Preproceedings of the 19th Conference on Computer
Science and Intelligence Systems (FedCSIS) pp. 219–224

LSTM-based Deep Neural Network With A Focus on Sentence Representation for Sequential Sentence Classification in Medical Scientific Abstracts

Phat Lam[*§], Lam Pham[†§], Tin Nguyen[*], Hieu Tang [‡], Michael Seidl[†], Medina Andresel[†], Alexander Schindler[†]
[*]Ho Chi Minh University of Technology
ORCIDs: 0009-0003-5105-5976, 0009-0006-4615-5624
Email: {phat.lamhcmutddk21, tin.nguyen112101bku}@hcmut.edu.vn
[†]Austrian Institute of Technology, Austria
ORCIDs: 0000-0001-8155-7553, 0000-0003-4109-1335, 0009-0002-4424-7817, 0000-0002-4881-6741
Email: {lam.pham, michael.seidl, medina.andresel, alexander.schindler}@ait.ac.at
[‡]FPT University, Vietnam
ORCID: 0009-0006-7922-4040
Email: hieutq10@fpt.edu.vn
[§] These authors contributed equally.

***Abstract*—The Sequential Sentence Classification task within the domain of medical abstracts, termed as SSC, involves the categorization of sentences into pre-defined headings based on their roles in conveying critical information in the abstract. In the SSC task, sentences are sequentially related to each other. For this reason, the role of sentence embeddings is crucial for capturing both the semantic information between words in the sentence and the contextual relationship of sentences within the abstract, which then enhances the SSC system performance. In this paper, we propose a LSTM-based deep learning network with a focus on creating comprehensive sentence representation at the sentence level. To demonstrate the efficacy of the created sentence representation, a system utilizing these sentence embeddings is also developed, which consists of a Convolutional-Recurrent neural network (C-RNN) at the abstract level and a multi-layer perception network (MLP) at the segment level. Our proposed system yields highly competitive results compared to state-of-the-art systems and further enhances the F1 scores of the baseline by 1.0%, 2.8%, and 2.6% on the benchmark datasets PudMed 200K RCT, PudMed 20K RCT and NICTA-PIBOSO, respectively. This indicates the significant impact of improving sentence representation on boosting model performance.**

***Keywords*— sentence representation, sequential sentence classification, bidirectional long short-term memory network, multiple feature branches.**

I. Introduction

WHEN researching a large-scale source of scientific papers, it is necessary to skim through abstracts to identify whether papers align with the research interest. This process becomes more straightforward when abstracts are organized with semantic headings such as "background", "objective", "methods", "results", and "conclusion". Therefore, automatically categorizing each sentence in a scientific abstract into a relevant heading, known as the task of Sequential Sentence Classification (SSC), significantly facilitates the information retrieval process within large-scale data. In medical domain, research abstracts present a large volume and have grown exponentially. Manually sorting through these documents to find relevant insights presents a time-consuming and labor-intensive task, highlighting the need for efficient information retrieval and summarization methods without entirely reading full-text content in medical scientific articles [1], [2]. Therefore, the result of the SSC tasks significantly enables researchers and learners to catch up and categorize research abstracts effectively. In other words, the SSC task significantly facilitates learners and researchers by accelerating their educational processes of literature review, information extraction, evidence-based decision-making, etc. Recently, the SSC task in medical scientific abstracts has drawn attention from NLP research community. Indeed, some large and benchmark datasets such as PubMed RCT [3] and NICTA-PIBOSO [4] were published. Additionally, a wide range of machine learning-based and deep learning-based models have been proposed for this task [5]. Traditional machine learning methods utilized hand-crafted feature extraction for individual sentences. These extracted features are related to lexical, semantic, and structural information of an individual sentence such as synonyms, bag-of-words, part of speech, etc. Then, sentences are classified by Hidden Markov Model (HMM) [6], Naive Bayes [7] or CRF [8]. While the traditional machine learning-based models present a limitation of exploring the relation among the sentences as using the hand-crafted features, leveraging deep neural networks in deep learning-based models allows to capture the patterns of contextual relationship among sentences in the same abstract that leads to a breakthrough on model performance. For example, Dernoncourt et al. [9] introduced a deep learning model that uses a CRF layer to optimize the predicted label sequence, where adjacent sentences have an impact on the prediction of each other. Jin and Szolovits [10] proposed a hierarchical sequential labeling network to further improve the semantic information within surrounding sentences for classification. Recently, Yamada et al. [11] and Shang et al. [5] introduced some methodologies to assign labels to span sequences at the span level, which achieved state-of-the-art

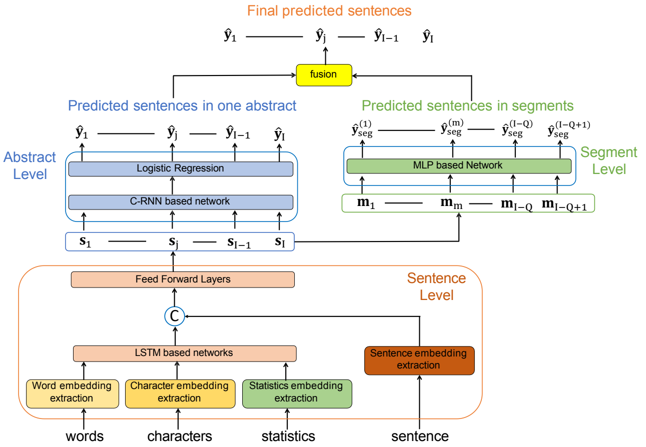
Final predicted sentences
$\hat{y}_1$ — $\hat{y}_j$ — $\hat{y}_{I-1}$ $\hat{y}_I$
fusion
Predicted sentences in one abstract
Predicted sentences in segments
$\hat{y}_1$ — $\hat{y}_j$ — $\hat{y}_{I-1}$ $\hat{y}_I$
$\hat{y}_{seg}^{(1)}$ — $\hat{y}_{seg}^{(m)}$ — $\hat{y}_{seg}^{(I-Q)}$ $\hat{y}_{seg}^{(I-Q+1)}$
Abstract Level
Logistic Regression
C-RNN based network
Segment Level
MLP based Network
m_1 — m_m — m_{I-Q} m_{I-Q+1}
s_1 — s_j — s_{I-1} s_I
Feed Forward Layers
Sentence Level
C
LSTM based networks
Sentence embedding extraction
Word embedding extraction
Character embedding extraction
Statistics embedding extraction
words
characters
statistics
sentence

Fig. 1. The overall architecture of the proposed system

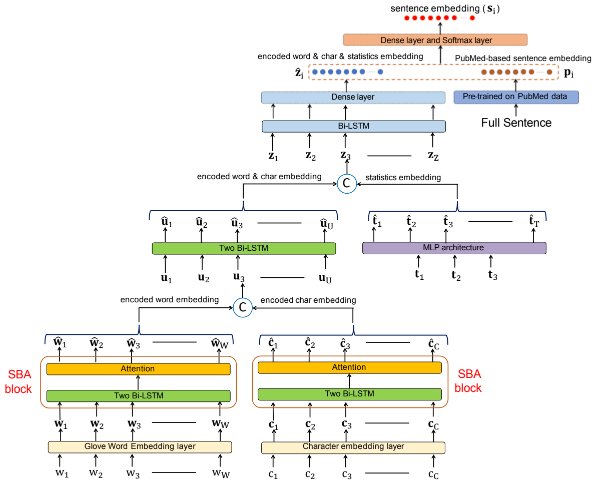
sentence embedding (s_i)
Dense layer and Softmax layer
encoded word & char & statistics embedding
PubMed-based sentence embedding
$\hat{z}_i$
p_i
Dense layer
Pre-trained on PubMed data
Full Sentence
Bi-LSTM
z_1 z_2 z_3 — z_Z
encoded word & char embedding
C
statistics embedding
$\hat{u}_1$ $\hat{u}_2$ $\hat{u}_3$ — $\hat{u}_U$
$\hat{t}_1$ $\hat{t}_2$ $\hat{t}_3$ — $\hat{t}_T$
Two Bi-LSTM
MLP architecture
u_1 u_2 u_3 — u_U
t_1 t_2 t_3
encoded word embedding
C
encoded char embedding
$\hat{w}_1$ $\hat{w}_2$ $\hat{w}_3$ — $\hat{w}_W$
$\hat{c}_1$ $\hat{c}_2$ $\hat{c}_3$ — $\hat{c}_C$
SBA block
Attention
Two Bi-LSTM
Attention
Two Bi-LSTM
SBA block
w_1 w_2 w_3 — w_W
c_1 c_2 c_3 — c_C
Glove Word Embedding layer
Character embedding layer
w_1 w_2 w_3 — w_W
c_1 c_2 c_3 — c_C

Fig. 2. The Sen-Model architecture for classification at the sentence level

results. However, these two systems consider all possible span sequences with various lengths, which is very computationally expensive on large datasets [11]. Importantly, these systems attempted to analyze the sequence of sentences at the span level without initially considering the improvement of the sentence representation, which is the fundamental component of this specific SSC task. At the sentence level, these systems leveraged sentence embeddings extracted from pre-trained BERT model [12], which is trained on biomedical text for various NLP tasks such as Name entity recognition, Sentence similarity, etc. Commonly, BERT models primarily focus on capturing syntactic meaning and contextual dependencies of words within individual sentences or pairs of sentences [13]. To some extent, the extracted sentence embeddings may lack the ability to grasp dependencies between sentences in a wider context (e.g. abstracts, documents), which is one of the most crucial task-specific properties of the SSC task. To the best of our knowledge, there has been very little to no research dedicated to independently improving sentence embeddings specifically for the SSC task.

In this paper, we therefore aim to improve the sentence representation and explore its impact on the performance of SSC task in medical scientific abstracts. We propose a deep neural network with a focus on extracting well-presented sentence embeddings. In particular, we explore the independent features of sentence, word sequence, character sequence, and statistic information of sentences in one abstract. Then, we develop a LSTM-based deep neural network with multiple-feature branches for classifying individual sentences. The network is then used to extract the comprehensive sentence embeddings. Given these sentence embeddings, a system including a Convolutional-RNN based network (C-RNN) at the abstract level and a Multi-layer Perception network (MLP) at the segment level (i.e. a segment includes a fixed-length group of consecutive sentences) is introduced to extensively learn the contextual patterns of sentences in the same abstract. Finally, the results of C-RNN and MLP models are fused to achieve the final predicted sentences in an abstract. We evaluate our proposed models on two benchmark datasets, PubMed RCT [3] and NICTA-PIBOSO [4]. The experimental results indicate that exploiting multiple features extracted from sentences such as word sequence, character sequence, and statistical information of sentences in the abstract potentially helps to generate well-presented sentence embeddings at the sentence level. Both C-RNN network at the abstract level and MLP network at the segment level respectively further improve the performance when leveraging these well-presented sentence embeddings.

II. The Overall Proposed System

The proposed system in this paper for the task of sequential sentence classification in medical scientific abstracts is generally presented in Fig. 1.

As Fig. 1 shows, the proposed network comprises of three main sub-networks, referred to as the classification model (Sen-Model) at the sentence level, the regression model at the abstract level (Abs-Model) and the classification model at the segment level (Seg-Model). At the sentence level, we establish the task sentence classification for individual sentences. The proposed LSTM-based classification model at the sentence level (Sen-Model) presents 4 branches, each of which explores the distinct feature from the full sentence, the words in the sentence, the character in the sentence, and the statistical information of the sentence in one abstract, aiming to achieve the comprehensive and contextually adaptive sentence representation. Given the classification model at the sentence level, we extract the sentence embeddings $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_S]$, where each $\mathbf{s}_i$, $i = 1, 2, ..., S$, represents an individual sentence. The sentence embeddings are then utilized in the regression model at the abstract level (Abs-Model) and the classification model at the segment level (Seg-Model) to further improve the tasks of sentence classification by exploiting the properties of the well-presented sentence representation at higher contextual levels. Both the classification model at the sentence level and the regression model at the abstract level leverage RNN-based architecture, attention mechanism, and multi-layer perception (MLP) architecture which are comprehensively presented in next sections.

A. The classification model at the sentence level (Sen-Model)

The proposed LSTM-based network focuses on improving sentence representation at the sentence level (Sen-Model) is comprehensively presented in Fig. 2. Given a sentence including W words $[w_1, w_2, ...w_W]$ and C characters $[c_1, c_2, ...c_C]$. We make use of the Glove [14] model to extract a sequence of word embeddings $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, ...\mathbf{w}_W]$, where $\mathbf{w}_w \in \mathbb{R}^{d_w}$ presents a word embedding and d_w is the dimension of a word embedding. Regarding the sequence of characters in one sentence, the character embedding is randomly initialized in the uniform distribution to extract the character embeddings $\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, ...\mathbf{c}_C]$, where $\mathbf{c}_c \in \mathbb{R}^{d_c}$ presents a character embedding and d_c is the dimension of a character embedding.

The sequence of word embeddings $\mathbf{W}$ and the sequence of character embeddings $\mathbf{C}$ are fed into stacked Bi-LSTM-Attention encoder blocks, referred as SBA blocks, to generate the encoded word embeddings $\widehat{\mathbf{W}} = [\widehat{\mathbf{w}}_1, \widehat{\mathbf{w}}_2, ...\widehat{\mathbf{w}}_W]$ and the encoded character embeddings $\widehat{\mathbf{C}} = [\widehat{\mathbf{c}}_1, \widehat{\mathbf{c}}_2, ...\widehat{\mathbf{c}}_C]$, where $\widehat{\mathbf{w}}_w, \widehat{\mathbf{c}}_c \in \mathbb{R}^{d_h}$ and d_h is the hidden state dimension. The SBA block includes a Bi-LSTM network which comprises of two stacked Bidirectional LSTM layers, followed by a Scaled Dot-Product Attention layer [15]. Each Bidirectional LSTM layer takes the output sequence of the previous layer as input, which allows the capture of more complex lexical, syntactic, and semantic information between words and characters in an individual sentence. Given the sequential word representation and the sequential character representation extracted from the Bidirectional LSTM layers, we apply linear transform to create query, key and value matrix $\mathbf{Q} \in R^{N_q \times d_l}$, $\mathbf{K} \in R^{N_k \times d_l}$, $\mathbf{V} \in R^{N_v \times d_v}$, where N_q, N_k, N_v are the number of queries, keys and values, d_l and d_v are the dimension of query and key, the dimension of value, respectively. The output matrix of the Scaled Dot-Product attention layer is computed as:

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{QK}^T}{\sqrt{d_l}}\right)\mathbf{V} \quad (1)$$

Two encoded embeddings $\widehat{\mathbf{W}}$ and $\widehat{\mathbf{C}}$ extracted from SBA blocks of words and characters are concatenated to generated the word-char embedding $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, ...\mathbf{u}_U]$ where $U = W + C$ and $\mathbf{u}_u \in \mathbb{R}^{d_h}$. The word-char embedding $\mathbf{U}$ is then fed into the word-char encoder block to generate the encoded word-char embeddings $\widehat{\mathbf{U}} = [\widehat{\mathbf{u}}_1, \widehat{\mathbf{u}}_2, ...\widehat{\mathbf{u}}_U]$. The word-char encoder block reuses the two stacked Bidirectional LSTM layers from the SBA block without using the attention layer.

Besides the lexical, syntactic and semantic information for each sentence extracted from the word and character branches, we consider the statistical information of individual sentence: the number of sentences in the same abstract, the index of sentence in the abstract, and the number of words in the sentence, which are represented by one-hot vectors $\mathbf{t_1}, \mathbf{t_2}, \mathbf{t_3}$. The statistical information equips each sentence with the ability to capture sequential and contextual properties related to other sentences within abstract. The statistical vectors are fed into a Multi-layer perception (MLP) to generate encoded statistic embeddings $\widehat{\mathbf{T}} = [\widehat{\mathbf{t}}_1, \widehat{\mathbf{t}}_2, ...\widehat{\mathbf{t}}_T]$. The encoded statistic embeddings $\widehat{\mathbf{T}}$ are then concatenated with the encoded word-char embedding $\widehat{\mathbf{U}}$ to generate the word-char-stat embedding $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, ...\mathbf{z}_Z]$ where $Z = U + T$ and $\mathbf{z}_z \in \mathbb{R}^{d_h}$. Again, one Bi-LSTM layer and one Dense layer are used to learn the sequence of word-char-stat embeddings, which combine both semantic, syntactic information of word-char encoded embedding and statistical information of statistic embedding, to generate the encoded word-char-stat embeddings $\widehat{\mathbf{Z}}$.

To further enhance the representation of sentences in term of language comprehension specifically in biomedical domains, we utilize BiomedBERT [16], which was pretrained on the PubMed corpus. The PubMed-based sentence embedding $\mathbf{P} = [\mathbf{P}_1, \mathbf{P}_2, ...\mathbf{P}_P]$ is concatenated with the encoded word-char-stat embedding $\widehat{\mathbf{Z}}$ before feeding into a Dense layer followed by a Softmax layer for classification. After training the Sen-Model at the sentence level, for each sentence, we extract the output of the Dense layer before the final Softmax layer and consider it as the final sentence-level embedding $\mathbf{s}_i \in R^{d_L}$, where d_L is the dimension of a sentence-level embedding (i.e. the Softmax layer presents L outputs which match L labels of sentences in an abstract). The final sentence representation of the entire dataset is $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, ...\mathbf{s}_S]$, where S is the number of sentences in the dataset and $\mathbf{s}_i \in R^{d_L}$. The Sen-Model is optimized using Categorical Cross-Entropy:

$$L_{\text{Sen}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{L} y_{ij} \log \widehat{y}_{ij} \quad (2)$$

where $\mathbf{y}_i$, $\widehat{\mathbf{y}}_i$ and N are the true label, the predicted probability vector of sentence $\mathbf{s}_i$ and the batch number, respectively. To examine the efficacy of the extracted sentence embeddings, we constructed two networks aimed at enhancing performance at higher levels by leveraging these embeddings. The networks are presented in the next subsections.

B. The regression model at the abstract level (Abs-Model)

Given the original dataset comprising of S sentences $[s_1, s_2, ...s_S]$, each sentence is now represented by a sentence-level embedding $\mathbf{s}_i$, $i = 1, 2, ..., S$, extracted from the Sen-Model at the sentence level. To explore the sequential and contextual properties of sentences in one abstract, we group sentence embeddings in the same abstract to create the abstract representation $\mathbf{A} = [\mathbf{s}_1, \mathbf{s}_2, ...\mathbf{s}_I]$ where $\mathbf{s}_i \in \mathbb{R}^{d_L}$ and I is the number of sentences in one abstract. The abstract representation $\mathbf{A}$ is a sequence of sentence-level embeddings which is fed into the regression model at the abstract level (Abs-Model). The regression model at the abstract level (Abs-Model) is comprehensively presented at the left corner of the upper part of Fig. 1. The network includes two parts: Convolutional-Recurrent Neural Network (C-RNN) and Logistic Regression classifier.

Each abstract representation $\mathbf{A}$ is now considered as a two-dimensional tensor which is fed into the convolution layers to extract essential features represented for neighbour sentences in one abstract. The two 2D-convolution layers in the C-RNN

TABLE I
MLP BASED NETWORK FOR CLASSIFICATION AT THE SEGMENT LEVEL

Blocks	Layers	Output Shape
F1	Dense (512) - Elu - BN -Dr(0.5)	512
F2	Dense (256) - Elu - BN - Dr(0.5)	256
F3	Dense (128) - Elu - BN - Dr(0.5)	128
F4	Dense (64) - Elu - BN - Dr(0.5)	64
F5	Dense (L) - Softmax	L

present similar settings in terms of kernel size, padding and the number of filters in each layer. Next, the Bi-RNN decoder is used for learning sequential relationship feature maps extracted from the convolutional layers. Finally, the Logistic Regression classifier receives the feature maps from the Bi-RNN decoder as input and generate predicted values $\widehat{\mathbf{Y}}_{\text{abs}} = [\widehat{\mathbf{y}}_1, \widehat{\mathbf{y}}_2, ...\widehat{\mathbf{y}}_{\text{I}}]$ corresponding to the ground truth $\mathbf{Y}_{\text{abs}} = [\mathbf{y}_1, \mathbf{y}_2, ...\mathbf{y}_{\text{I}}]$ where $\widehat{\mathbf{y}}_i, \mathbf{y}_i \in R^{d_L}$. Regarding the predicted value and the ground truth of one abstract, we form a predicted sequence $\widehat{\mathbf{y}}_{\text{abs}} \in \mathbb{R}^{d_L \times I}$ and $\mathbf{y}_{\text{abs}} \in \mathbb{R}^{d_L \times I}$ by concatenating all the vectors of $\widehat{\mathbf{Y}}_{\text{abs}}$ and $\mathbf{Y}_{\text{abs}}$, respectively. Then, the abstract-level Abs-Model is optimized using Binary Cross Entropy (BCE) loss on these predicted sequences, which can be written as:

$$L_{\text{Abs}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d_L \times I} \left(y_{ij}\log(\widehat{y}_{ij}) + (1 - y_{ij})\log(1 - \widehat{y}_{ij})\right) \tag{3}$$

where N is the batch number and the innermost sum presents the BCE loss for one abstract.

C. The classification model at the segment level (Seg-Model)

Given the extracted sentence embeddings, instead of generating all the segments with various lengths, we create fixed-length segments with the size of Q by grouping every Q consecutive sentences in one abstract. Each abstract of I sentences has $I - Q + 1$ segments. The i^{th} segment representation is described as $\mathbf{m}^{(i)} = [\mathbf{s}_{Qi}\, \mathbf{s}_{Qi+1} \ldots \mathbf{s}_{Qi+Q-1}]$, which is formed by concatenating Q continuous sentence embeddings. The corresponding label vector $\mathbf{y}_{\text{seq}}^{(i)}$ of the i^{th} segment is defined as:

$$\mathbf{y}_{\text{seq}}^{(i)} = \frac{\displaystyle\sum_{q=Qi}^{Qi+Q-1} \mathbf{y}_q}{\displaystyle\sum_{q=Qi}^{Qi+Q-1}\sum_{l=1}^{\text{L}} y_{ql}} \tag{4}$$

where $\sum_{l=1}^{\text{L}} y_{ql}$ is the sum of elements in the label vector $\mathbf{y}_q$ of the sentence $\mathbf{s}_q$. The fixed-length Q is set to 3 based on empirical experiments. The Seg-model at the segment level uses the same labels as which of the sentence level, meaning that all the sentences in a segment receive the label of that segment.

To classify segments, we use the MLP network which is shown in detail at table I. The network consists of five fully-connected blocks. The first four blocks present the same layers which perform Dense layer, ELU activation, Batch Normalization and Dropout, respectively. The output of the last block is used for segment-based classification task. Since the labels of segment embeddings are no longer one-hot encoded, we use the Kullback-Leibler (KL) divergence loss for the segment-based classification task, which is defined as:

$$L_{\text{Seg}}(\theta) = \sum_{n=1}^{N} \mathbf{y}_{\text{seq}}^{(n)} \log \frac{\mathbf{y}_{\text{seq}}^{(n)}}{\widehat{\mathbf{y}}_{\text{seq}}^{(n)}} + \frac{\lambda}{2}||\theta||_2^2 \tag{5}$$

where θ is the trainable parameters of the network, λ denotes the l_2 regularization coefficient experimentally set to 0.0001, N is the batch number, $\mathbf{y}_{\text{seq}}^{(n)}$ and $\widehat{\mathbf{y}}_{\text{seq}}^{(n)}$ denote the ground-truth and the network output in a batch, respectively.

TABLE II
DATASET STATISTICAL INFORMATION

Dataset	$\|C\|$	$\|V\|$	Train	Validation	Test
PubMed 20k	5	68k	15k/180k	2.5k/30k	2.5k/30k
PubMed 200k	5	331k	190k/2.2M	2.5k/29k	200/29k
NICTA-BIBOSO	6	17k	720/7.7k	80/0.9k	200/2.2k

D. Inference with the entire system

Given the predicted labels of Abs-Model at the abstract level and Seg-Model at the segment level, referred to as $\hat{\mathbf{Y}}_{\text{abs}}$ and $\hat{\mathbf{Y}}_{\text{seg}}$, the final predicted labels of our proposed system is defined as:

$$\hat{\mathbf{Y}} = \lambda_{\text{abs}}\hat{\mathbf{Y}}_{\text{abs}} + \lambda_{\text{seg}}\hat{\mathbf{Y}}_{\text{seg}} \tag{6}$$

where λ_{abs} and λ_{seg} are the hyperparameters to control the predicted labels at the abstract level and the segment level.

III. EXPERIMENT AND RESULTS

A. Datasets

In this paper, we evaluate our proposed deep neural networks on two benchmark datasets: PubMed RCT [3] and NICTA-PIBOSO [4].

PubMed RCT: This dataset presents the largest and published dataset of text-based medical scientific abstracts. In particular, the PubMed dataset presents approximately 200,000 abstracts of randomized controlled trials. The total sentences in the PubMed dataset is around 2.3 million. Each sentence of each abstract is labeled with 'BACKGROUND', 'OBJECTIVE', 'METHOD', 'RESULT', or 'CONCLUSION' which matches its role in the abstract. The PubMed dataset proposed two sets of PubMed 20K and PubMed 200K, each of which presents three subsets of Training, Validation, and Test for training, validation and test processes, respectively.

NICTA-PIBOSO: This dataset is the official dataset of the ALTA 2012 Shared Task. The task was to build classifiers which automatically divide sentences to a pre-defined set of categories in the domain of Evidence Based Medicine (EBM), which are 'BACKGROUND', 'INTERVENTION', 'OUTCOME', 'POPULATION', 'STUDY DESIGN', 'OTHER'. Table II presents statistics information of these above datasets, where $|C|$ denotes the number of classes, $|V|$ denotes the vocabulary size. In the train, validation and test sets, we indicate the number of abstracts and the number of sentences separated by the slash (e.g. 15k/180k).

TABLE III
COMPARE OUR PROPOSED SYSTEMS WITH THE BASELINE ON THE TEST SET (F1 SCORE/PRESISION/RECALL)

Systems	PubMed 20K	NICTA-PIBOSO
bi-ANN [9] (baseline)	90.0/-/-	82.7/-/-
Sen-Model w/ word only	84.0/84.2/83.9	69.9/70.3/69.8
Sen-Model w/ word & char	84.2/84.2/84.2	70.0/70.3/69.8
Sen-Model w/ word & char & stat	89.5/89.7/89.3	77.9/77.9/77.9
Sen-Model w/ pre-trained sentence only	87.0/87.1/87.0	78.5/78.8/78.5
Sen-Model w/ sentence & word & char & stat	**91.1/91.9/90.9**	**81.8/81.8/81.8**
Abs-Model w/ word only	90.6/91.2/90.4	81.5/83.4/80.3
Abs-Model w/ word & char	90.7/91.3/90.5	81.4/83.1/80.5
Abs-Model w/ word & char & stat	91.5/91.8/91.2	81.2/82.6/80.1
Abs-Model w/ pre-trained sentence only	91.9/92.1/91.7	82.5/84.0/81.5
Abs-Model w/ sentence & word & char & stat	**92.7/93.2/92.6**	**84.6/85.5/84.1**

TABLE IV
COMPARE OUR PROPOSED SYSTEMS ON DIFFERENT LEVELS WITH THE BASELINE ON THE TEST SET OF TWO BENCHMARK DATASETS(F1 SCORE/PRECISION/RECALL)

Systems	PubMed 20K	NICTA-PIBOSO
bi-ANN [9] (baseline)	90.0/-/-	82.7/-/-
Sen-Model (Sentence)	91.1/91.9/90.9	81.8/81.8/81.8
Abs-Model (Abstract)	92.7/93.2/92.6	84.6/85.5/84.1
Seg-Model (Segment)	91.0/92.5/89.6	79.5/80.7/78.5
Combine-Model	**92.8/93.4/92.7**	**85.3/86.5/84.5**

TABLE V
COMPARE OUR BEST MODEL WITH THE STATE-OF-THE-ART SYSTEMS ON TEST SET OF PUBMED 20K DATASET

Authors	Systems	F1-score
Yamada et al. [11]	Semi-Markov CRFs	**93.1**
Athur Brack et al. [18]	Transfer /Multi-task learning	93.0
Cohan et al. [19]	Pretrained BERT	92.9
Xichen Shang et al. [5]	SDLA	92.8
Jin and Szolovits [10]	HSLN	92.6
Gaihong Yu et al. [20]	MSM	91.2
Gonçalves et al. [21]	CNN-GRU	91.0
Dernoncourt et al. [9]	bi-ANN	90.0
Agibetov et al. [22]	fastText	89.6
Our proposed model	BiLSTM-CRNN-MLP	92.8

TABLE VI
COMPARE OUR BEST MODEL WITH THE STATE-OF-THE-ART SYSTEMS ON THE TEST SET OF NICTA-PIBOSO DATASET

Authors	Systems	F1-score
Xichen Shang et al. [5]	SDLA	**86.8**
Athur Brack et al. [18]	Transfer /Multi-task learning	86.0
Yamada et al. [11]	Semi-Markov CRFs	84.4
Jin and Szolovits [10]	HSLN	84.3
Sarker et al. [23]	SVM	84.1
Cohan et al. [19]	Pretrained BERT	83.0
Dernoncourt et al. [9]	bi-ANN	82.7
M Lui [24], [25]	Feature stacking + Metalearner	82.0
Our proposed model	BiLSTM-CRNN-MLP	85.3

B. Evaluation metric

In this paper, we follow the original paper [3], [4] which proposed the PubMed RCT and NICTA-PIBOSO datasets. We then use Precision, Recall, and F1 scores as evaluation metrics.

C. Experimental settings

We construct our proposed deep neural networks with the TensorFlow framework. While the deep neural network used for Sen-Model is trained for 30 epochs, we train Abs-Model and Seg-Model with 60 epochs. All deep neural networks in this paper are trained with the Titan RTX 24GB GPU. We use the Adam [17] method for the optimization. The learning rate for Sen-Model, Abs-Model and Seg-Model are 0.001, 0.003 and 0.001, respectively. A reduce learning rate scheme by a factor of 0.1 is set during training. The Bi-RNN decoder at Abs-Model uses Bi-LSTM for PudMed dataset and Bi-GRU for NICTA-PIBOSO dataset, respectively. The hyperparameters λ_{abs} and λ_{seg} are empirically set to 1 and 0.2, respectively. The two 2D-convolution layers in the C-RNN has the same padding with kernel size and number of filters set to $(8, 3)$ and 16, respectively. The hidden states dimension d_h of all LSTM layers in Sen-Model is set to 128.

D. Experimental results

We first evaluate our proposed models at the sentence level with different input features: using only word sequence (Sen-Model w/ word only); using both word and character sequences (Sen-Model w/ word & char); using word, character, and statistics (Sen-Model w/ word & char & stat); using only sentence embeddings extracted from the pre-trained PudMed model (Sen-Model w/ sentence only); using all input features of word, character, statistics, sentence embeddings (Sen-Model w/ sentence & word & char & stat). The experimental results shown in Table III highlight that each input feature helps to further improve the performance of the Sen-Model at the sentence level. The best performance at the sentence level is from the combination of all input features of word, character, statistics, sentence embeddings (Sen-Model w/ sentence & word & char & stat), presenting the F1/Precision/Recall scores of 91.1/91.9/90.9 and 81.8/81.8/81.8 on PubMed 20K and NICTA-PIBOSO datasets, respectively. This combination outperforms the model that uses only pre-trained sentence embeddings from the BERT model (Sen-Model w/ pre-trained sentence only), which scores 87.0, 87.1, and 87.0 on PubMed 20K, and 78.5, 78.8, and 78.5 on NICTA-PIBOSO. These results demonstrate the effectiveness of the proposed LSTM-based network in generating high-quality sentence representations. This is achieved by combining task-specific features based on words, characters, and statistical information within a specific context, along with pre-trained embeddings from the BERT model, which has a comprehensive understanding of medical domain language from large-scale medical corpora. In other words, the proposed LSTM-based network effectively integrates synergistic and diverse features, allowing model to consider both the overarching medical knowledge and the specific details of each sentence, resulting in superior sentence representations.

Leveraging sentence embeddings from the proposed Sen-model, the model at the abstract level (Abs-Model) further enhances the system performance. We achieve the F1/Precision/Recall scores of 92.7/93.2/92.6 on PubMed 20K and 84.6/85.5/84.1 on NICTA-PIBOSO datasets as shown in the lower part of Table III. The model at the segment level (Seg-Model), when being integrated into the system, shows efficiency in considering coherent dependencies of sentences in local regions within segments and recorrect sentences at the boundary of two label classes. The best system (Combine-Model), which combines of Abs-Model and Seg-Model, achieves the best result of 92.8/93.4/92.7 on PudMed

20K and 85.3/86.5/84.5 on NICTA-PIBOSO as shown in Table IV. This model also outperforms the baseline [9] by 1.0%, 2.8%, and 2.6% on PubMed 200K, PubMed 20K, and NICTA-PIBOSO datasets in terms of F1 scores, respectively.

Compared with the state-of-the-art systems as shown in Table V and Table VI, although our best system presents fundamental network architectures at the abstract level and the segment level when leveraging the well-presented sentence embeddings at the sentence level, we achieve very competitive results (top-4 on PubMed 20K and top-3 on NICTA-PIBOSO). This indicates that the role of the LSTM-based network (Sen-Model) at the sentence level is important to achieve comprehensive sentence representation, which can be effectively set as an initial foundation and leveraged in higher levels of segment level and abstract level to improve the model performance. Therefore, our future work is to investigate novel methods for further improving the model performance based on the well-presented sentence representation.

IV. Conclusion

This paper has presented a deep learning system for the Sequential Sentence Classification (SSC) task in medical scientific abstracts based on the motivation of improving sentence representation. By conducting extensive experiments, we achieved the best system that outperforms the baseline by 1.0%, 2.8%, and 2.6% on the benchmark datasets of PubMed 200K RCT, PubMed 20K RCT, and NICTA-PIBOSO regarding F1 scores, respectively. The results are highly competitive to the state-of-the-art systems on these two datasets. Particularly, our proposed LSTM-based network at the sentence level proves a vital role in generating comprehensive sentence representation, which can be served as a strong foundation for further exploring and improving the performance of the SSC task on higher contextual levels.

Acknowledgment

The work described in this paper is performed in the H2020 project STARLIGHT ("Sustainable Autonomy and Resilience for LEAs using AI against High Priority Threats"). This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 101021797.

References

[1] A. Rai *et al.*, "Query specific focused summarization of biomedical journal articles," in *2021 16th Conference on Computer Science and Intelligence Systems (FedCSIS)*, 2021, pp. 91–100, doi:10.15439/2021F128.

[2] H. S. Nguyen *et al.*, "Semantic explorative evaluation of document clustering algorithms," in *2013 Federated Conference on Computer Science and Information Systems*, 2013, pp. 115–122.

[3] F. Dernoncourt and J. Y. Lee, "PubMed 200k RCT: a dataset for sequential sentence classification in medical abstracts," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, Nov. 2017, pp. 308–313.

[4] I. Amini, D. Martinez, and D. Molla, "Overview of the ALTA shared task," in *Proceedings of the Australasian Language Technology Association Workshop 2012*, 2012, pp. 124–129.

[5] X. Shang, Q. Ma, Z. Lin, J. Yan, and Z. Chen, "A span-based dynamic local attention model for sequential sentence classification," in *Proc. ACL-IJCNLP*, 2021, pp. 198–203, doi: 10.18653/v1/2021.acl-short.26.

[6] J. Lin, D. Karakos, D. Demner-Fushman, and S. Khudanpur, "Generative content models for structural analysis of medical abstracts," in *Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology*, 2006, pp. 65–72.

[7] P. Ruch, C. Boyer *et al.*, "Using argumentation to extract key sentences from biomedical abstracts," *International Journal of Medical Informatics*, vol. 76, no. 2, pp. 195–200, 2007, doi: https://doi.org/10.1016/j.ijmedinf.2006.05.002.

[8] S. N. Kim *et al.*, "Automatic classification of sentences to support evidence based medicine," in *BMC bioinformatics*, vol. 12, no. 2, 2011, pp. 1–10, doi: https://doi.org/10.1186/1471-2105-12-S2-S5.

[9] F. Dernoncourt, J. Y. Lee, and P. Szolovits, "Neural networks for joint sentence classification in medical paper abstracts," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 2017, pp. 694–700.

[10] D. Jin and P. Szolovits, "Hierarchical neural networks for sequential sentence classification in medical scientific abstracts," in *Proc. EMNLP*, 2018, pp. 3100–3109, doi: 10.18653/v1/D18-1349.

[11] K. Yamada, T. Hirao, R. Sasano, K. Takeda, and M. Nagata, "Sequential span classification with neural semi-Markov CRFs for biomedical abstracts," in *Findings of the Association for Computational Linguistics: EMNLP*, 2020, pp. 871–877, doi: 10.18653/v1/2020.findings-emnlp.77.

[12] Y. Peng, S. Yan, and Z. Lu, "Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets," in *Proceedings of the 18th BioNLP Workshop and Shared Task*, 2019, pp. 58–65, doi: 10.18653/v1/W19-5006.

[13] Devlin *et al.*, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018, doi: 10.18653/V1/N19-1423.

[14] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *in Proc. EMNLP*, 2014, pp. 1532–1543, doi: 10.3115/v1/D14-1162.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023.

[16] Y. Gu *et al.*, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Transactions on Computing for Healthcare (HEALTH)*, vol. 3, no. 1, pp. 1–23, 2021, doi: 10.1145/3458754.

[17] P. K. Diederik and B. Jimmy, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2015.

[18] A. Brack *et al.*, "Sequential sentence classification in research papers using cross-domain multi-task learning," *International Journal on Digital Libraries*, pp. 1–24, 2024, doi: https://doi.org/10.1007/s00799-023-00392-z.

[19] Cohan *et al.*, "Pretrained language models for sequential sentence classification," in *Proc. EMNLP-IJCNLP*, 2019, pp. 3693–3699, doi: 10.18653/v1/D19-1383.

[20] G. Yu, Z. Zhang, H. Liu, and L. Ding, "Masked sentence model based on bert for move recognition in medical scientific abstracts," *Journal of Data and Information Science*, vol. 4, no. 4, pp. 42–55, 2019, doi: 10.2478/jdis-2019-0020.

[21] S. Gonçalves, P. Cortez, and S. Moro, "A deep learning classifier for sentence classification in biomedical and computer science abstracts," *Neural Comput. Appl.*, vol. 32, no. 11, p. 6793–6807, 2020, doi: 10.1007/s00521-019-04334-2.

[22] A. Agibetov, K. Blagec, H. Xu, and M. Samwald, "Fast and scalable neural embedding models for biomedical sentence classification," *BMC bioinformatics*, vol. 19, pp. 1–9, 2018, doi: https://doi.org/10.1186/s12859-018-2496-4.

[23] A. Sarker, D. Mollá, and C. Paris, "An approach for automatic multi-label classification of medical sentences," in *Proceedings of the 4th International Louhi Workshop on Health Document Text Mining and Information Analysis. Sydney, NSW, Australia*, 2013.

[24] M. Lui, "Feature stacking for sentence classification in evidence-based medicine," in *Proceedings of the Australasian Language Technology Association Workshop*, 2012, pp. 134–138.

[25] D. Mollá, "Overview of the alta shared task: Piboso sentence classification, 10 years later," in *Proceedings of the The 20th Annual Workshop of the Australasian Language Technology Association*, 2022, pp. 178–182.