

Domain Classification of Biomedical Research Articles based on BiLSTM for Recommendation System

Sahaya Chithra.E¹, Dr. S. John Peter²

¹Research Scholar, Manonmaniam Sundaranar University, Tirunelveli, India.

²Associate Professor, Department Of Computer Science, St. Xavier's College, Tirunelveli, India.

ABSTRACT

Nowadays, Researchers and Scientists have plenty of resources for Research papers research repository available on the web. When the information amount increased day by day the researchers faced a new problem that they confused for selecting most relevant research paper actually they want to access. PubMed, for example, has an enormous collection of medical related Research papers. The Recommendation system helps researchers to retrieve the relevant paper and keeping the track of their research field. To improve the better response and achieving an accurate Recommendation system this proposed Domain Classification system is to classify similar domain of articles. This study focuses to extract the relevant domain from PubMed database by proposes a multilayered Recurrent Neural Network (RNN) model based on Bidirectional Long Short-Term Memory (BiLSTM). Experimental result shows RNN based BiLSTM on PubMed domain classification outperforms the traditional classifications for RS.

Keywords:

Recommendation system (RS), Internet, Domain Classification, Recurrent Neural Network (RNN), Bidirectional Long Short-Term Memory (BiLSTM).

Introduction

The growing amount of literature reviews lead to information overload problem for new researchers. The Recommendation System (RS) for research articles is an essential process that supports researchers in keeping track of their field in research. The initial process of the Recommendation systems (RS) is to classify matching products or services, such as books, music, movie based on implicit interests about the user, or the recommended products. Personalization system estimates users' preferences on research articles and facilitating a better user experience that researchers might search. For individuals, after this proposed study of Domain Classification, RS allows users to make more effective for selecting the relevant Research article [3].

Usually, three types of Recommendation Systems can be dealt with by Researchers such as collaborative filtering (CF) based methods, and hybrid methods. User profiles that are registered by the user and the product descriptions from the website are the key elements of the Content-based method. A. V. D. Oord [4] proposed the Content-based Recommendation system for Music. The Collaborative Filtering methods [5, 6, and 7] use the past activities expressed by the user or preferences, such as user ratings on items, and feedback reviews without using user's or product's content information. Hybrid methods [8, 9, 10] can be applied by combining content-based and Collaborating Filtering-based methods.

In the recent Internet era, some promising results of Domain classifications have been achieved by the use of Deep Learning techniques. Domain classification for Recommendations could be enriched by Deep learning approaches. Here the user experiences and their interests are improved

to attain better Recommendations. Now the exceptional performance of the Deep Learning algorithms in treating the application fields like speech and face recognition, object detection, and natural language processing proved by world-famous organizations such as Google, Facebook, and Microsoft.

The proposed model can classify similar Domains of PubMed articles for the Recommendation system. The rest of this paper is organized as follows: Section 2 discussed the related work of why RS needs classification. Section 3 is reserved for the Deep Learning-based approaches which are impacted by the LSTM and BiLSTM. This classification is followed by the identification of the new challenges in the Deep learning-based Recommendation System in the future. In section 4, the preprocessed model is trained and compared the Result with LSTM and BiLSTM.

Related Work

Domain classification play a crucial role in NLP and Information Extraction research. In an attempt to address the Recommendation System, the Researcher must begin with Domain Classification. Evan Cox and Marcelo Worsley [11] proposed a multi domain text classification system. This study has taken random of five thousand product reviews for analysis. Among these samples the model were trained by Support Vector Machine and Naïve Bayes.

Kyo-Joong Oh et al. [12] proposed an out-of-domain detection in dialogue system. This study detects the OOD utterances from a dialogue and made a classification of sentences or utterances as in-domain (ID) and OOD. This study used the sentence embedding vector with domain features for dialogue domain classification. Ryu et al. [13] used an LSTM network trained for domain classification of neural sentence embedding system for processing unknown words in the OOD problem. This study used the learned sentence representations to train an auto-encoder that detects OOD sentences based on their reconstruction errors.

ChhayaChoudhary et al. [14] proposed a system is to classify domains and categorizing the Domain Generation Algorithms generated domains according to their family. This research trained a model of malware detection family with Random forest to extract features from domain names.

Han Guo [15] presented a study of multiple domain distance measures to solve domain adaption problem. FangliRen [16] presented deep learning framework ATT-CNN-BiLSTM for identifying and detecting DGA domains to alleviate the threat. This study used Convolutional Neural Network (CNN) and bidirectional Long Short-Term Memory (BiLSTM) to extract the features of the domain sequences information.

Materials and Methods

3.1 Recurrent Neural Network

Deep Learning, a subset of machine learning is a mimic for the human brain. A Simple neural network cannot be used in sequence problems. RNN can be implemented through multi-layered neural networks. All sequence models are dealt with RNN to solve NLP problems such as autocompletion of words, sequence translation. In the field of NLP, for vectorizing a word in

sequence model RNN is used to make computations quickly. Now, plentiful recommendation models are raising with the help of RNN. [23].

The traditional Recommendation system avoids the time complexity completely. In feed-forward networks, inputs are independent of each other. With the new exposure of deep learning, all inputs are connected to each other in [Recurrent Neural Networks](#) (RNNs). The RNN was first developed in the 1980s [17–19]. RNN can be thought of as multiple copies of the same network, each passing a message to a successor.

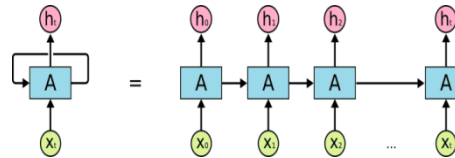


Figure.1. RNN Loop [20]

As it can be seen in the figure 1, x_0 is the input and h_0 is the output of this sequence and then h_0 and x_0 are the input of next step of this sequence. Similarly, h_1 and x_1 are the input of next step and so on. In the last few years, there has been incredible success when applying RNNs to a variety of sequence to sequence translation problems.

3.2 LSTM (Long Short-Term Memory)

LSTM ([Hochreiter and Schmidhuber, 1997](#)) is a special variant of the RNN which solves short-term memory problems and handling long-term dependencies. According to Olah [21], LSTM is designed for sequence problems and has a hidden layer which is called short-term memory.

LSTM is a chain-like structure but the repeating module has a different structure. LSTM cell state which handles long-term memory is the key feature of LSTM. This is represented by the horizontal line running through the top of the diagram. The information that has not needed is thrown away from the LSTM memory is decided initially. This decision is made by the sigmoid function in the forget gate. The sigmoid layer outputs numbers between zero and one. A value one means “let everything through” while a value zero means “let nothing thorough”. Now a tan layer creates a vector of new values that could be added to the cell state. Then these two are combined and an updated cell state is created. Finally, a filtered output will be based on the cell state is generated.

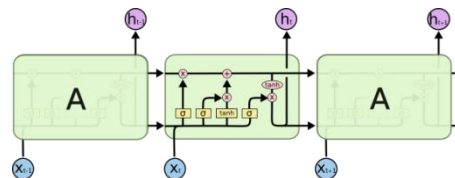


Figure. 2. LSTM cell unit [21]

The LSTM cell can compute current hidden state $\mathbf{h}_t$ based on current vector $\mathbf{x}_t$, earliest hidden state $\mathbf{h}_{t-1}$ and previous cell state $\mathbf{c}_{t-1}$. The operation of input gate i_t , forget gate f_t , output gate o_t and memory cell state $\mathbf{c}_t$ are defined as:

$$i_t = \delta(W_{xi} x_t + W_{hi} h_{t-1} + W_{ci} c_{t-1} + b_i) \quad (1)$$

$$f_t = \delta(W_{xf} x_t + W_{hf} h_{t-1} + W_{cf} c_{t-1} + b_f) \quad (2)$$

$$c_t = \tanh(W_{xc} x_t + W_{hc} h_{t-1} + b_c) \quad (3)$$

$$c_t = f_t * c_{t-1} + i_t * c_t \quad (4)$$

$$i_t = \delta(W_{xt} x_t + W_{ht} h_{t-1} + W_{ct} c_{t-1} + b_t) \quad (5)$$

$$o_t = \delta(W_{xo} x_t + W_{ho} h_{t-1} + W_{co} c_{t-1} + b_o) \quad (6)$$

$$h_t = o_t * \tanh(c_t) \quad (7)$$

where δ is the element-wise sigmoid function, $*$ is the element-wise product, $\mathbf{W}(\cdot)$ are the weight metrics, and $b(\cdot)$ are the biases. For each vector $\mathbf{x}_t$, the hidden state $\mathbf{h}_t$ preserves or removes epast information.

3.3 BiLSTM (Bidirectional Long Short-Term Memory Unit)

Bidirectional LSTMs are an extension of traditional LSTMs that can be used to improve the performance of sequence classification problems. The bidirectional LSTM (BiLSTM) architecture (Gers et al., 2000) consists of two LSTM. One LSTM is used to capture inputs from the forward direction while another LSTM is used to get inputs from the backward direction.

The Bidirectional wrapper is used with an LSTM layer, so input propagates left to right as well as the right to left in time and then concatenates the outputs [22]. One LSTM layer propagates past input to forward direction while another LSTM propagates the future input to the backward direction. BiLSTM structure provides two additional training by traversing in both direction and this additional training in the layers lead to parameter better tuning and finally outperforms than the single LSTM [14].

Methodology for Domain Classification

4.1 Dataset

The output of this component is the preprocessed data, which is collected from PubMed Database. PubMed is free digital repository which contains millions of biomedical scholarly citations. This study randomly selects title and abstract of similar Research papers on different categories. The PubMed API is called the [Entrez Database](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&retmode=json&retmax=20&sort=relevance&term=cold). If the researchers want to search for the term 'cold', the following URL is used. <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&retmode=json&retmax=20&sort=relevance&term=cold>

Research papers were retrieved using the above link and downloading all the research paper that had relevance according to the category. For Domain classification process similarity between papers are analyzed. This process needs abstract and title only not full text of Paper. Selecting the entire text of Paper leads to slacking off the process.

4.2 Pre-processing

In Pre-processing, all punctuation marks, apostrophes, exclamation marks, question marks, hyphens, periods are removed and, turning the sentences into space-separated sequences of words. Stop words are removed by using the NLTK library. Lemmatization uses lexical knowledge bases to get the correct base forms of words that are converting them into root words. Finally, the tokens are lowercased, stop words removed, header-footer removed, and lemmatized. After tokenization, the next step is to turn those tokens into lists of sequences. Here the sequences

to be in the same size, so padding is important for text classification. The Padding process truncates each sentence into a given standard size.

4.3 Word Embedding

Word embedding provides a dense representation of words and their relative meanings. This study uses the Global Vectors for Word Representation (**GloVe**) algorithm for word embedding. Lemmatization takes longer than stemming but effectively reduces the sizes of the bag-of-words matrix. Both reduce dimensions though. Lemmatizing a word can help infer more useful vectors when using GloVe. GloVe has embedding vector sizes fifty, hundred, two hundred and three hundred dimensions. Here this study chose the hundred dimensional one. Usually the size of the words would grow larger and larger as a result, sparse matrix arose. One solution is to fix a standard length of hundred for all instances **truncating** longer ones while shorter ones are **padded**.

4.4 Layered Structure of BiLSTM

Total 400000-word vectors in Glove 6B 100d.

Model fitting - simplified convolutional neural network

Table 1. Layered output using BiLSTM

Layer (type)	Output Shape	Param#
Input_2 (Input Layer)	(None, 1000)	0
Embedding_2 (Embedding)	(None, 1000, 100)	984800
Bidirectional_1(Bidirectional)	(None, 100)	60400
Dense_3(Dense)	(None,128)	12928
Dense_4 (Dense)	(None, 5)	645

Total params: 1,058,773

Trainable params: 1,058,773

4.5 A Proposed Biomedical Articles Domain Classification Work flow

1. Receive set of N titles and abstracts $A = \{a_1, a_2, \dots, a_N\}$
2. Preprocess documents $A = \text{preprocess}(\{a_1, a_2, \dots, a_N\})$ using stemming to removing English stop words.
3. Apply GloVe word embedding for word representation.
4. Apply BiLSTM layered structure for sequential classification of research paper's domain.

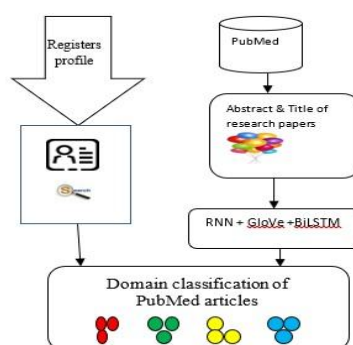


Figure 4. Proposed Flow diagram of Biomedical articles Domain Classification

4.6 Experimental Results

This study discusses the performance metrics like Accuracy, Precision, Recall and FI score to monitor and measure the performance of model. The formula to calculate Precision, recall and F1score are defined by equation (8), (9), and (10).

$$\text{precision} = \frac{TP}{(TP+FP)} \quad (8)$$

$$\text{recall} = \frac{TP}{(TP+FN)} \quad (9)$$

$$\text{F1score} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad (10)$$

Using these results, finally the accuracy will be calculated by the formula (11).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (11)$$

Where TP is True Positive, TN is True Negative, FP is False Positive and FN is False Negative. The Classified metrics results of LSTM and BiLSTM representations are compared in Table 1.

Table 2. Metrics comparative results of LSTM and BiLSTM

	LSTM	BiLSTM
Accuracy	0.72	0.82
Precision	0.772131	0.828054
Recall	0.745158	0.818785
F1Score	0.709372	0.815864

Based upon the above results the graph is constructed in Figure 4 to analyze, LSTM and BiLSTM results which shows BiLSTM model performs 82% accuracy, 83% Precision, 82% Recall and 82% FI Score than LSTM model for achieving the most promising Domain classification system.

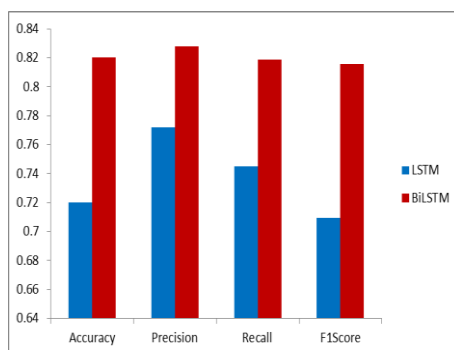


Figure 4. Metrics Results w.r.t LSTM and BiLSTM

Conclusion

This study proposed similar Biomedical Research Articles from the PubMed database based on RNN with BiLSTM and GloVe. The experimental results obtained by this study are suitable in classifying the domains which are relevant to the researcher's interest. A key feature of this system is the easy handling of the most promising biomedical domains. That is, a researcher only has to give preferred keywords as input into this system, then by using the researcher's preferred keywords preferences, this system classifies similar domains with the help of BiLSTM and GloVe. Thus, this system will help to get the most relevant biomedical domain area to enrich the effectiveness of the Recommendation System.

References

- [1] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay," Deep Learning based Recommender System: A Survey and New Perspectives". ACM Computing Surveys, Vol. 1, No. 1, Article 1. July 2018. DOI: 0000001.0000001.
- [2] Joeran Beel¹, Bela Gipp², Stefan Langer¹, and Corinna Breiting¹," Research-Paper Recommender Systems: A Literature Survey ".
- [3] Rim Fakhfakh ,Anis Ben Ammar , Chokri Ben Amar , "Deep Learning-based Recommendation: Current Issues and Challenges".
- [4] Oord, A.V., Dieleman, S., &Schrauwen, B."Deep content-based music recommendation." NIPS (2013).
- [5] Hao Wang, Naiyan Wang, Dit-Yan Yeung," Collaborative Deep Learning for Recommender Systems".
- [6] Jian Wei, Jianhua He, Kai Chen, Yi Zhou, Zuoyin Tang," Collaborative Filtering and Deep Learning Based Recommendation System for Cold Start Items".
- [7] KrutiJani, Dr. V.M. Chavda, "Personalizing Movie Recommendation Using Semantic Contents in collaborative Filtering", GJRA - Global Journal for Research Analysis, Page no 103.
- [8] ZhenghuaXu, Thomas Lukasiewicz, Cheng Chen, Yishu Miao, XiangwuMeng," Tag-Aware Personalized Recommendation Using a Hybrid Deep Model", Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17).
- [9] Shuai Zhang, Lina Yao, and XiweiXu,"AutoSVD++: An Efficient Hybrid Collaborative Filtering Model via Contractive Autoencoder",ACM. 978-1-4503-5022-8/17/08 (2017), DOI: <http://dx.doi.org/10.1145/3077136.3080689>
- [10] SahayaChithra. E, Dr. S. John Peter,"A Literature Review and Classification of Semantic Web Approaches for Web Personalization Research", IRJET, Volume: 06 Issue: 10 | Oct 2019, Page 307.
- [11] Evan Cox, Marcelo Worsley," In Pursuit of an Efficient Multi-Domain Text Classification Algorithm".
- [12] Kyo-Joong Oh, DongKun Lee, "Out-of-Domain Detection Method Based on Sentence Distance for Dialogue Systems", 2375-9356 /2018 IEEE DOI 10.1109/BigComp.2018.00123.
- [13] S. Ryu, S. Kim, J. Choi, H. Yu, G. Lee, "Neural Sentence Embedding using Only In-domain Sentences for Out-of-domain Sentence Detection in Dialog Systems", arXiv:1807.11567v1.
- [14] S. Siarni-Namini, N. Tavakoli and A. S. Namin, "The Performance of LSTM and BiLSTM in Forecasting Time Series," 2019 IEEE International Conference on Big Data (Big Data), Los

- Angeles, CA, USA, 2019, pp. 3285-3292, doi: 10.1109/BigData47090.2019.9005997.
- [15] Han GuoRamakanthPasunuruMohitBansal,” Multi-Source Domain Adaptation for Text Classification via DistanceNet-Bandits”, The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20).
 - [16] FangliRen, Zhengwei Jiang, XurenWang and Jian Liu1,” A DGA domain names detection modeling method based on integrating an attention mechanism and deep neural network”, Cybersecur 3, 4 (2020), <https://doi.org/10.1186/s42400-020-00046-6>.
 - [17] Rumelhart, D., Hinton, G. & Williams, R. Learning representations by back-propagating errors. Nature 323, 533–536 (1986). <https://doi.org/10.1038/323533a0>
 - [18] Werbos, P., “Generalization of backpropagation with application to a recurrent gas market model.” Neural Networks 1 (1988): 339-356.
 - [19] JeffreyL.Elmán,” Finding Structure in Time “, Cognitive Science 14, 179-211 (1990)
 - [20] Olah, C.” Understanding LSTM Networks.”, Available online: <http://colah.github.io/posts/2015-08-Understanding-LSTMs/>
 - [21] Yan, S.,” Understanding LSTM and Its Diagrams.”, Available online: <https://medium.com/mlreview/understanding-lstm-and-its-diagrams-37e2f46f1714>.
 - [22] RaghavAggarwal, Bi-LSTM. Available online: <https://medium.com/@raghavaggarwal0089/bi-lstm-bc3d68da8bd0>.
 - [23] SahayaChithra.E, Dr. S. John Peter,”Deep Learning In Personalized Recommendation System – A Comparative Study”, Studies in Indian Place Names, ISSN: 2394-3114 ,Vol-40,Issue - 70,March -2020.