The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 334-342

基於階層式編碼架構之文本可讀性預測

A Hierarchical Encoding Framework for Text Readability Prediction

翁詩諺　Shi-Yan Weng
國立臺灣師範大學資訊工程學系
Department of Computer Science and Information Engineering
National Taiwan Normal University
40547041S@ntnu.edu.tw

曾厚強　Hou-Chiang Tseng
國立臺灣師範大學資訊工程學系 / 心理與教育測驗研究發展中心
Department of Computer Science and Information Engineering/ Research Center for Psychological and Educational Testing
National Taiwan Normal University
ouartz99@gmail.com

宋曜廷　Yao-Ting Sung
國立臺灣師範大學教育心理與輔導學系
Department of Educational Psychology and Counseling
National Taiwan Normal University
sungtc@ntnu.edu.tw

陳柏琳　Berlin Chen
國立臺灣師範大學資訊工程學系
Department of Computer Science and Information Engineering
National Taiwan Normal University
berlin@ntnu.edu.tw

摘要

以教育的角度來看，為了幫助學生獲得更好的學習效果，對每個年級安排適當的難度文本是非常重要的。因此，長久以來，陸續有許多學術機構致力於可讀性模型或特徵的研究。為了解決這個問題，在先前的研究常使用一些人為定義的特徵，例如難詞頻率或字數等特徵來對文本進行可讀性難易程度預測。然而，這些特徵可能太淺層而不能表示文本的語法、語意或更深層的內涵。近期，由於深度學習或表示學習技術的蓬勃發展，使得更有代表性的語言特徵能從文本中被萃取出來增進可讀性難易程度的準確性。延伸此技術發展趨勢，在本論文中我們設計並實作出具有階層式編碼的類神經網路來做為可讀性預測模型，以擷取在文本中的詞彙到語句、語句到文本的語意和結構表示資訊。此外，

我們並嘗試在此模型中額外加入傳統的人為定義特徵作為輔助資訊。從實驗結果可以發現，我們提出的可讀性預測模型具有良好的效能表現，而加入傳統的人為定義特徵，亦可以進一步增進其預測的準確性。

Abstract

From an educational perspective, it is important to provide students of different grades with reading material of appropriate difficulty for better learning retention. To deal with this problem, it is common practice to use a set of handcrafted features, for example, hard word rate or word count, to distinguish articles into different readability levels. However, these traditional readability features are often too shallow to represent deeper semantic or syntactic structures of the articles. In view of this, we present a modeling approach that leverages a recurrent neural network to hierarchically encode both the semantic or syntactic structures of a given article for better readability classification. Furthermore, we also seek to make extra use of traditional handcrafted feature as side information to further boost the performance.

關鍵詞：可讀性、語言特徵、表示學習法、卷積式類神經網路、遞迴式類神經網路

Keywords: Readability、Language Feature、Representation Learning、Convolutional Neural Network、Recurrent Neural Network

一、緒論

可讀性(Readability)是指閱讀材料能夠被讀者所理解的程度[1],[2],[3],[4]；當讀者閱讀較高可讀性的文本時，會產生較好的理解及學後保留效果[2],[3]。西方的可讀性公式發展的非常早[5],[6]，據 Chall 與 Dale[7]在 1995 年的統計，到 1980 年為止相關的可讀性公式就已經超過 200 多種。這些傳統的可讀性研究大多使用較淺層的語言特徵來發展線性的可讀性公式。例如著名的 Flesch Reading Ease 公式以詞彙音節數做為語意的指標、以語句的長度作為語法的指標，透過計算詞彙的平均音節數與文本的平均語句長度來評估文本的可讀性難易程度：當文本的詞彙音節數愈多、語句愈長時，則該文本在閱讀理解時愈為困難。然而，傳統可讀性公式所採用的淺層語言特徵，並不足以反映文本閱讀理解的難度。Graesser、Singer 和 Trabasso[8]便指出，傳統可讀性公式無法反映閱讀的真實歷程，並沒有考量文本的深層特性，例如語句的前後順序。Collins-Thompson[9]亦指出傳統可讀性公式僅著重在使用文本的表淺資訊，而忽略文本其它的重要特徵，例如文

本結構資訊或 *N*-連語言樣式(*N*-gram Patterns)。這也讓傳統可讀性公式在預測文本可讀性的結果常遭受到質疑。甚至因為可讀性公式所採用的可讀性特徵過少，導致容易受到有心人士為了達到特定的可讀性數值，而刻意針對可讀性特徵的特性來修改文本，使得文本呈現許多簡短而破碎的句子，反而降低了文本的流暢度與連貫性，因此增加閱讀難度[10]。直至今日，可讀性的研究仍持續不斷。研究人員為了克服傳統可讀性公式的缺點，嘗試利用較複雜的機器學習演算法來發展出更細緻、非線性的可讀性模型；同時，亦納入更多元的可讀性指標來共同評量文本的可讀性，除了可以提升可讀性模型的效能，亦可防止有心人士去操弄文本的可讀性[11],[12],[13]。

而在近年，由於深度學習(或表示學習)在自然語言處理領域的蓬勃發展，讓我們能夠藉由相關技術從文本中萃取出更深層的語意或結構特徵。例如，在自動文件摘要的研究上，有所謂的序列對序列(Sequence-to-Sequence, Seq2Seq)模型架構提出，透過在不同階段使用卷積式類神經網路(Convolutional Neural Network, CNN)與遞迴式類神經網路(Recurrent Neural Network, RNN)來達成語句或文本(文本)的固定長度語意向量表示[14]，並藉此產生出一份較為簡短扼要的摘要來代表原始文件。而在[15]中，研究人員採用了更複雜的序列對序列模型並加入了強化學習(Reinforcement Learning)使得自動摘要的表現更好。

基於上述的研究與相關技術發展，在本論文中我們設計並實作出具有階層式編碼的類神經網路來做為可讀性預測模型。首先，透過以卷積式類神經網路(CNN)以擷取每一句語句內部的局部詞彙使用特徵；接著，經由遞迴式類神經網路(RNN)依照語句向前或向後讀取卷積式類神經網路所產生之語句的語意向量表示，來產生同時含有文本語句結構資訊之文本的語意向量表示，最後用於於文本可讀性難易程度預測。同時，我們並嘗試在此模型中額外加入傳統的人為定義特徵作為輔助資訊以期能更進一步提升可讀性預測模型的效能表現。本論文接下來的安排如下：第二節描述我們提出的文本階層式編碼模型架構；第三節將呈現實驗材料及相關的實驗設定及結果；最後第四節是總結及未來研究的方向。

二、階層式編碼架構

基於階層式編碼模型架構之文本可讀性預測模型是如圖一所意示，它主要可分為兩個部分：首先為結合卷積式類神經網路(CNN)和遞迴式類神經網路(RNN)所組成的文本編碼

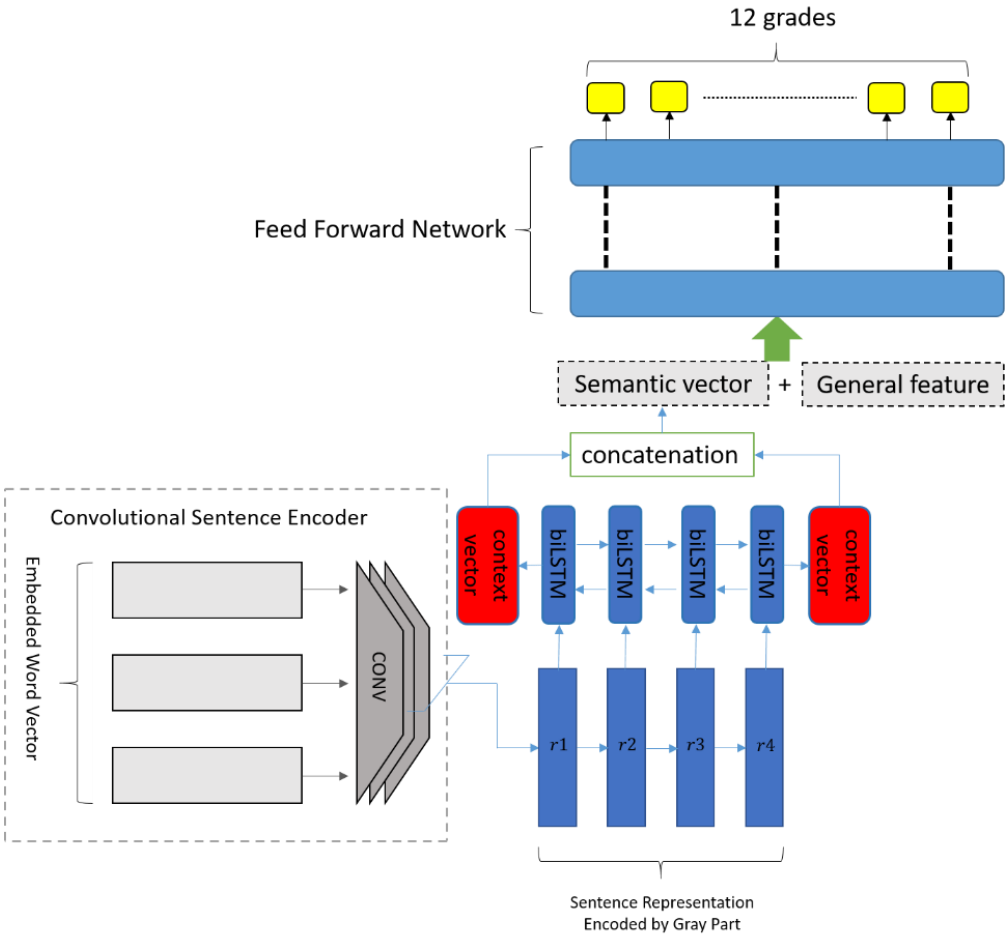
(Encoding)結構，此部分主要的目標在於得到對於每一篇文本的高階特徵；此結構的優點在於融合卷積式類神經網路與遞迴式類神經網路的長處: 卷積神經網路能夠捕捉語句內局部的 *N*-連語言樣式(*N*-gram Patterns)，而遞迴式類神經網路則可以捕捉長距離的語句與語句之間的語意和結構關係。透過上述依序結合使用的方式，可以彌補單純將文本視為一個詞彙序列(忽略語句間結構資訊)並僅單獨使用卷積式類神經網路或遞迴式類神經網路來產生文本固定長度向量表示的不足處。更詳細地說，在實作時輸入文本中的每一句語所含的詞彙會先經過詞嵌入(Word Embedding)轉換方法，例如 Skip-Gram 和 CBOW，轉換成固定長度的詞彙語意向量[16]。再者，每一句語句中詞彙的語意向量會經過含三層卷積式類神經網路，每一層含有不同數量和大小(由大到小)的核函數(Kernel Function)以萃取不同階段的語句內部的語意向量表示。

語句會經過三層核函數由大到小的卷積式類神經網路以取得對每個句子的向量表達(Sentence Representation)，再將其送入雙向長短期記憶網路(Bi-directional LSTM)，雙向長短期記憶網路能夠達成捕捉每一句對其他句之間的關係的目標，最後得出兩個方向相反的內文向量，將其串接之後，送入最後的前饋網路(Feed Forward Network)做出最後的預測。

在文本編碼及前饋網路之間，會考慮是否加入文本的一般特徵(表一)，此些一般特徵由考慮文本的不同面向統計而來。一般特徵會在得到對整個文本的特徵表達之後與之串接，得到一個包含同時具有文本語句結構資訊及一般統計特徵之文本的語意向量，以期作出最後的文本可讀性難度預測。

表一、文本之一般特徵共 15 個

	特徵名稱
詞相關特徵	詞數、動詞數、領域詞頻對數平均、負向連接詞數、中筆劃字元數、副詞數、實詞數、低筆劃字元數、二字詞數、複雜語意類別數、正向連接詞數、難詞數
句相關特徵	單句數比率、複雜結構句數、複雜語意類別句子數



圖一、階層式編碼之文本可讀性預測模型架構圖

三、實驗設定與結果

(一)、實驗材料

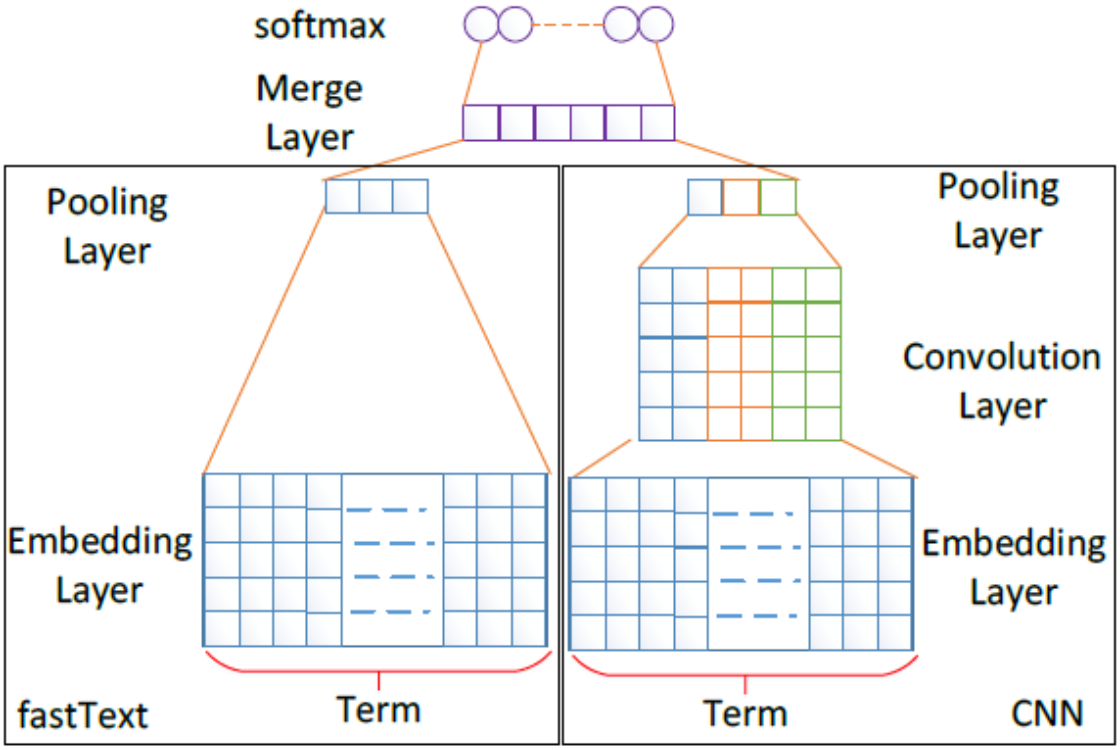
本研究材料選自 98 年度臺灣 H、K、N 三大出版社所出版的 1-12 年級審定版的國語科、社會科和自然科等三個領域的教科書全部共計 4,648 篇[17]，各版本教科書均經由專家根據課程綱要編制而成，其實驗材料的年級分佈如表一所示，模型的目標則是預測輸入屬於 12 個年級中何者。

表一、實驗材料在各年級的數量分佈

年級	1	2	3	4	5	6	7	8	9	10	11	12
國文科	24	67	61	71	69	70	37	34	28	84	41	47
自然科	0	0	72	67	67	62	172	175	157	211	355	295
社會科	0	0	80	74	85	81	389	407	325	340	331	270

(二)、實驗結果

我們的結果顯示在表二中。在表二中，我們將結果與[17]進行比較。在[17]中提出了一種結合了快速文本(fastText)和 CNN 的混合架構。[17]中提到，若在訓練可讀性模型的過程中可以同時融合快速文本(fastText)和 CNN 這兩種演算法，其訓練可讀性模型的過程中就可以相互分享資訊，為可讀性模型帶來更豐富的資訊去評估文本的可讀性。基於這個想法，快速文本(fastText) [18]用作學習可用於對語義進行分類的功能的角色。利用類神經網路來融合卷積神經網路及快速文本兩種不同的表示學習演算法，其示意圖如圖二所示，在[17]的可讀性模型的訓練過程中，卷積式類神經網路和快速文本所產生的特徵會以向量的形式在融合層進行相加、相乘、平均或串聯等不同的運算方式進行融合，而融合後的特徵便可以讓可讀性模型在訓練的過程中享有不同表徵學習法所帶來資訊。



圖二、融合卷積神網路及快速文本的可讀性模型架構

在表二中，可以觀察到我們的模型表現在沒有加上一般特徵的情況下表現略輸於[17]約 0.5 個百分點，但在加入一般特徵輔助判斷後我們的模型表現在準確率進步了 1.91%，相鄰準確率進步了 2.71%，表現都略優於[17]，可見的一般特徵還是有提供資訊

增進模型判斷的能力，相鄰準確率則是考慮了正確答案及前後一個年級:如果正確答案為 5 年級，則可以容忍模型的預測為 4、5、6 年級。

表二、實驗結果

Readability model	準確率	相鄰準確率
Tseng et al.,(2018)	79.42%	91.59%
Hierarchical Encoding w/o general feature	78.81%	89.64%
Hierarchical Encoding w/ general feature	**80.72%**	**92.35%**

我們還在前饋網路(feed-forward network)中比較不同數量的層數(6,7,8,9)對模型的影響。在表 3 中，我們可以看到，在前饋網路中如果層數更多，我們可以略微提高準確性。但太多層沒有對模型表現有更多的增進，可能還會有過度擬合的疑慮。

表三、前饋網路不同層數之結果

層數	準確率	相鄰準確率
6	79.12%	91.65%
7	79.82%	91.24%
8	**80.72%**	**92.35%**
9	80.04%	91.86%

四、結論

本研究結合卷積神經網路及遞歸神經網路並兼取兩者之長處，訓練出一個能夠抽取更深層特徵的可讀性模型。實驗結果顯示，此模型不考慮文本的一般特徵就已經可以達到不錯的模型效能，但若在模型中考慮一般特徵則能夠在得到些微的提升。在未來，我們希望能夠不只用遞歸神經網路來取得句與句之間的關係，而是更精細的計算句子間的相關性，並且探討相關性是否能可讀性研究有所幫助，並考慮使用更細緻的方法來結合文本的各種特徵，使的模型能夠更全面的評估文件的可讀性難易程度。

致謝

本論文之研究承蒙行政院科技部研究計畫 (MOST 105-2221-E-003-018-MY3 和 MOST 107-2221-E-003-013-MY2、MOST 108-2221-E-003-005-MY3 和 MOST 108-2634-F-008 -004 -) 之經費支持，謹此致謝。

參考文獻

[1] E. Dale and J. S. Chall, “The concept of readability,” Elementary English, vol. 26, pp. 19–26, 1949.

[2] G. R. Klare, “Measurement of Readability,” 1963.

[3] G. R. Klare, “The measurement of readability: useful information for communicators,” ACM Journal of Computer Documentation (JCD), vol. 24, pp. 107-121, 2000.

[4] G. H. McLaughlin, “SMOG grading: A new readability formula,” Journal of reading, vol. 12, pp. 639–646, 1969.

[5] B. A. Lively and S. L. Pressey, “A method for measuring the vocabulary burden of textbooks,” Educational administration and supervision, vol. 9, pp. 389–398, 1923.

[6] M. Vogel and C. Washburne, “An objective method of determining grade placement of children's reading material,” The Elementary School Journal, pp. 373–381, 1928.

[7] J. S. Chall and E. Dale, Readability Revisited: The new Dale-Chall Readability Formula, Brookline Books, 1995.

[8] A. C. Graesser, M. Singer, and T. Trabasso, “Constructing inferences during narrative text comprehension,” Psychological Review, vol. 101, pp. 371, 1994.

[9] K. Collins-Thompson, “Computational assessment of text readability: A survey of current and future research,” International Journal of Applied Linguistics, vol. 165, pp. 97–135, 2014.

[10] B. Bruce, A. Rubin, and K. Starr, “Why readability formulas fail,” IEEE Transactions on Professional Communication, pp. 50-52, 1981.

[11] S. E. Petersen and M. Ostendorf, “A machine learning approach to reading level assessment,” Computer Speech & Language, vol. 23, pp. 89–106, 2009.

[12] L. Feng, M. Jansche, M. Huenerfauth, and N. Elhadad, “A comparison of features for automatic readability assessment,” in Proceedings of the 23rd International Conference on Computational Linguistics: Posters, 2010, pp. 276–284.

[13] Y. T. Sung, J. L. Chen, J. H. Cha, H. C. Tseng, T. H. Chang, and K. E. Chang, “Constructing and validating readability models: the method of integrating multilevel linguistic features with machine learning,” Behavior research methods, vol. 47, pp. 340 – 354, 2014.

[14] Ilya Sutskever, Oriol Vinyals, Quoc V. Le, “Sequence to Sequence Learning with Neural Networks”, in Proc. NIPS, Montreal, CA , 2014

[15] Yen-Chun Chen, Mohit Bansal, “Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting”, in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics(ACL’2018), Pages 675-686

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

[17] Hou-Chiang Tseng, Berlin Chen, Yao-Ting Sun, “Exploring Combination of FastText and Convolutional Neural Networks for Building Readability Models”, in Proceedings of the the 2018 Conference on Computational Linguistics and Speech Processing, Pages 116-125

[18] A.Joulin, E.Grave, P.Bojanowski, and T.Mikolov, “Bag of tricks for efficient text classification,”arXiv preprint arXiv:1607.01759. 2016