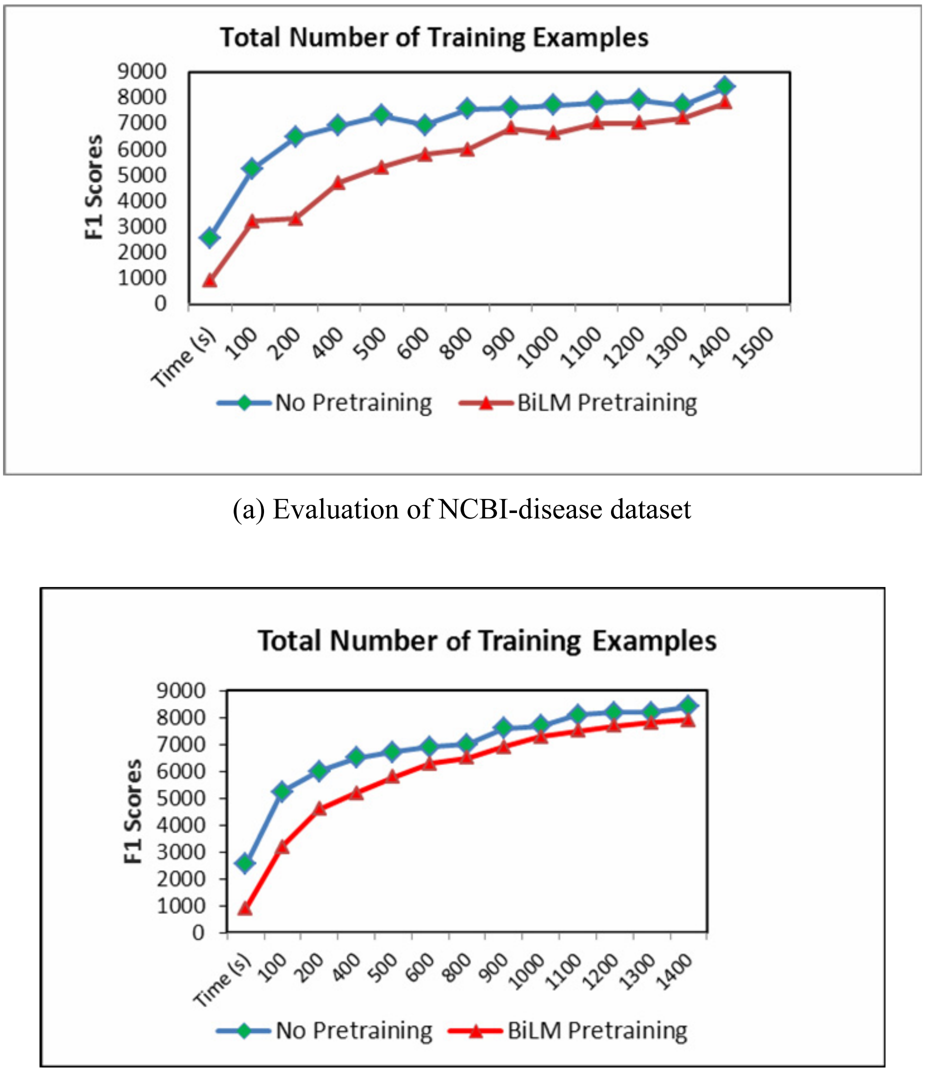


Fig. 6. Entity Extraction and Network Model

The Fig. 7 gives the studied case of NCBI disease which describes the qualitative results of the BiLM NER dataset. The collected information has been extracted from series of different biomedical documents. As shown in Fig. 8, the number of NCBI-disease is combined with the training and the development set shows performances up to *136 score* on disease entities. There are only *11* unique occurrences of disease names and the BiLM pretraining NER model is able to correctly predict more than *152 cases*.



(a) Evaluation of NCBI-disease dataset

(b) Evaluation of BC5CDR dataset

Fig. 7. Results of F1 score versus by increasing the number of BiLM pretraining/No pretraining examples.

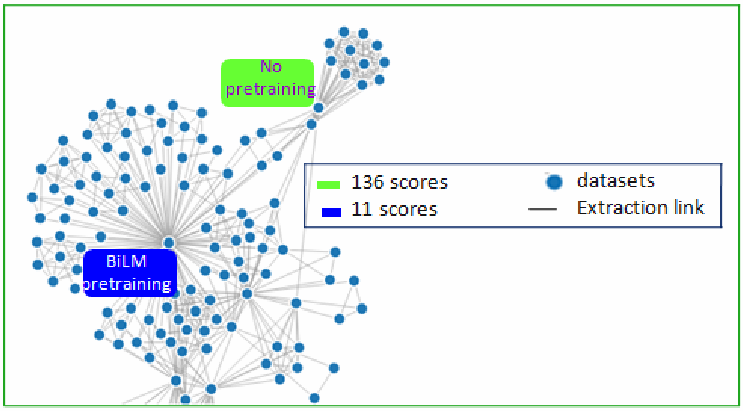


Fig. 8. Number of Entity Extraction for NCBI-disease dataset

The results of the learning curve detailed in Fig. 8 shows the results of *F1 score* versus by increasing the number of BiLM in pretraining. From both cases, the models on NCBI-disease and BC5CDR datasets, we can see that the BiLM pretraining model setting number is always optimal (higherthan *F1 score*).

4.2.4 Comparative Study between NER-BiLM, CCA and Canonical Correlation

To get a generalized impact compatible between our NER-BiLM F1 score and Global vector technique and CCA, the principle adopted in simulation works are defined on biomedical named entity recognition (Bio-NER). We choose the context "word embedding" strategy approach with five-dimensional (*10, 30, 50, 80*, and *100*).

Fig. 9, Fig. 10 and Fig. 11 indicate the experimental results of our proposed NER-BiLM. These results are obtained with various dimensions (from 3 to 11) and window sizes.

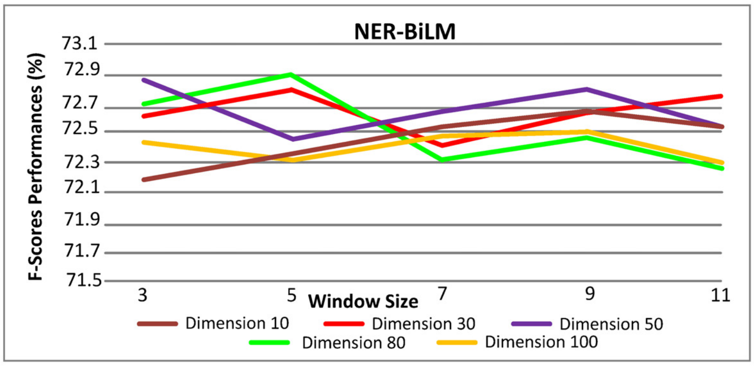


Fig. 9. Experimental results of Proposed NER-BiLM. X-axis stands for window size, and Y-axis represents F1 Performances. The Five lines correspond to five dimension sizes.

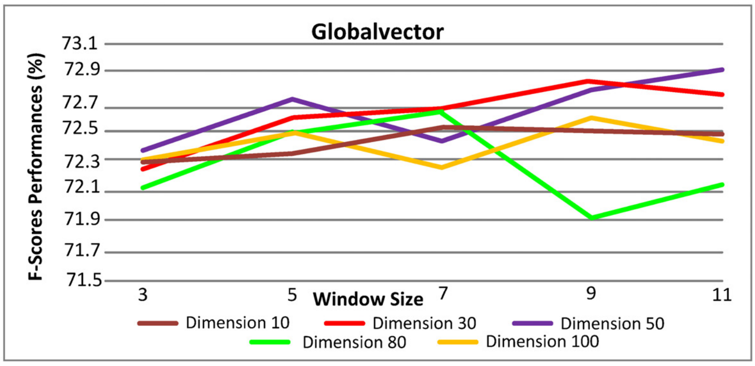


Fig. 10 Experimental results of Globalvector. The X-axis stands for window size, and Y-axis represents the F1 performance. Five lines correspond to five dimension sizes

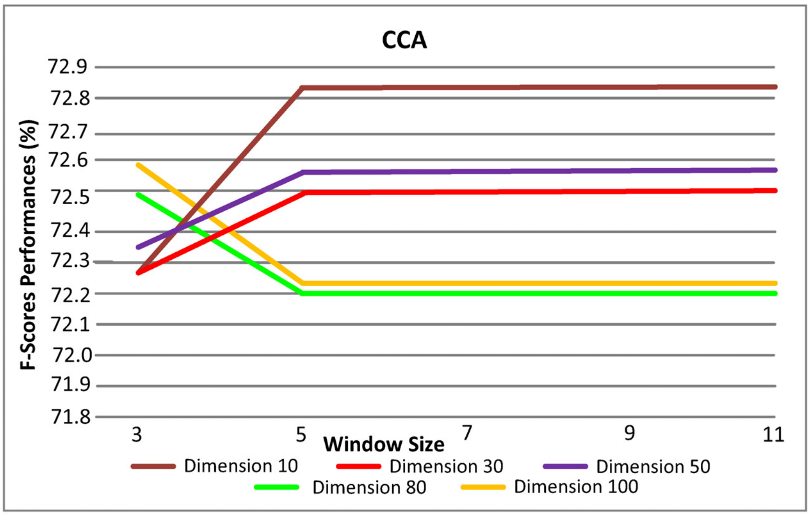


Fig. 11. Experimental results of CCA. The X-axis stands for window size, and Y-axis represents the F1 performance. Five lines correspond to five dimension sizes

NER-BiLM shows high performance when the dimension size approaches *80%* at the window size corresponding to 5. However, the same line presents also lowest performance when size increase to 7. In Fig. 9, the line corresponding to *dimension size 50* shows presents a moderately and stable high performance for all window sizes displayed. NER-BiLM does not appear to require high-dimensional representation, in the same context, lower-dimensional representations show an increase in performance proportional to the window size. Higher-dimensional representations don't exhibit specific characteristics in our NER-BiLM.

In Global vector, the corresponding behavior of *50 dimensions* and 11 window sizes display the highest performance. As it appears in NER-BiLM, Global vector approach also shows relatively stable and high performance when its *size* of a dimension is approaching *50*. But 30 dimensional score also is considered as higher due to the dimensional evaluation context.

Fig. 11 performance shows that lower-dimensional representations have approximately higher F1 performance than higher-dimensional cases when CCA is typically used in the embedding context.

To point out our approach technique estimation, Table 4 gives global details for a comparative approach between our proposed NER-BiLM method, Global vector technique, and CCA. Our system shows an F1 score of up to *72.82%*, which is a similar performance observed in all figures (Fig. 9, Fig. 10 and Fig. 11).

Table 4. Resume of Performance comparison betweenNER-BiLM, Global vector and CCA using BioNLP/NLPBA 2004 corpus environment

System Platform	Approachadopted	F1 score (%)
Our approach	NER-BiLM	72.82
	Globalvector	72.75
	CCA	72.73
Zhou and Su [25]	HMM, SVM	72.50%
Song et al. [26]	HMM and CRF	66.28%
Ponomareva et al. [27]	HMM environment	65.7%
Saha et al. [28]	HMM environment	67.41%
Finkel et al. [29]	HMM environment	69.8%
Settles [30]	HMM environment	70.2%
Tsai et al. [31]	HMM environment	870.82%

The final performance results are resumed in Table 4 which presents the main comparative performances on recent work developed in the same research area; Zhou and Su [25] adopted the methodology based on Hidden Markov Models (HMMs), and to achieve *72.50%* performance on biomedical named entity recognition (Bio-NER) system, and this performance has been the highest until now. In the same area, to improve the quality, Song et al. [26] achieved *66.28%* by using HMM and CRF. Ponomareva et al. [27] achieved a performance of *65.7%* limiting his investigation in HMM environment. Saha et al. [28] based the F1 performance by adopting Maximum Entropy factor and get *67.41%* of scores. Other relative works show some similar results, the case of Finkel et al. [29], Settles [30], and Tsai et al. [31], where the final results obtained were respectively *69.8%, 70.2%*, and *70.82%.*

Our final performance shows *72.82%*, which represent a gap of *0.3%* compared with Zhou and Su score on NER-BiLM. The comparative approach with Word embedding opens another particularity due to its advantage by offering automatic construction through unsupervised learning systems. But our approach does not require any domain knowledge or dictionary. However, our approach outperforms highly all kind of Word embedding and can be used for all tested methods.

5 Conclusion and Future Work

5.1 Conclusion

Text Mining and data extraction in biomedical and clinical domain presents today critical lackson reproducibility. This observation affects a certain rigor requirements in most of recent published researches. In this article, a new model based-transfer learning technique with BiLM for biomedical task is proposed. With the adoption of Biomedical NER, this article implements four filters in the purpose to extract a high level of the words-level BiLSTMF1 scores. The designed BiLM model shows a performed training architecture weights result with a good initialized strategy to optimize *F1 scores,* by providing consistent improvements with the adoption of four model variation Weights datasets (*No-derivative, LM_F-derivative, LM_B-derivative* and *BiLM derivative*).

To confirm this new performance, BiLSTM-NER is compared with CCA and Canonical Correlation two based machine-learning approaches. The result confirms the efficient position of BiLSTM-NER with a gap of 0.3%. This proposed model is suitable for scholars, it offer consistent interpretation with stable configuration during the network analysis process. There is a large possibility to obtain concise numbers of data strategically. This new approach can be used as a helpful learning tool to evaluate biomedical extraction data in machine learning and get appropriate clustering designs.

5.2 Future Works

For future work, new investigation will be defined on three prerogatives; first, the design of a large optimized based biomedical named entity recognition (Bio-NER) artificial framework. The purpose is to achieve the limitation of learning knowledge observed in this paper. The second contribution will consist to develop unsupervised learning methods for the Bio-NER. This aims to satisfy the new condition of corpus and although RNN requirement. Third, embedding work need more optimization on various linguistic resources for domain knowledge. For this reason, we propose to build a performed word embedding methods for resources extraction.

Acknowledgements

We would like express deep acknowledgement to Wuhan University of Technology (Computer Science School) team to all contribution on analysis and subject implementation; special thanks also for all colleagues from Beijing Universities to their special contribution in evaluation and simulation interpretation process.

References

[1] Y. Zhang, M. Chen, L. Liu, A review on text mining, in: Proc. 2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS), 2015.

[2] G. Balikas, A.Krithara, I. Partalas, G. Paliouras, BioASQ: a challenge on large-scale biomedical semantic indexing and question answering, in: Proc. International Workshop on Multimodal Retrieval in the Medical Domain (MRDM), 2015.

[3] H. Shatkay, N. Chen, D. Blostein, Integrating image data into biomedical text categorization, Bioinformatics 22(14)(2006) e446-453.

[4] A. Ahmed, E.P. Xing, W.W. Cohen, R.F. Murphy, Structured correspondence topic models for mining captioned figures in biological literature, in: Proc. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009.

[5] S.V. Gaikwad, A. Chaugule, P. Patil, Text mining methods and techniques, International Journal of Computer 85(17)(2014) 0975-8887.

[6] S.A. Salloum, M. Al-Emran, A.A. Monem, K. Shaalan, A Survey of Text Mining in social media: Facebook and twitter perspectives, Advances in Science, Technology and Engineering Systems Journal 2(1)(2017) 127-133.

[7] N.S.B. Ramez, Data warehousing and data mining, Database System 7(2)(2007) S33-S40.

[8] G. V. Lehal, A survey of text mining techniques and applications, Journal of Emerging Technologies in Web Intelligence 1(1)(2009) 60-76.

[9] S. Saha, A. Ekbal, U.K. Sikdar, Named entity recognition and classification in biomedical text using classifier ensemble, Data Mining and Bioinformatics 11(4)(2015) 365-391.

[10] D. Campos, S. Matos, J. L. Oliveira, Biomedical named entity recognition: a survey of machine-learning tools, Theory Applications for Advanced Text Mining 2(8)(2012) 175-195.

[11] K.R. Rahem, N. Omar, Rule-based named entity recognition for drug-related crime news documents, Journal of Theoretical & Applied Information Technology 77(2)(2015) 001-700.

[12] R.M.R. Zavala, P. Martinez, I. Segura-Bedmar, A hybrid Bi-LSTM-CRF model for knowledge recognition from health documents, in: Proc. Third Workshop on Evaluation of Human Language Technologies for Iberian Languages, 2018.

[13] M. Talha, S. Boulaknadel, D. Aboutajdine, Enhancing performance of hybrid named entity recognition for amazighe language, in: A.E. Hassanien (Ed.), Machine Learning Paradigms: Theory and Application, Springer, Cham, Switzerland, 2018, pp.211-232.

[14] E.F. Tjong, K. Sang, F. De Meulder, Introduction to the CoNLL-2003 shared task: language-independent named entity recognition, in: Proc. Seventh Conference on Natural Language Learning at HLT-NAACL 2003, 2003.

[15] G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, R. Weischedel, The Automatic Content Extraction (ACE) program – tasks, data, and evaluation, in: Proc. Fourth International Conference on Language Resources and Evaluation (LREC’04), 2004.

[16] R. Grishman, B. Sundheim, Message Understanding Conference-6: a brief history, in: Proc. 16th Conference on Computational Linguistics - Volume 1, 1996.

[17] M. Yehia, L. Fattouh, M. Abulkhair, Text Mining and Knowledge Discovery from Big Data: Challenges and Promise, Computer Science 13(3)(2016) 54-61.

[18] Q. Yang, X. Wu,10 Challenging Problems in Data Mining Research, Information Technology & Decision Making 5(4)(2016) 597-604.

[19] Y. Shen, H. Yun, Z.C. Lipton, Y. Kronrod, A. Anandkumar, Deep Active learning for named entity recognition. <https://arxiv.org/abs/1707.05928>, 2017.

[20] S. Pradhan, A. Moschitti, N. Xue, O. Uryupina, Y. Zhang, CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes, in Proc. Sixteenth Conference on Computational Natural Language Learning (CoNLL2012), 2012..

[21] Y. Zhang, M. Chen, L. Liu, A review on text mining, in: Proc. 2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS), 2015.

[22] R. Leaman, C.-H. Wei, Z. Lu, tmChem: a high-performance approach for chemical named entity recognition and normalization, Chemical informatics 7(1)(2015) S. DOI: 10.1186/1758-2946-7-S1-S3.

[23] D. Zeng, C. Sun, L. Lin, B. Liu, LSTM-CRF for drug-named entity recognition, Entropy 19(6)(2017) 283.

[24] J.P.C. Chiu, E. Nichols, Named entity recognition with bidirectional LSTM-CNNs, Transactions of the Association for Computational Linguistics 4(2)(2016) 357-370.

[25] G.D. Zhou, J. Su, Exploring deep knowledge resources in biomedical name recognition, in: Proc. International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP), 2004.

[26] Y. Song, E. Kim, G.G. Lee, B.-k. Yi, POSBIOTM-NER in the shared task of BioNLP/NLPBA, in: Proc. International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP), 2004.

[27] N. Ponomareva, F. Pla, A. Molina, P. Rosso, Biomedical named entity recognition: a poor knowledge HMM-based approach, in: Proc. Natural Language Processing and Information Systems, 12th International Conference on Applications of Natural Language to Information Systems, 2007.

[28] S.K. Saha, S. Sarkar, P. Mitra, Feature selection techniques for maximum entropy-based biomedical named entity recognition, Journal of Biomedical Informatics 42(5)(2018) 905-911.

[29] J. Finkel, S. Dingare, H. Nguyen, M. Nissim, C. Manning, G. Sinclair, Exploiting context for biomedical entity recognition: From syntax to the Web, in: Proc. International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP), 2004.

[30] B. Settles, Biomedical named entity recognition using conditional random fields and rich feature sets, in: Proc. International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP), 2004.