Deep Sequence Labelling Model for Information Extraction in Micro Learning Service

Jiayin Lin[1], Zhexuan Zhou[1], Geng Sun[1], Jun Shen[1,2], David Pritchard[2],Tingru Cui[3], Dongming Xu[4], Li Li[5],Ghassan Beydoun[6]
[1]School of Computing and Information Techonology,University of Wollongong, Wollongong, Australia
[2]Research Lab of Electronics, Massachusetts Institute of Technology, Cambridge, MA
[3]University of Melbourne, Melbourne, Australia
[4]UQ Business School, The University of Queensland, Brisbane, Australia
[5]Faculty of Computer and Information Science, Southwest University, Chongqing, China
[6]School of Information,System and Modelling, University of Technology Sydney, Sydney, Australia
{jl461, zz827}@uowmail.edu.au, {gsun, jshen}@uow.edu.au, dpritch@mit.ed, tingru.cui@unimelb.edu.au, d.xu@business.uq.edu.au, lily@swu.edu.cn, ghssan.beydoune@uts.edu.au

***Abstract*—Micro learning aims to assist users in making good use of smaller chunks of spare time and provides an effective online learning service. However, to provide such personalized online services on the Web, a number of information overload challenges persist. Effectively and precisely mining and extracting valuable information from massive and redundant information is a significant pre-processing procedure for personalizing online services. In this study, we propose a deep sequence labelling model for locating, extracting, and classifying key information for micro learning services. The proposed model is general and combines the advantages of different types of classical neural network. Early evidence shows that it has satisfactory performance compared to conventional information extraction methods such as conditional random field and bi-directional recurrent neural network, for micro learning services.**

Keywords—information extraction, sequence labelling, micro learning, neural network, machine learning

I. INTRODUCTION

Micro learning is a novel e-learning style, which aims to provide effective online learning activities within fragmented time frames. Learning activities in this paradigm are typically completed within fifteen minutes through the Internet. Micro learning has attracted extensive attention from researchers in both education and computer science areas. As a form of the Web service, micro learning targets massive online users at different ages, from different locations, and with different learning demands. Depending on the learners' learning purpose and the credit they would obtain after the completion of a full course, this learning service could be formal, informal, or non-formal [1]. Hence, the volume of learning materials involved in this service and information generated associated with them would also conform with the standard definition of 'big data' [2]. However, operating and maintaining such online service poses the challenges in effectively managing dynamic and massive information from both user-side and resource-side. In addition, sometimes, for a particular single application scenario, a large percentage of information is redundant or useless. Hence, to effectively analyse information stream and precisely extract valuable information becomes a significant and necessary pre-processing stage for the micro learning service. Such pre-processing stage can not only save cost for learners, release the heavy workload but also improve the quality of the online micro learning service.

In a broader sense, information extraction refers to the task of automatically analysing, locating, distilling, summarizing, and extracting useful information from massive unstructured multimedia documents. The workflow of the micro learning service can be realized via three important modules: non-micro learning material segmentation, learning material annotation and learning material recommendation [2]. Based on the workflow of micro learning, the utility of information extraction technique could play a vital role in the pre-processing stage of each intelligent module mentioned above. Extracting relevant information about boundaries between adjacent knowledge points guarantees the accuracy of segmentation results; extracting keywords from learning resources makes the annotation results more descriptive; and extracting latent information such as users' knowledge level and the difficulty level of learning resource makes recommendation results more personalized.

In this study, we are dedicated to using sequence labelling model to automatically analyse the content of the information stream and then identify, locate, and classify the valuable information for the micro learning service. The relationship between sequence labelling strategy and information extraction task will be discussed in the following section. We herein propose a deep sequence labelling model (namely, deep Bi-LSTM-CNNs-CRF) for information extraction. This model tries to insightfully depict different aspects of the online micro learning scenario, summarises them together, and extracts valuable information for assisting the follow-up different intelligent processing modules of the online learning service.

The remainder of this paper is organized as follows. Related works about information extraction and micro learning will be firstly introduced and discussed in Section II. Then in section III, the architecture of the proposed deep sequence labelling model will be presented. The functionalities of each significant component of this model will be described in detail in section III as well. The relevant

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

experiments, evaluation metrics, involved datasets, and test results of this study will be analysed in Section IV. We conclude this paper and discuss future research in Section V.

II. Related Work

A. Information Extraction, Sequence Labelling Problem, and Micro learning

For an online learning service, sequence information refers to any chronologically ordered information. This could be historical learning records of an online learner, or it could simply be the metadata of learning resource with chronological characters such as text information.

In one previous work about cloud-based micro learning system [3], the authors pointed out that keyword extraction was important to learning resource modelling, and real-time data extraction and inferencing were vital for constructing personalized learner model. In [4], the authors argued that extracting certain types of information like a user's profile was essential to solve the cold-start problem for recommending online educational resource. According to [5], for a learning service, intelligent components such as entity extraction, relationship extraction, and resources disambiguation are based on the successful extraction of useful information from massive information stream.

Intuitively, the process of an information extraction task can be roughly divided into two individual steps: locating the valuable information from a multimedia document and then classifying the located information into the predefined categories. However, to construct two separate models for each processing step will make the whole processing procedure error-prone. To design a model which is capable of dealing with two above-mentioned sub-tasks at the same time is vital to maintain the robustness of the extraction procedure and effectiveness of the personalized service. In micro learning, this even poses more difficulty to researchers and practitioners.

Hence, to solve the information extraction problem in micro learning requires an end-to-end solution, and very few existing studies can be found in the literature. Such information extraction solution takes the input of the raw information and outputs the distilled valuable information, with no intermediate output or operation being involved. Moreover, the end-to-end model does not require much feature engineering tasks and domain knowledge [6, 7], which makes it yielding satisfactory generalization performance. Many prior studies suggested using sequence labelling model to solve the information extraction problem. In a sequence labelling model for micro learning, the information stream is fed into the model in chronological order, and the output is the sequence of relevant labels for data elements, each label can indicate the usefulness and category of each data element at the same time. To this end, several previous representative studies about sequence labelling will be discussed in the following subsections.

B. Hidden Markov Model and Conditional Random Field

Hidden Markov Model (HMM), Conditional Random Field (CRF) and their variants have been widely used in many studies for modelling various sequential and temporal problems [8-10]. These models perform well in modelling local characteristics and constraints. However, such modelling strategies are based on linear information and heavily relying on handcraft features, which make them lacking abilities to model complex tagging problems in a more general case, such as online non-formal learning.

C. Recurrent Neural Network

Unlike other types of neural network, recurrent neural network (RNN) shows outstanding ability in modelling the temporal dynamic behaviour of the target problem. It can memorize historical information and combine such historical information with the currently received information, then make predictions. As a representative class of RNN, the Long Short-Term Memory (LSTM) shows satisfactory performance in modelling both short-term memory and long-term memory [11]. Moreover, bidirectional Long Short-Term Memory (Bi-LSTM) can further utilize 'future' input information for modelling sequential patterns [12]. Hence, it can not only mine forward sequence information but also mine backward sequence information at the same time. However, such type of model does not perform well in modelling local constraints, especially when adjacent outputs in a sequence have a strong influence on each other.

D. Combination of RNN and CRF

Based on the advantages and disadvantages of various sequential modelling strategies, LSTM-CRF and its variants are the state-of-art solutions in dealing with the sequence modelling problem. As pointed in [13], such network structure has the ability to efficiently use past input features via LSTM layer and sequence level local information via a CRF layer. With the advantages of Bi-LSTM, bidirectional LSTM-CRF (Bi-LSTM-CRF) models are used in many studies for sequence tagging or labelling task [7, 13, 14]; and in many cases, it outperforms the LSTM-CRF based model.

III. Model Design

In this section, we describe the design of the proposed deep sequence labelling model designed for our micro learning problem. Firstly, from a high-level perspective, we introduce the overall architecture of this model. Next, for each low-level vital component of this model, we discuss its characteristics in detail.

A. The Network Architecture

The proposed network contains four important layers and one block. The embedding layer is used for mapping the one- or multi-hot raw data to dense representations. The Bi-LSTM layer is used for modelling the temporal pattern of the embedded input sequence. The convolution neural network (CNN) layer is used for extracting adjacent latent features from the embedded input data. The CRF layer is used for adding extra local constraints, which are not captured by the previous layers. Moreover, the fusion block is used for combining different types of latent features. In the proposed network, the CNN layer and the Bi-LSTM layer model the embedded input separately. The input of this model is a sequence of vectors in the high dimension space. Each vector

represents the selected features of an individual raw micro learning resource. The outputs of the CNN layer and the Bi-LSTM layer are jointly fed into the fusion block, which contains several non-linear transformation layers for better information fusion. The general architecture of the proposed neural network is shown in Figure 1.

B. *Embedding Layer for Semantic Modelling and Dimension Reduction*

Embedding technique is an effective method for dimensionality reduction and feature representation, which transfers the points lying in a high dimensional space to a low dimensional one, while approximately preserving pairwise distances between points [15]. Such technique shows satisfactory performance in reducing the data and model complexity, and has been used in various machine learning-related tasks, such as information retrieval [16], multimedia data processing [17], and data mining [18]. For sequential signal processing, especially in natural language processing (NLP), the embedding layer also contributes to modelling the semantic information of raw input data [19]. In our proposed model, an embedding layer is used to map high-dimensional sparse raw data into low-dimensional continuous but dense one and extract the first-step semantic information.

C. *CNN Layer for Latent Feature Extraction*

Because of the competitive performance in extracting the latent feature, CNN is widely used not only in the research area of computer vision but in many other applications like NLP [20]. Many previous studies [7, 11, 21] proposed using CNN to extract and model character-level semantic information such as prefix and suffix.

However, not all types of information contain character-level semantic information. As a result, for constructing a more general model towards the information extraction task, after embedding layer, two continuous CNN layers are used to further mining and summarizing local features from adjacent inputs (object-level) in our proposed model. The utility of the CNN layers is quite different from the researches mentioned above. We may assume that such CNN layer can capture different types of information and boost the performance of Bi-LSTM-CRF. In computer vision, sliding window strategy is widely used for restricting the amount of information to be involved in each summarization process [22]; in our proposed model, we use a fix-size sliding window in each of the CNN layer. As shown in Figure 2, which is a single CNN layer for extracting latent features of adjacent embedded inputs, E_i is the i-th embedding vector generated from the previous embedding layer; the information from adjacent input embedding vectors is summarized. The rectangle is the sliding-widow moving from left to right, and at each iteration the information inside the window will be summarised.

D. *Bi-LSTM Layer for Sequence Labelling*

In our proposed model, one RNN layer is added as the core component after the embedding layer. This layer aims to extract and model chronological features. For a typical information extraction task, we may assume that both past and future inputs can provide valuable information in recognizing and locating information needing to be extracted. Hence, in our model, a Bi-LSTM structure is used in the RNN layer. The workflow of Bi-LSTM is shown in Figure 3. The embedded information is fed into the Bi-LSTM layer in both successive order and reverse order; then for each time step, the Bi-LSTM layer will output a prediction based on the 'past' and 'future' information of the current 'moment'.

E. *Fusion Block for Combining Different Types of Latent Feature*

As discussed in [23], fusing different types of features properly can make the information representation more reliable and more accurate. Inspired by this concept, a fusion block is added behind the CNN and Bi-LSTM components, and this block aims to better combine those different types of latent features. In the fusion block of the new model, latent features extracted from CNN layer and Bi-LSTM layer are firstly merged together by a concatenation operation. Then several non-linear transformation layers are used to further combine the latent features into fine-grained features. Such non-linear transformation layer could vary greatly respective to the problem domain and the complexity of the input data. In this study, we used another Bi-LSTM layer and fully connected neural network to model this non-linear

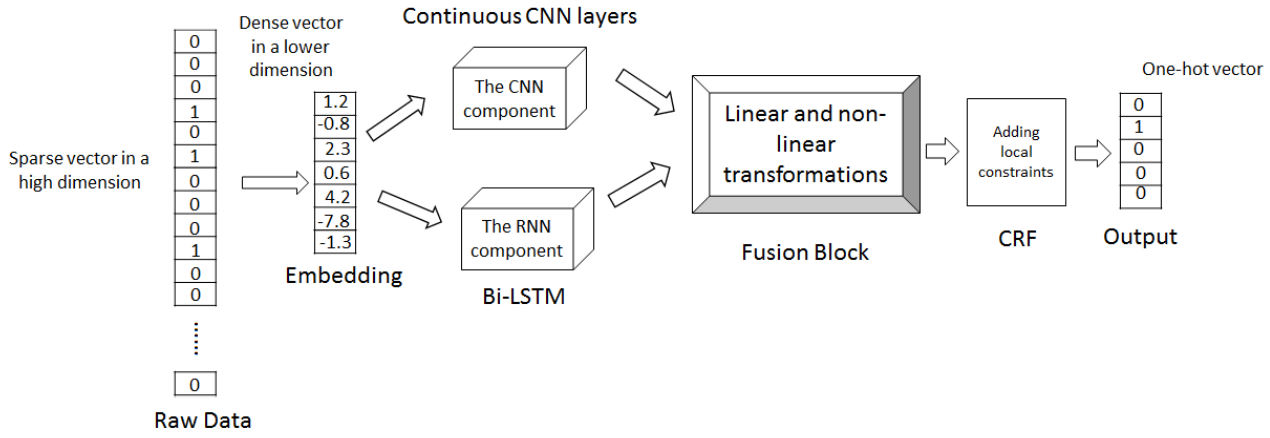

Figure 1. The Overall Network Structure of the Proposed Deep Bi-LSTM-CNNs-CRF Model

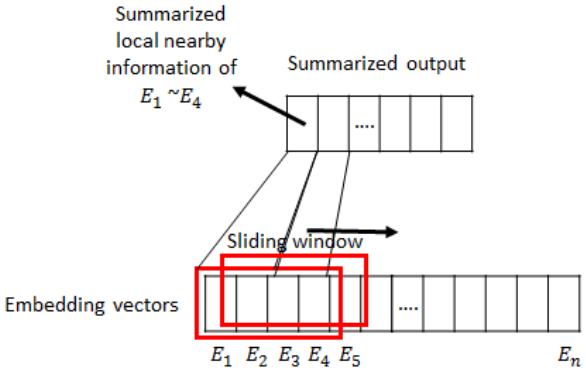

Figure 2. Convolutional Neural Network for Summarizing and Extracting Latent Features

transformation. The structure detail of this fusion block is shown in Figure 4.

F. CRF Layer for Adding Local Constrains to the Sequential Model

As discussed earlier, a pure RNN model has its own disadvantage in modelling local constraints. Hence, a CRF layer is used prior to the final output layer of the whole model.

In the CRF model, for a given sequence x, the probability of output y could be simply formulated as equation (1). From this equation, we can easily observe that y_i and y_{i-1} influences each other, which indicates that this probability value considers the correlation between outputs in neighbourhoods. The network structure of CRF is shown in Figure 5. Herein the prediction of second output Y_2 is not only determined by the second input X_2but also influenced by the first output Y_1. Function $t(Y_{i-1}, Y_i)$ and $s(Y_i, X_i)$ models the state transitions and emissions, respectively. This CRF layer aims to add more local constraints, especially the local constraints of the output sequence, which is not captured by former embedding, Bi-LSTM, and CNN layers [7].

$$p(\mathbf{y}|\mathbf{x}) = \frac{\prod_{i=1}^{n} \exp(y_{i-1}, y_i, \boldsymbol{x})}{\sum_{y \in y(x)} \prod_{i=1}^{n} \exp(y_{i-1}, y_i \boldsymbol{x})} \quad (1)$$

IV. Experiments

We have compared the performance of CRF, LSTM, Bi-LSTM, Bi-LSTM-CRF, our proposed deep Bi-LSTM-CNNs-CRF model with different experiment settings. We used three different evaluation metrics: precision, recall, and f1-score.

A. Evaluation Metrics

Precision is the number of true positive predictions divided by the total number of positive elements that predicted, which reflects how accurate the model is getting out of the predicted positives. The recall is the number of true positive predictions divided by the total number of positive elements in the data set, which reflects how many of the actual positives that model predicted through all the positives. F1-score considers both the precision value and the recall value, which can better reflect the model performance when the class distribution is uneven.

The receiver operating characteristic curve (ROC) is about plotting the true positive rate (TPR) against the false positive rate (FPR). The area under curve (AUC) can reflect the ability of how much a model can distinguish different information.

B. Datasets

The data used in the experiment comes from different fields of documents where we aim to facilitate the information extraction task of online learning service that covers a variety of disciplines. There are two separate datasets used in this study; one is labelled for training the model except for the embedding layer, which contains four different types of information that need to be extracted. Another dataset is unlabelled used for training the embedding layer, whereas the training is unsupervised. The unlabelled dataset is about a hundred times bigger than the labelled counterparts. Both datasets are encoded sequential of data and coming from the same source.

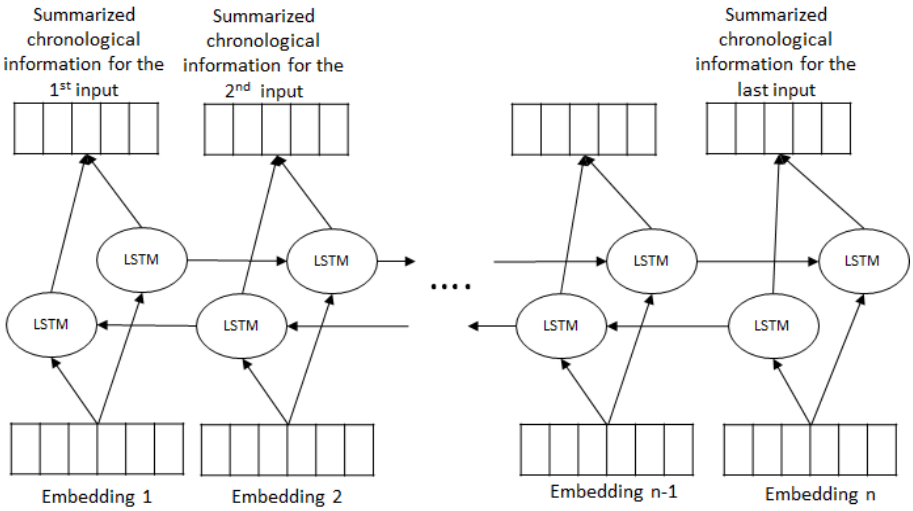

Figure 3. Bidirectional Long Short Memory for Modelling Sequential Pattern

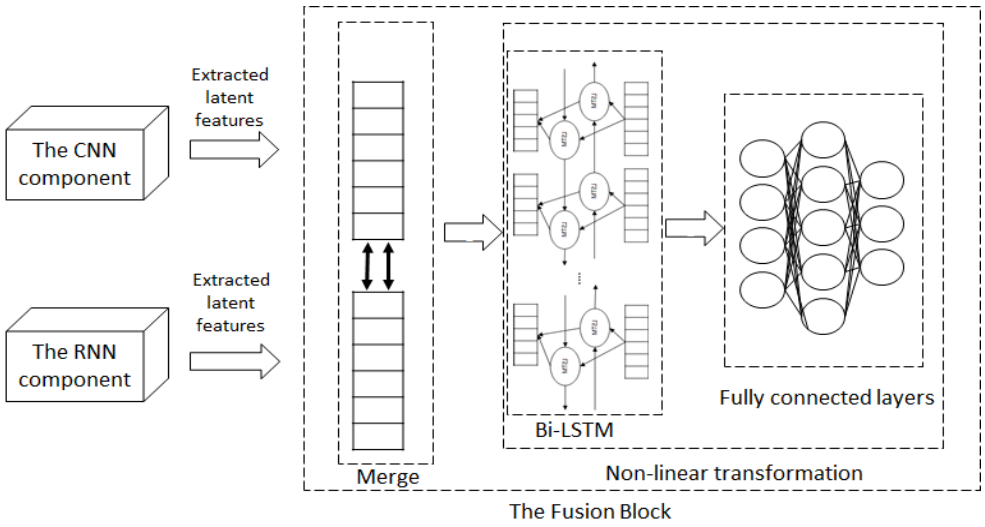

Figure 4. The Structure Detail of the Fusion Block

For a single case of a certain online learner, most of the online information is redundant. As a result, in this experiment, the information that needs to be extracted is very sparse. Also, in order to better reflect the real-world task where information with different significant levels or in different types could have totally different distributions, in the datasets, these four types of information are in different distributions. The statistical information about each type of information involved in the dataset is shown in Table 1.

Moreover, to maintain the generalization ability of our proposed deep Bi-LSTM-CNNs-CRF model, no handcraft domain-relevant feature is involved in this study. All the features that contain domain information are automatically selected, or extracted, or summarized by different layers of the neural network. There is no pre-trained model used in this study, even some of them are powerful and can greatly simplify the training process, such as BERT [24] and ELMo [25]. Our goal is to demonstrate that the proposed model is not restricted to any domain-relevant datasets or any domains. And how to effectively apply it to a specific online educational scenario or a problem domain is beyond the scope of this paper and will be considered as the future work.

C. Experimental Setup

In order to make these models comparable, we fixed the hyper-parameters of each neural network component during the training process of each model based on the pre-experiment results. The embedding technique used in this study is the classic unsupervised word2vec model [26], and the embedding dimension is 32. The output dimension of the first CNN layer is 128 and the second CNN layer is 64, and the sliding window size for both CNN layers is 5. The output dimensions of the Bi-LSTM outside and inside the fusion block is 128 and 64 respectively, and the maximum sequence length for modelling the chronological pattern is 128. All the neural network layers held 0.3 dropout rate, and we used default initialized settings for other parameters such as activation function and weight initialization method. Ten-fold cross-validation is used in the experiments.

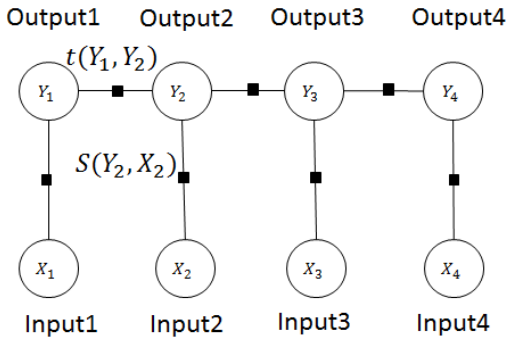

Figure 5. CRF Network [14]

TABLE 1 STATISTICAL INFORMATION ABOUT THE DATASET

	Information A	Information B	Information C	Useless Information	Total
Number	45,736	34,655	26,221	761,237	867,849
Percentage	5.26%	3.99%	3.02%	87.71%	100%

TABLE 2 EXPERIMENT RESULTS OF DIFFERENT MODELS

Metrics / Model	Recall	Precision	F1	# Results (TP+TN)
CRF	0.5164	0.3682	0.4299	5,238
LSTM	0.3016	0.1994	0.2401	4,858
BiLSTM	0.3231	0.2509	0.2824	5,704
BiLSTM-CRF	0.8168	0.7908	0.8089	7,113
Proposed model (without fusion block)	0.8083	0.7851	0.7951	7,136
Proposed model	**0.8223**	**0.8188**	**0.8206**	7,316
Proposed model (use softmax function to replace the CRF layer)	0.8052	0.7897	0.7974	7,205

D. Experiment Results and Discussions

Table 2 illustrates the most representative results obtained from the experiments. For better representing the performance of each model, despite the recall, precision, and f1-score, we also involved the total number of information that each model extracted containing true positive (TP) and true negative (TN) results. The total number of ground truth information needs to be extracted is 7,345.

1) The Significance of CRF Layer

Although the neural network has demonstrated its superb ability in modelling complex problems, based on our experiment results, we can easily find out that pure CRF model greatly outperforms pure RNN-based models (LSTM and Bi-LSTM). This result highlights the significance of the CRF layer, especially for the problems where the adjacent outputs have strong correlation or connections with each other.

In this study, we also compared the performance of the CRF method and the widely used softmax function, we keep all the settings of the model constant, only replacing the final CRF layer with the softmax function. From the bottom two rows of Table 2, we can observe that, for extracting information from the information sequence, CRF layer is greatly outperforming the softmax function.

2) Viterbi Algorithm and Cross-entropy

In this study, we also compared the model performance between using the Viterbi algorithm combined with negative log-likelihood and using cross-entropy for calculating the loss during the model optimization stage. Viterbi algorithm is mainly used for dynamically searching the best path for sequential predictions.

The performances of three representative models, pure CRF, Bi-LSTM-CRF, and the proposed model with different optimization strategies are shown in Table 3. The results demonstrate that using the Viterbi algorithm combined with negative log-likelihood for optimizing model greatly outperforms using cross-entropy. Simply using cross-entropy is prone to involving more truth negative predictions, which

TABLE 3 COMPARISON OF THE VITERBI ALGORITHM AND CROSS-ENTROPY

Metrics / Model	Recall	Precision	F1	# Results (TP+TN)
CRF (Viterbi)	0.5164	0.3682	0.4299	5,238
CRF (Cross-entropy)	0.2535	0.1999	0.2236	5,795
Bi-LSTM-CRF(Viterbi)	0.8168	0.7908	0.8089	7,113
Bi-LSTM-CRF(Cross-entropy)	0.7367	0.7319	0.7343	7,299
Proposed model (Viterbi)	0.8223	0.8188	0.8206	7,316
Proposed model (Cross-entropy)	0.8057	0.8056	0.8057	7,346

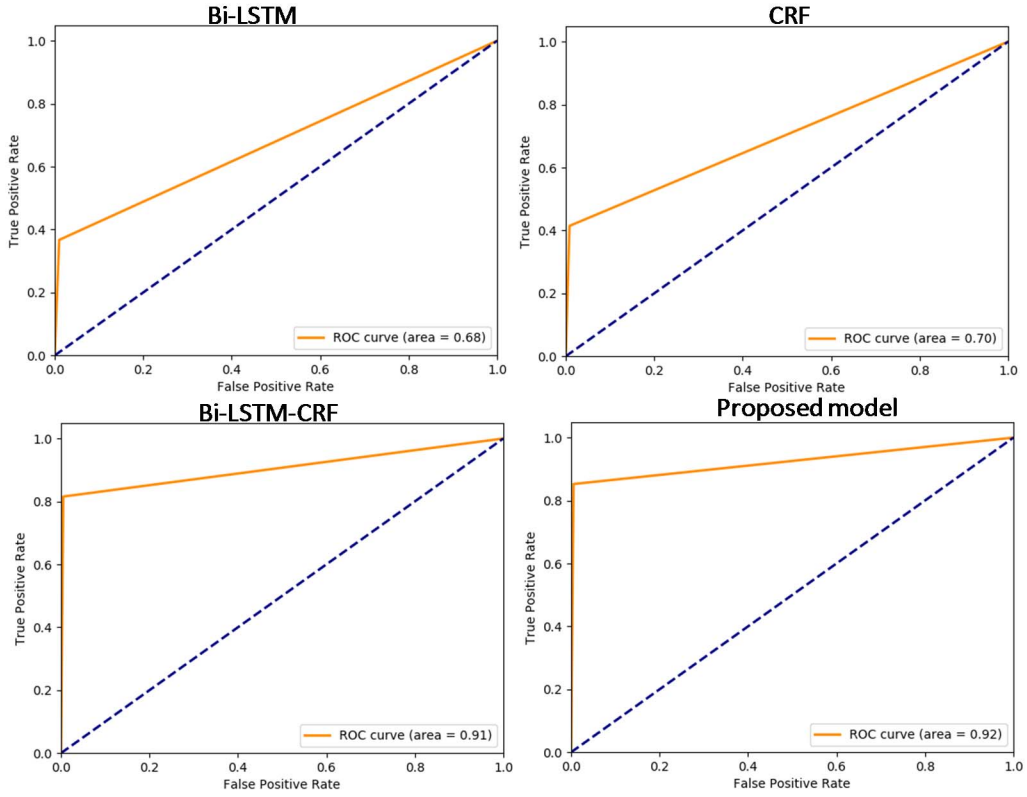

Figure 6. The AUC of Total Extracted Information

also indicates that, the Viterbi algorithm adds some constraints to eliminate wrong predictions during the prediction process, as well.

The reason for this phenomenon is because the outputs are not independent, and the Viterbi algorithm has better ability in finding the best path for sequential output.

3) The Bi-directional RNN and the CNN Layer

The experiment results show that pure Bi-LSTM model performs better than pure LSTM mode. This result confirms our assumption in section III.C, that for a general sequential modelling problem, 'future information' does help the model in making more accurate predictions and it, therefore, should be involved in the model.

From the last column of Table 2, we can find that the number of information extracted by Bi-LSTM is higher than CRF and LSTM, but the performance of Bi-LSTM is much worse than CRF. This indicates that Bi-LSTM does have the ability in extracting more information compared to LSTM, but it is also prone to involving more truth negative information. In general, Bi-LSTM has the potential to extract more diverse information but requires a constraining module or layer to filter out truth negative predictions.

The model combining Bi-LSTM and CRF shows a huge leap in recall, precision, and F1 score. This proves that Bi-LSTM-CRF model can remain the strength of the pure CRF model and pure Bi-LSTM model and also eliminate the drawbacks of these two models in the meantime.

The model proposed in this study outperforms the Bi-LSTM-CRF model. This result suggests that adding a convolution layer and then using appropriate fusion block is really achieving the extraction of some latent information, which is not captured by former embedding, RNN and later CRF layers. We maintain the reason for the improvement of the model is towing to the structure difference of the CNN layer and the RNN layer, and this makes it easier to model and represent different aspects of the same problem.

4) The Significance of the Fusion Block

Moreover, by comparing the model with the fusion block and without the fusion block, we can see that without the fusion block, the model with CNN layer performs even worse than Bi-LSTM-CRF model. The model extracts more truth negative information when adding the CNN layer but without using fusion block. This result indicates that, even though the CNN layer can extract some other latent information, the model still requires a proper information fusion procedure to combine different aspects of information. Simply adding a fusion block does increase the recall, precision, and F1-score in this experiment.

5) The Information Distinguishing Ability

As discussed in the previous section, the distribution of different types of information could vary greatly. For a personalized online learning service like micro learning, the ability to distinguish and extracting different types of information is significant for capturing characteristics of both learners and learning materials. From the other perspective to demonstrate the robustness of the proposed deep Bi-LSTM-CNNs-CRF model, we also compared such ability of representative models on different types of information. The

overall information distinguishing ability of these models is shown in Figure 6.

The proposed deep sequence labelling model greatly outperforms the pure CRF model and pure Bi-LSTM model, and the AUC score of our proposed model is about 2% higher than the mainstream Bi-LSTM-CRF model. For different types of information, the proposed model shows satisfactory performance than any other models. This result indicates that our model has great robustness, which can precisely locate and classify different types of information with different distributions. The details of the model performance in distinguishing our pre-defined types of information are shown in Figure. 7, where the class 1 to 3 represent the information A, B, and C, respectively (Table 1); the class 0 represents the useless information. The distributions of these types of information are shown in Table 1.

6) The Efficiency of the Proposed Model

Lastly, in this study, we also compared the training efficiency of three most complex models, Bi-LSTM-CRF, the proposed model, and the proposed model without fusion block. Due to the simplicity of the model structure and less satisfying performance, the training efficiency of pure CRF, LSTM, Bi-LSTM are not compared in the experiments. The decreasing of the loss value during the training process is shown in Figure 8. According to Figure 8, three models converge within similar training steps. Comparing with the mainstream sequence labelling model Bi-LSTM-CRF, the proposed model with additional CNN layer and fusion block does not show any obvious drop in the training efficiency. And the difference of training efficiency between the proposed model with and without fusion block is also very unnoticeable. Hence, our model does not require more extra training steps to reach an optimal state.

V. Conclusions and the Future Work

In this paper, we proposed a deep Bi-LSTM-CNNs-CRF model for extracting valuable information from massive and redundant information stream for micro learning services. We compared our proposed model with several classical and widely adopted sequence labelling models. The experiments conducted in this study demonstrated the robustness of the proposed model. The model can identify new latent features and outperforming the mainstream Bi-LSTM-CRF based model.

From our experimental results, we can conclude five specific points about information extraction or sequence labelling: 1. the CRF layer is vital for sequence modelling, for some cases, pure CRF model performs even better than pure RNN models such as LSTM or Bi-LSTM; 2. for a general application case, using bi-directional RNN such as Bi-LSTM

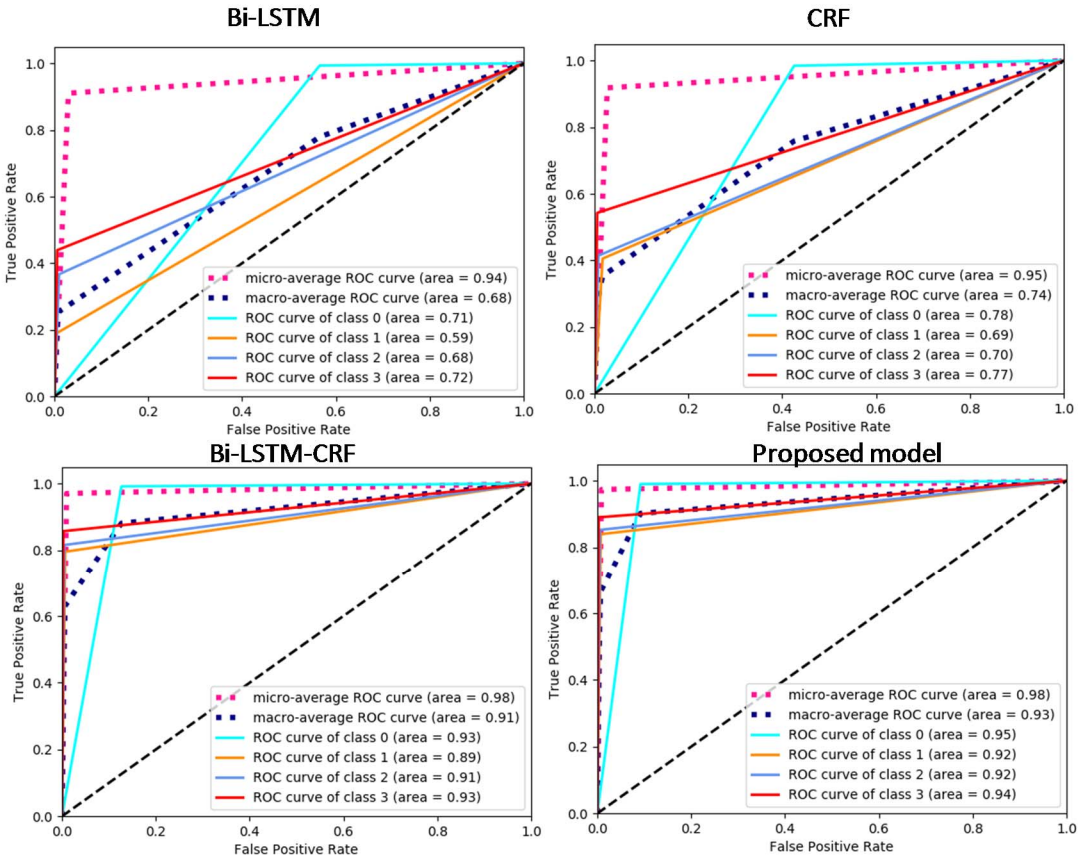

Figure 7. The ROC and AUC of Different Types of Extracted Information

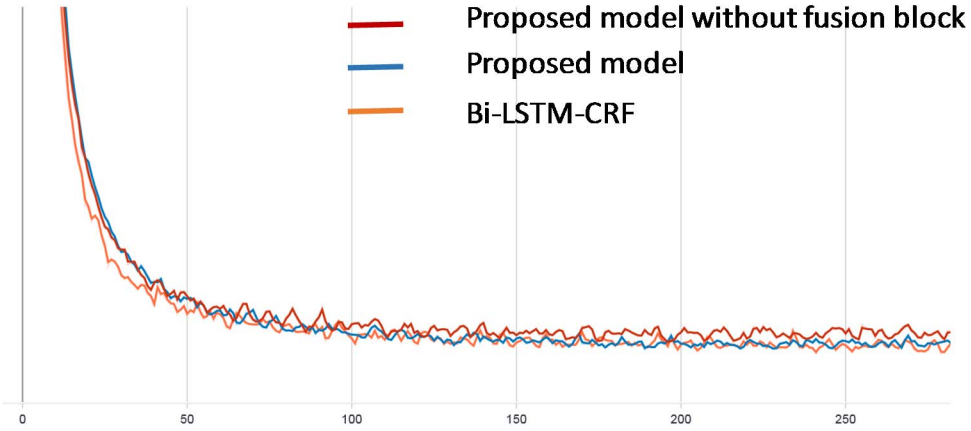

Figure 8. The Changes of the Loss Values for Each Training Epoch.
(Vertical and horizontal coordinate refers to the learning loss and learning epoch, respectively)

is a better choice than using single direction RNN; as 'future' input can anyhow provide some information for the sequence modelling; 3. CNN is useful in mining supplementary information and further boosting the performance of the current model; 4. the model proposed in this study demonstrates that it has the ability to efficiently mine useful information for online service. 5. compared with other representative models, our model shows satisfactory characteristics in both efficiency and robustness.

As discussed in a recent study [27], a learning service always has underlying pedagogical issues. Different subjects or disciplines present different contexts and may require significantly different information. Hence, in order to better adapt for a general micro learning service, it is pertinent to involve more global information in the information extraction process. In the future, we plan to continue improving the proposed information extraction model by mining and involving global information about the education and learning domain.

ACKNOWLEDGMENT

This research has been carried out with the support of the Australian Research Council Discovery Project, DP180101051, and Natural Science Foundation of China, no. 61877051, and UGPN RCF 2018-2019 project between University of Wollongong and University of Surrey. The work was also partially conducted during authors' collaborative visit to MIT and CSIRO.

REFERENCES

[1] H. Eshach, "Bridging in-school and out-of-school learning: Formal, non-formal, and informal education," Journal of science education and technology, vol. 16, no. 2, pp. 171-190, 2007.

[2] J. Lin, G. Sun, J. Shen, T. Cui, P. Yu, D. Xu, and L. Li, "A Survey of Segmentation, Annotation, and Recommendation Techniques in Micro Learning for Next Generation of OER," in 2019 IEEE 23rd International Conference on Computer Supported Cooperative Work in Design (CSCWD),2019, pp. 152-157.

[3] G. Sun, T. Cui, J. Yong, J. Shen, and S. Chen, "MLaaS: A Cloud-Based System for Delivering Adaptive Micro Learning in Mobile MOOC Learning," IEEE Transactions on Services Computing, no. 2, pp. 292-305, 2018.

[4] G. Sun, T. Cui, F. Dong, D. Xu, J. Shen, S. Chen, and J. Lin, "(WIP) Evaluation of a Cloud-Based System for Delivering Adaptive Micro Open Education Resource to Fresh Learners," in 2018 IEEE 11th International Conference on Cloud Computing (CLOUD),2018, pp. 586-589.

[5] G. Sun, T. Cui, G. Beydoun, S. Chen, F. Dong, D. Xu, and J. Shen, "Towards massive data and sparse data in adaptive micro open educational resource recommendation: a study on semantic knowledge base construction and cold start problem," Sustainability, vol. 9, no. 6, pp. 898, 2017.

[6] H. Guo, R. Tang, Y. Ye, Z. Li, and X. He, "DeepFM: a factorization-machine based neural network for CTR prediction," in Proceedings of the 26th International Joint Conference on Artificial Intelligence,2017, pp. 1725-1731.

[7] X. Ma, and E. Hovy, "End-to-end sequence labeling via bi-directional lstm-cnns-crf," arXiv preprint arXiv:1603.01354, 2016.

[8] A. McCallum, D. Freitag, and F. C. Pereira, "Maximum Entropy Markov Models for Information Extraction and Segmentation," in Icml,2000, pp. 591-598.

[9] J. Zhu, Z. Nie, J.-R. Wen, B. Zhang, and W.-Y. Ma, "2d conditional random fields for web information extraction," in Proceedings of the 22nd international conference on Machine learning,2005, pp. 1044-1051.

[10] T. Li, M. Choi, K. Fu, and L. Lin, "Music sequence prediction with mixture hidden markov models," arXiv preprint arXiv:1809.00842, 2018.

[11] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in Advances in neural information processing systems,2014, pp. 3104-3112.

[12] C. Dyer, M. Ballesteros, W. Ling, A. Matthews, and N. A. Smith, "Transition-based dependency parsing with stack long short-term memory," arXiv preprint arXiv:1505.08075, 2015.

[13] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," arXiv preprint arXiv:1508.01991, 2015.

[14] J. Yang, S. Liang, and Y. Zhang, "Design Challenges and Misconceptions in Neural Sequence Labeling," in Proceedings of the 27th International Conference on Computational Linguistics,2018, pp. 3879-3889.

[15] A. Abdullah, R. Kumar, A. McGregor, S. Vassilvitskii, and S. Venkatasubramanian, "Sketching, Embedding and Dimensionality

Reduction in Information Theoretic Spaces," in Artificial Intelligence and Statistics,2016, pp. 948-956.

[16] B. Mitra, and N. Craswell, "Neural text embeddings for information retrieval," in Proceedings of the Tenth ACM International Conference on Web Search and Data Mining,2017, pp. 813-814.

[17] Y. Liu, L. Li, and J. Liu, "Bilateral neural embedding for collaborative filtering-based multimedia recommendation," Multimedia Tools and Applications, vol. 77, no. 10, pp. 12533-12544, 2018.

[18] N. Zhou, W. X. Zhao, X. Zhang, J.-R. Wen, and S. Wang, "A general multi-context embedding model for mining human trajectory data," IEEE transactions on knowledge and data engineering, vol. 28, no. 8, pp. 1945-1958, 2016.

[19] Z. Zhang, and V. Saligrama, "Zero-shot learning via semantic similarity embedding," in Proceedings of the IEEE international conference on computer vision,2015, pp. 4166-4174.

[20] W. Yin, K. Kann, M. Yu, and H. Schütze, "Comparative study of CNN and RNN for natural language processing," arXiv preprint arXiv:1702.01923, 2017.

[21] Z. Zhai, D. Q. Nguyen, and K. Verspoor, "Comparing CNN and LSTM character-level embeddings in BiLSTM-CRF models for chemical and disease named entity recognition," arXiv preprint arXiv:1808.08450, 2018.

[22] G. Papandreou, I. Kokkinos, and P.-A. Savalle, "Modeling local and global deformations in deep learning: Epitomic convolution, multiple instance learning, and sliding window detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2015, pp. 390-399.

[23] Z. Chen, and W. Li, "Multisensor feature fusion for bearing fault diagnosis using sparse autoencoder and deep belief network," IEEE Transactions on Instrumentation and Measurement, vol. 66, no. 7, pp. 1693-1702, 2017.

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers),2019, pp. 4171-4186.

[25] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep Contextualized Word Representations," in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers),2018, pp. 2227-2237.

[26] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.

[27] D. Wu, J. Lu, and G. Zhang, "A fuzzy tree matching-based personalized e-learning recommender system," IEEE transactions on fuzzy systems, vol. 23, no. 6, pp. 2412-2426, 2015.