運用集成式多通道類神經網路於科技英文寫作評估

Scientific Writing Evaluation Using Ensemble Multi-channel Neural Networks

王昱翔 Yuh-Shyang Wang, 李龍豪 Lung-Hao Lee
國立中央大學電機工程學系
Department of Electrical Engineering
National Central University
yswang135@gmail.com, lhlee@ee.ncu.edu.tw

林柏霖 Bo-Lin Lin, 禹良治 Liang-Chih Yu
元智大學資訊管理學系
Department of Information Management
Yuan Ze University
ss27713084@gmail.com, lcyu@saturn.yzu.edu.tw

摘要

現在有許多母語非英語人士撰寫的科學論文，幫助作者撰寫科學論文的自動化工具產生了巨大的需求。國際科技英文寫作評估評測任務藉由評估一個論文中的英文句子，是否需要語言編輯為任務目標，幫助開發自然語言處理工具，用以改善科技英文寫作的品質。本研究透過實驗設計比較通道數、模型架構和集成數，提出一個集成式多通道類神經網路架構，在該評測資料集下獲得 F1 分數 63.28，比當時參與評測的系統有更好的效能。

Abstract

A huge number of scientific papers have been authored by non-native English speakers. There is a large demand for effective computer-based writing tools to help writers composing scientific articles. The Automated Evaluation of Scientific Writing (AESW) shared task seeks to promote the use of NLP tools for improving the quality of scientific writing in English by predicting whether a given sentence needs language editing or not. In this study, we propose an ensemble multi-channel BiLSTM-CNN model based on a series of experiments in comparing the number of channels, network architectures, and ensemble size. Our model achieved an F1 score of 63.28 outperforms participating systems in the AESW 2016 task.

關鍵詞：集成學習、多通道神經網路、自動寫作評估, 科技英文

Keywords: Ensemble Learning, Multi-channel Neural Networks, Automated Writing Evaluation, Scientific English

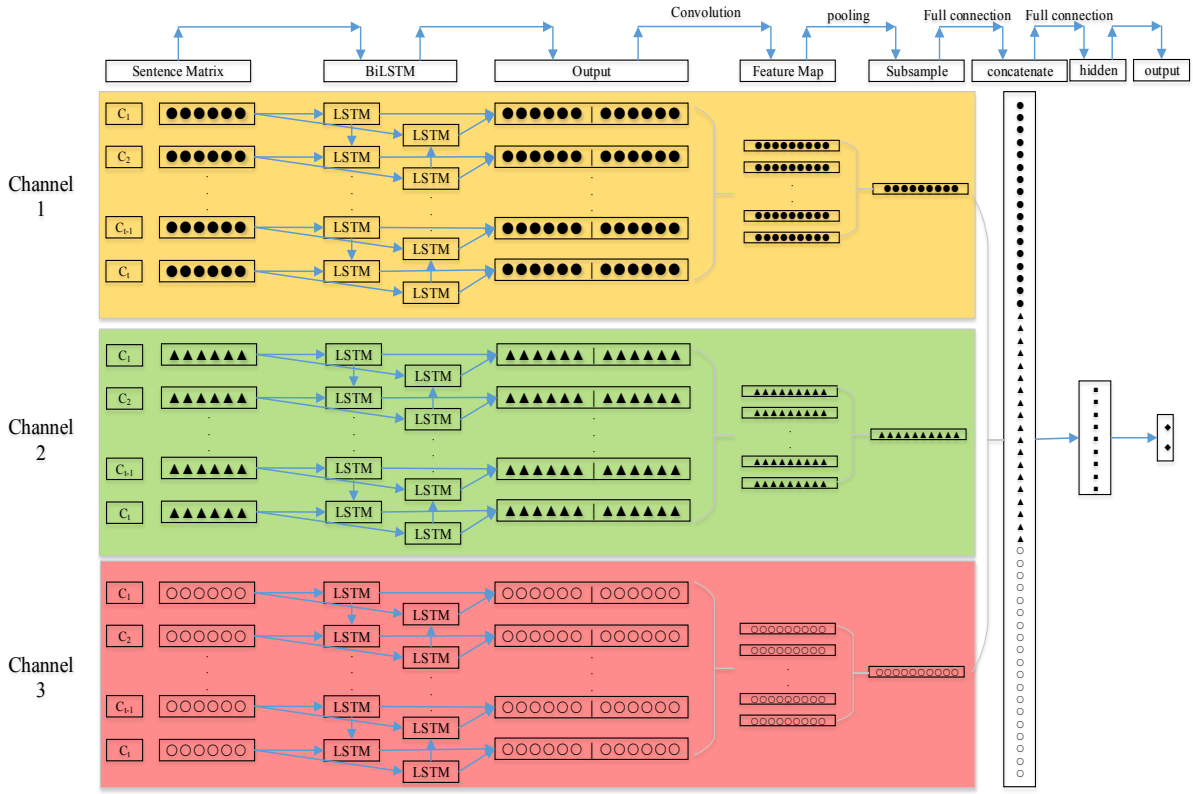
一、緒論

英文是世界上最常被使用的語言之一，有 67 個國家將英文列為他們的官方語言，非英語系的國家通常都將英文視為第一外語或第二外語學習。語言學習有四個重點「聽、說、讀、寫」，在正式的場合如信件、商業合約、論文等等，對於寫作者需要額外的人力去校稿。近年有越來越多非英文母語者撰寫的科學論文，因而對科技英文寫作的自動化評測工具的需求與日俱增。一個有效的自動化寫作工具，可以幫助寫作者減少語意表達上的錯誤，提升寫作品質，進而減少校稿的時間以及人力成本。為了建立自動化工具，語法錯誤檢測和更正是必要的一環，也有許多的任務競賽：Helping Our Own (HOO)是一系列用於寫作者文法錯誤校正任務 [1][2]。CoNLL 2013/2014 的評測任務則是英文為外語學習者的文法錯誤更正[3][4]。

本研究使用的資料來源為科技英文寫作評估競賽(Automated Evaluation of Scientific Writing, AESW)。AESW 2016 評測任務的目的是分析科學寫作的語言特徵，促進科技論文的自動化寫作評估工具的發展[5]。任務內容分為兩個子任務：一是二元分類預測，輸入的句子判定是否需要語言校正，如果是則預測為 True，反之為 False；二是機率估算，系統需要估計句子需要被校正的機率值。有鑑於不同的詞嵌入以及神經網路模型在語法偵錯任務上各有優缺點，本研究透過完整的實驗流程，結合各種詞嵌入(word embedding)與卷積神經網路（Convolutional Neural Network, CNN）以及長短期記憶神經網路（Long Short-Term Neural Network, LSTM）兩個深度學習的基礎模型，藉由實驗比較通道數、模型架構、集成數對子任務一分類效能的差異，建構出集成式多通道類神經網路，在測試集上達到 63.28 的 F1 分數，與競賽時的方法相比較，有更好的分類成效。

二、模型架構

我們提出的方法為集成式多通道類神經網路（Ensemble Multi-Channel BiLSTM-CNN）模型，其架構圖如圖一，由詞向量輸入、多通道輸入、雙向長短期記憶網路、卷積神經網路所組成。



圖一、集成式多通道類神經網路架構

(一)、詞嵌入向量 (Word Embedding)

詞向量是在自然語言處理中常用的方法，要將語言輸入給機器運算前，需要將其數值化，詞向量就是將文句中的詞數值化的方式，將每一個單詞以一個向量表示。最簡單的方式維 One-hot Encoding 是用一個維度等同文本中詞彙數的向量來表示，向量中只包含一個 1 與多個 0，字典中第一個詞表示為[1,0,0,……,0]、第二個詞為[0,1,0,0,……,0]以此類推。這種表示的缺點為當文本、字典較大時，代表每個詞的維度就會變得極為巨大，引發維度災難，造成運算上的困難，而且這種表示方式對於詞與詞之間的關係沒有代表性。

為了解決該上述問題，Hinton 於 1986 年提出 Distributed Representations，將所有詞向量組合成一個詞向量空間，每個向量則為空間上的一點，點與點之間的距離即為詞之間的相似性[6]。常用的傳統靜態詞向量工具有 Word2vec, GloVe, fastText。Word2Vec 為 Mikolov 於 2103 年提出，使用連續詞袋模型(continuous bag of words, CBOW)和 skip-gram，以非監督式學習的算法學習單詞的含義[7]。GloVe 為史丹佛大學於 2014 年提出，基於文本內詞彙的共現矩陣(Co-occurrence Matrix)，對共現矩陣進行訓練，計算出詞向量[8]。fastText 為 Facebook 於 2016 提出，相較於 Word2vec，fastText 引入了 N-gram 考

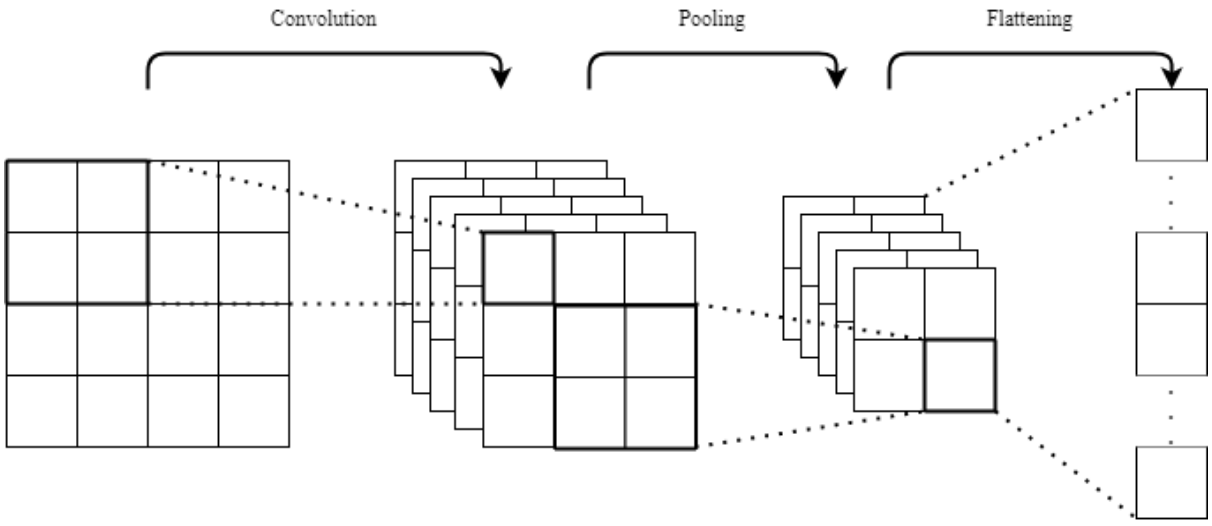
慮詞序特徵，並使用 subword 來處理未登錄詞[9]。

本研究採用三個詞向量模型分別為：Word2vec 官方利用 GoogleNews dataset 所訓練的詞向量模型，內含各為 300 維向量的 300 萬個字詞。GloVe 官方利用 Common Crawl 資料集所訓練的 glove.840.300d 詞向量模型，包含 300 維的 220 萬個字詞。fastText 官方提供使用 Wikipedia 所訓練的 294 個不同語言的詞向量中的英文詞向量，其維度也是 300 維。

(二)、卷積神經網路(Convolutional Neural Network, CNN)

卷積神經網路(CNN)示意圖如圖二，在圖像處理上有出色的表現，也能有效處理自然語言中的語意分析、分類、預測等任務。在 2016 年 DeepMind 透過結合蒙地卡羅搜尋法(MCTS) 與深度卷積神經網路(DCNN)提出的演算法開發出 AlphaGo，並與韓國職業九段棋士世界冠軍李世乭對弈以四勝一敗獲勝，引起世界大量關注。

CNN 有兩個主要部分：卷積層(Convolution Layer)和池化層(Pooling Layer)，卷積層透過在輸入圖像上的數個卷積核滑動計算並提取資料的特徵，各個卷積核可以分別得出一個特徵圖(Feature Maps)。而池化層將前面獲得的特徵圖做次採樣(Subsampling)，最常見的方法為最大池化(Max Pooling)，將特徵圖切為多個矩形，輸出各區的最大值，透過池化可以保留顯著的特徵並降低特徵的數量。

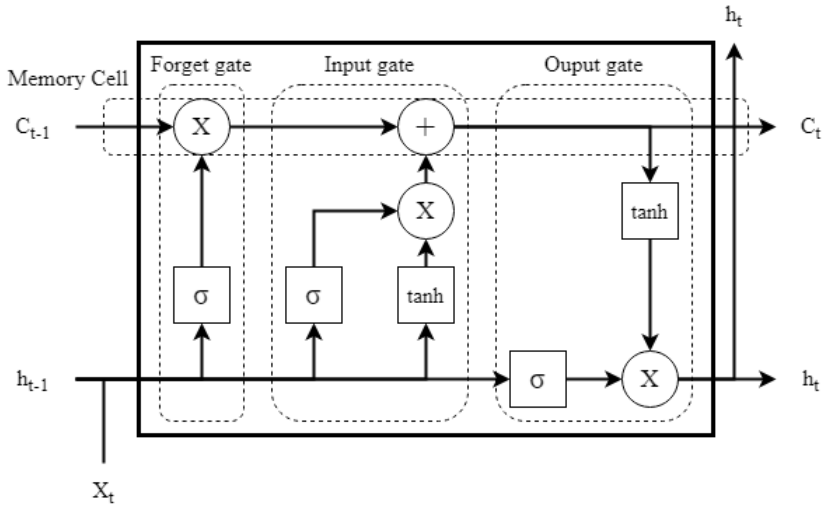


圖二、卷積神經網路

(三)、長短期記憶網路 (Long Short-Term Memory, LSTM)

1、 單向長短期記憶網路 (LSTM)

長短期記憶網路(LSTM) 是一種時間循環神經網路 (Recurrent Neural Network，RNN)。一般的 RNN 各節點的輸入為輸入的資料以及前一個節點的輸出，當序列較長時前面的資訊便無法完整傳遞到後面，因此 RNN 在處理短文句時會有很好的表現，但是在較長的句子時就會無法順利預測。

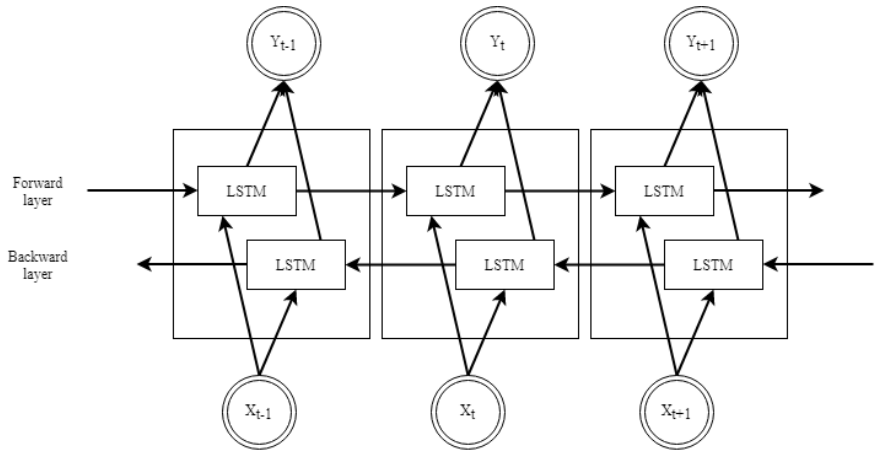


圖三、LSTM Cell

因此 1997 年 Hochreiter 和 Schmidhuber 提出了 LSTM，基於 RNN 架構引入了記憶單元來解決這個問題。LSTM Cell (如圖三)由四個元件組成輸入門(Input Gate)、輸出門(Output Gate)、遺忘門(Forget Gate)、記憶單元(Memory Cell)，輸入門決定要輸入到記憶單元的特徵、遺忘門決定要刪除的特徵訊息、而輸出門則是決定記憶單元內的特徵是否能輸出。

2、雙向長短期記憶 (Bi-directional Long Short-Term Memory, BiLSTM)

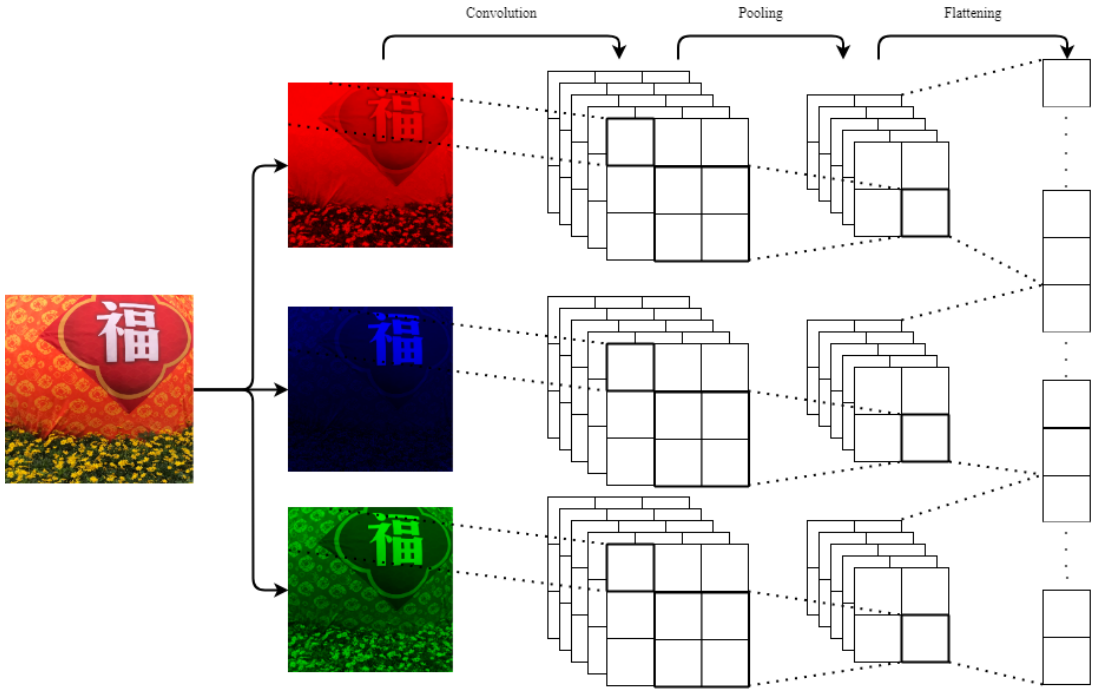
單向長短期記憶在學習到時間點 t 時，只能獲得該時間點之前的訊息，而雙向長短期記憶網路(架構圖如圖四)則能在每個時間點都獲得前後文的序列狀態。在英文文法偵錯時，需要同時注意前後文以確認時態、助動詞等是否正確使用，因此 BiLSTM 應較 LSTM 更適合本實驗。



圖四、雙向長短期記憶網路架構

(四)、多通道輸入 (Multi-Channel)

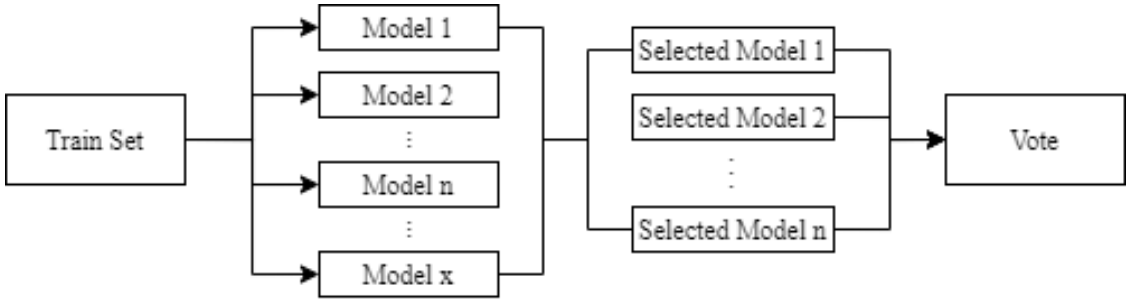
單通道僅能用將資料以一種表示方式輸入網路，若使用多通道則可將資料以多種方式輸入類神經網路之中。以彩色圖片使用多通道卷積神經網路進行分類如圖五為例：將圖片以 RGB 三元素表示時，將 R、G、B 數值分為三通道輸入，分別提取特徵，拓展卷積神經網路的視野。



圖五、多通道卷積神經網路

(五)、集成學習 (Ensemble Learning)

集成學習為使用多種學習算法來獲得比單獨使用一種學習算法更好的預測性能，如俗話說「三個臭皮匠勝過一個諸葛亮」。集成學習的常見方法有 Bagging、Boosting、Stacking 等，我們採用類似 Bagging(如圖六)的作法，將同一份訓練集，訓練多個模型並選擇表現較好的模型進行投票。



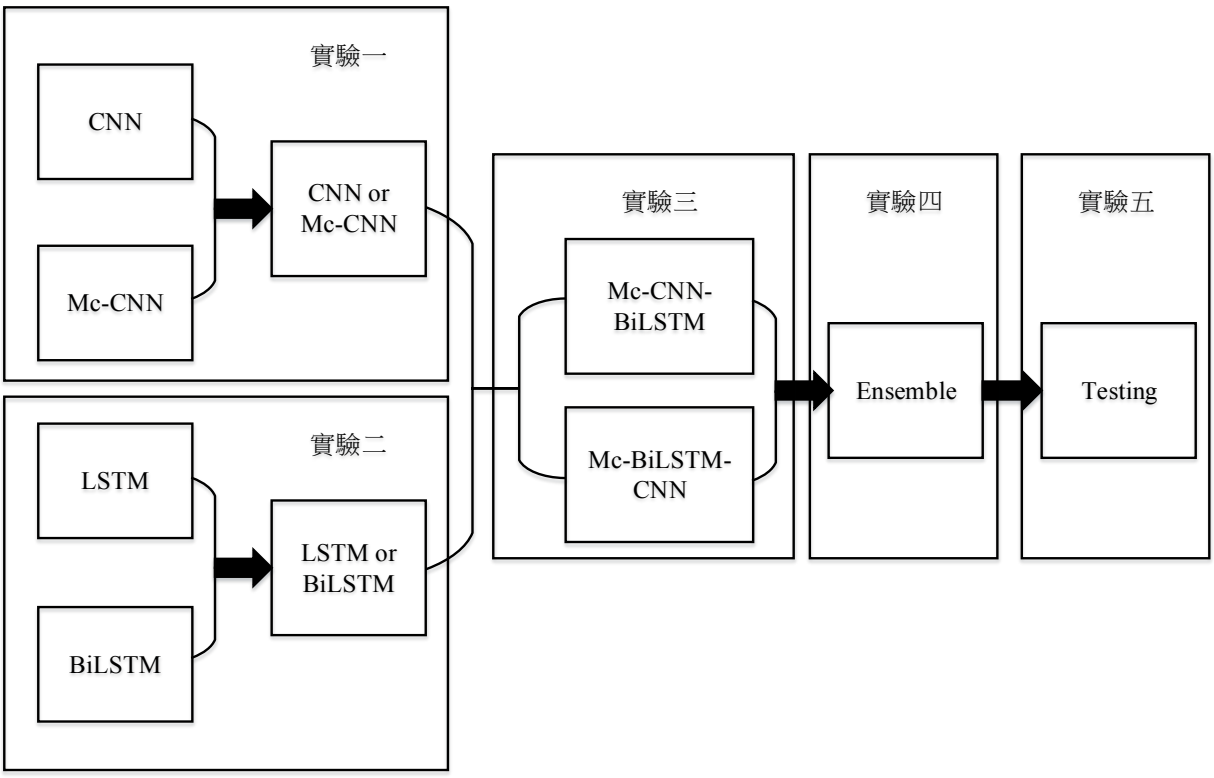
圖六、集成學習投票機制

三、實驗與結果

(一)、實驗設計

實驗使用的資料來源為 AESW 2016 科技英文自動評測國際競賽提供的資料，它是第一個大規模公開的科技寫作資料集，資料中包含需要語言編輯以符合科技論文寫作的體裁的文句，以及不須編輯便已符合體裁的文句，需要經過編輯的文句會附上如何編輯以符合體裁。<del>標籤中為需要刪除的字詞，而<ins>中則是需要加入的字詞。訓練集共有 1,196,940 筆，其中 466,672 筆為經語言編輯(True 標記)，發展集有 148,478 筆，其中 57,340 筆為經語言編輯，而測試集有 143,784 筆。三組資料中經語言編輯比例均約佔四成、未經修改的則為六成。該競賽目標為輸入一個未經編輯的文句，預測其是否需要語言編輯以符合寫作體裁，評估方式是將模型預測結果與實際答案做比對，以 F1 為主要評分標準。

本實驗分為五個階段(如圖七)，透過逐一比較效果並決定參數，實驗一比較單通道與多通道的表現，實驗二做單向與雙向 LSTM 的比較，實驗三再依照前兩個比較出較好的模型做組合，實驗四是集成多模型的結果，決定出最佳模型及參數。前四個實驗均是在發展集上調整參數，以三次實驗平均作為比較標準，最後實驗五再對測試集做預測。



圖七、實驗流程

(二)、實驗一結果

第一個實驗比較單通道、雙通道、三通道在 CNN 最佳參數時的平均表現，以決定後續實驗所使用的通道數，以及詞向量組合順序。不同通道數的表現結果分見於表一、表二和表三，單通道是以使用 GloVe 為詞向量時表現最佳，平均 F1 為 0.6392，而多通道包含雙通道及三通道中表現最佳的是三通道的 Word2vec+ FastText+GloVe 組合，平均 F1 為 0.6422，此處實驗可得知三通道在此實驗上表現較好，結合訓練方式各異的詞向量特徵，更能有效獲得語句中的訊息。

表一、單通道 CNN 於發展集的結果

Embedding	**Avg. Precision**	**Avg. Recall**	**Avg. F1**	**F1 Std.**
GloVe	0.5484	0.7661	**0.6392**	0.0004
fastText	0.5991	0.6729	0.6365	0.0012
Word2vec	0.5763	0.7065	0.6347	0.0016

表二、雙通道 CNN 於發展集的結果

Embedding	Avg. Precision	Avg. Recall	Avg. F1	F1 Std.
Word2vec + GloVe	0.5636	0.7354	0.6375	0.0014
GloVe + Word2vec	0.5753	0.7129	0.6367	0.0011
GloVe + fastText	0.5683	0.7266	0.6375	0.0031
fastText + GloVe	0.5507	0.7617	**0.6390**	0.0024
fastText + Word2vec	0.5647	0.7241	0.6341	0.0029
Word2vec + fastText	0.5746	0.7121	0.6360	0.0010

表三、三通道 CNN 於發展集的結果

Embedding	Avg. Precision	Avg. Recall	Avg. F1	F1 Std.
Word2vec + GloVe + fastText	0.5617	0.7473	0.6404	0.0028
GloVe + Word2vec + fastText	0.5638	0.7383	0.6391	0.0003
GloVe + fastText + Word2vec	0.5654	0.7376	0.6400	0.0023
fastText + GloVe + Word2vec	0.5423	0.7838	0.6403	0.0030
fastText + Word2vec + GloVe	0.5722	0.7264	0.6396	0.0036
Word2vec + fastText + GloVe	0.5524	0.7671	**0.6422**	0.0017

(三)、實驗二

第二個實驗比較單雙向長短期記憶網路(LSTM vs. BiLSTM)在此任務上的表現，作為後續實驗模型選擇的依據。單雙向模型的表現分別在表四及表五，LSTM 以 GloVe 為詞向量時表現最好，平均 F1 為 0.6419；BiLSTM 以 GloVe 為詞向量時表現最好，平均 F1 為 0.6473，勝過 LSTM。與前面推論相同，BiLSTM 能獲取前後文訊息，較適合語法語意偵錯的任務。

表四、LSTM 於發展集的結果

Embedding	Avg. Precision	Avg. Recall	Avg. F1	F1 Std.
GloVe	0.5201	0.8383	**0.6419**	0.0007
fastText	0.5151	0.8359	0.6373	0.0020
Word2vec	0.5221	0.8264	0.6399	0.0013

表五、BiLSTM 於發展集的結果

Embedding	Avg. Precision	Avg. Recall	Avg. F1	F1 Std.
GloVe	0.5529	0.7803	**0.6473**	0.0002
fastText	0.5458	0.7888	0.6452	0.0004
Word2vec	0.5443	0.7889	0.6441	0.0006

(四) 實驗三

第三個實驗根據前兩個實驗的結果，選用多通道、BiLSTM 與 CNN 組合，建立更複雜的網路，使用多通道表現最好的詞向量組合 Word2vec+fastText+GloVe，實驗結果記錄於表六。兩模型中前者為將詞向量經由 CNN 取得 feature map 再進入 BiLSTM，後者則反之。Mc-BiLSTM-CNN 的平均 F1 為 0.6536 較 Mc-CNN-BiLSTM 的 0.6491 來得好。

表六、Mc-BiLSTM-CNN 與 Mc-CNN-BiLSTM 於發展集的結果

Model	Avg. Precision	Avg. Recall	Avg. F1	F1 Std.
Mc-BiLSTM-CNN	0.5529	0.7803	**0.6473**	0.0002
Mc-CNN-BiLSTM	0.5458	0.7888	0.6452	0.0004

(五)、實驗四

第四個實驗為了驗證集成(ensemble)學習的效果，利用實驗三所決定的模型架構、參數，訓練了數十個模型，取出表現最佳的前 N 名，若超過（N/2）+1 個模型的預測結果認為該句有誤，便認為該句文法有誤，反之亦然。實驗結果如表七，可以看出 N 為 3 和 5 時表現提升不少，增加到 7 和 9 時幾乎沒有提升，11 時反而表現變差，因此決定採用最佳的前 9 名為最終參數。

表七、Ensemble Mc-BiLSTM-CNN 於發展集的結果

Ensemble (N)	Precision	Recall	F1
3	0.5606	0.7980	0.65860
5	0.5599	0.8010	0.65911
7	0.5586	0.8040	0.65928
9	0.5676	0.7866	0.**65994**
11	0.5652	0.7907	0.65924

(六)、實驗五

將以上實驗中的最佳模型，在 AESW2016 國際評測的測試集，驗證效能結果如下表八。與實驗預期相同，多通道比單通道好，深層網路架構，以及集成學習都是有效提升模型表現的方法。而表九則是與該競賽參賽隊伍的成績做比較，我們的最佳模型 Ensemble(N=9) Mc-BiLSTM-CNN 達到 F1 分數 0.6328，超越了表現最好的哈佛大學團隊的 0.6278。

表八、實驗中各模型測試結果

Method	**Precision**	**Recall**	**F1**
CNN	0.5274	0.7153	0.6071
Mc-CNN	0.5256	0.7405	0.6148
LSTM	0.4786	0.8316	0.6076
BiLSTM	0.5157	0.7735	0.6188
Mc-CNN-BiLSTM	0.5257	0.7581	0.6209
Mc-BiLSTM-CNN	0.5144	0.7988	0.6258
Ensemble (N=9) Mc-BiLSTM-CNN	0.5359	0.7724	**0.6328**

表九、測試結果與參賽隊伍的比較

Team	Method	Precision	Recall	F1
Hu	CNN, RNN, LSTM	0.5444	0.7413	0.6278
HITS	HMM, Logistic Regression	0.3765	0.9480	0.5389
ISWD	SVM, SubSet Tree kernel	0.4482	0.7279	0.5548
Knowlet	MaxEnt	0.6241	0.3685	0.4634
NTNU-YZU	CNN	0.5025	0.7785	0.6108
UW-SU	MaxEnt	0.4145	0.8201	0.5507
Ours	**Ensemble (N=9) Mc-BiLSTM-CNN**	0.5359	0.7724	**0.6328**

四、結論

本研究將科技英文寫作的句子是否需要語言編修，視為一個標準的二元分類問題，使用 AESW2016 國際競賽的資料集驗證方法成效。在文句分類時，最重要的就是文句的表示方式以及找到合適的模型，透過實驗由不同的詞向量(GloVe, Word2vec, fastText)、通道選擇(單通道、雙通道、三通道)、模型的挑選(CNN vs. Mc-CNN, LSTM vs. BiLSTM, Mc-CNN-BiLSTM vs. Mc-BiLSTM-CNN)，到 Ensemble 個數，最後提出了 Ensemble (N=9) Mc-BiLSTM-CNN 集成式多通道類神經網路模型，這是一種利用多種靜態詞向量，透過多個神經網路集成預測結果的分類模型，此方法在測試集上得到最好的 F1 分數為 0.6328，較當時參賽隊伍中表現最好的表現高。未來希望除了靜態詞向量之外，可以採用動態詞向量如：BERT 或者是 ELMO，或是使用 Transformer 進行分類。

致謝

This work was partially supported by the Ministry of Science and Technology, Taiwan under the grant MOST 108-2218-E-008-017-MY3 and MOST 107-2628-E-155-002-MY3.

參考文獻

[1] R. Dale and A. Kilgarriff. 2011. Helping Our Own: The HOO 2011 pilot shared task. In

Proceedings of the 13th European Workshop on Natural Language Generation, pages 242–249.

[2] R Dale, I Anisimoff, and G Narroway. 2012. A report on the preposition and determiner error correction shared task. In *Proceedings of the NAACL Workshop on Innovative Use of NLP for Building Educational Applications*.

[3] H. T. Ng, S. M. Wu, C. Hadiwinoto, and J. Tetreault. 2013. The CoNLL-2013 Shared Task on Grammatical Error Correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–12.

[4] H. T. Ng, S. M. Wu, T. Briscoe, C. Hadiwinoto, R. H. Susanto, and C. Bryant. 2014. The CoNLL-2014 Shared Task on Grammatical Error Correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14.

[5] V. Daudaravicius, R. E. Banchs, E. Volodina and C. Napoles. 2016. A report on the automated evaluation of scientific writing shared task. In *Proceedings of the 11th Workshop on the Innovative Use of NLP for Building Educational Applications*, pages 53-62.

[6] G. E. Hinton. Learning distributed representations of concepts. 1986. In *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*, pages 1-12.

[7] T. Mikolov, I. Sutskever, K. Chen, G. Corrado and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of Neural Information Processing Systems 2013*, pages 1-10.

[8] J. Pennington, R. Socher and C. D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Empirical Methods on Natural Language Processing*, pages 1532-1543.

[9] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.