Computational Linguistics and Chinese Language Processing
Vol. 23, No. 2, December 2018, pp. 67-84

未登錄詞之向量表示法模型於中文機器閱讀理解之應用[1]

An OOV Word Embedding Framework for Chinese Machine Reading Comprehension

羅上堡*、李青憲+、涂家章+、陳冠宇*

Shang-Bao Luo, Ching-Hsien Lee, Jia-Jang Tu and Kuan-Yu Chen

摘要

在使用深度學習(Deep Learning)方法於自然語言處理的問題時，我們通常會先將每一個詞以一個相對應的詞向量(Word Embedding)表示，再輸入至各式神經網路模型。當遭遇未登錄詞(Out-of-Vocabulary, OOV)的問題時，最常見的處理方式是略去該未登錄詞、以一個零向量表示或是用一個隨機產生的向量表示這個未登錄詞。就我們所知，在目前的研究裡，似乎仍未有一套合理且快速的做法，用於產生未登錄詞的詞向量表示法，並進一步地探索未登錄詞的詞向量對於任務成效的影響性。因此，本論文提出一套新穎的詞向量表示法學習技術，其目標是為未登錄詞產生一個較為合理且可靠的低維度向量表示法；除此之外，我們將進一步地把此一技術運用於中文機器閱讀理解任務之中，探究未登錄詞對於中文機器閱讀理解任務之影響，並驗證本論文所提出的詞向量表示法學習技術之成效。

關鍵詞：自然語言處理、詞向量表示法、未登錄詞、機器閱讀理解

---

[1] This research was partially supported by the Project H367B83300 (ITRI) under the sponsorship of the Ministry of Economic Affairs, Taiwan, R.O.C.

* 國立臺灣科技大學系資訊工程系
Department of Computer Science & Information Engineering, National Taiwan University of Science and Technology
E-mail: {M10615012, kychen}@mail.ntust.edu.tw

+ 工業技術研究院巨資中心
Computational Intelligence Technology Center, Industrial Technology Research Institute
E-mail: {C.H.Lee, santu}@itri.org.tw

Abstract

When using Deep Learning methods in NLP-related tasks, we usually represent a word by using a low-dimensional dense vector, which is named the word embedding, and these word embeddings can then be treated as feature vectors for various neural network-based models. However, a major challenge facing such a mechanism is how to represent OOV words. There are two common strategies in practiced: one is to remove these words directly; the other is to represent OOV words by using zero or random vectors. To mitigate the flaw, we introduce an OOV embedding framework, which aims at generating reasonable low-dimensional dense vectors for OOV words. Furthermore, in order to evaluate the impact of the OOV representations, we plug the proposed framework into the Chinese machine reading comprehension task, and a series of experiments and comparisons demonstrate the good efficacy of the proposed framework.

Keywords: Natural Language Processing, Word Embedding, Out-of-vocabulary, Machine Reading Comprehension

1. 緒論 (Introduction)

機器閱讀理解(Machine Reading Comprehension, MRC)是一個自然語言處理(Natural Language Processing, NLP)領域中相當重要的任務，其目標是希望讓機器像人類一樣進行文本閱讀，並根據對該文本之理解，進而回答相關的問題。讓電腦幫助人類在大量文本中找到想要的答案，可以減輕資訊獲取的成本、加速資訊處理的速度、以及提升資訊的利用率。進一步地，如果電腦能具備相當高水準的閱讀理解能力，許多應用將會有更進一步的發展，例如問答(Question Answering)、對話系統(Dialogue System)以及搜尋引擎(Search Engine)等。因此機器閱讀理解不論在學術界或產業界都有極高的研究價值。

目前機器閱讀理解的研究主要有完型填空(Cloze Style)與文本段(Text Span)預測等兩種型式。完型填空是去掉文本中的某個詞語，讓系統進行填空，但答案往往是單一的字詞，並不需要對於整段文本進行理解，因此這類型的回答形式較難以延伸應用於實際生活中。有鑑於完型填空之不足，2016 年時，一個大規模的文本段類型數據集 SQuAD (The Stanford Question Answering Dataset) (Rajpurkar, Zhang, Lopyrev & Liang, 2016)應運而生。SQuAD 資料集包含十萬多個問題答案組，文本來至維基百科，因此答案就是維基百科文本中的一個小段落；更明確地，文本段的預測方式為給定文本與問題後，機器需以文本中一個連續的小片段來回答給定的問題。

由於深度學習的蓬勃發展並且在許多領域中取得空前的好成績，目前多數的機器閱讀理解模型皆是建構於深度學習的方法上。基於深度學習之機器閱讀理解模型，主要可區分成五個模塊，嵌入層(Embedding Layer)、嵌入編碼層(Embedding Encoder Layer)、段落問題注意機制(Passage-question Attention)、注意力編碼層(Attention Encoder Layer)以及

輸出層(Output Layer)。嵌入層主要是將每一個詞轉換成一個相對應的詞向量表示法，有些模型會額外加入各式語言特徵(Linguistic Features)來豐富每一個詞的語意或語言資訊；嵌入編碼層目標於探究詞與詞之間的上下文關係，並基於這個資訊，為每一個詞產生一個新的低維度向量表示法；段落問題注意機制是藉由注意力機制(Attention Mechanism)將每一段落（或文本）表示為含有問題意識之段落表達(Question-aware Passage Representation)；注意力編碼層(Attention Encoder Layer)則是堆疊在段落問題注意機制之後，利用段落問題注意機制所產生的詞向量與雙向循環神經網路並搭配各式注意力機制，產生更進階的詞向量表示法；在文本段預測的機器閱讀理解模型裡，輸出層通常採用指針網路(Pointer Network)生成兩個分別代表開始與結束的機率分布，藉由簡單的搜尋演算法，將文本中被預測為答案開始與結束的位置標示出來。值得一提的是，段落問題注意機制可視為機器閱讀理解模型中最重要且關鍵的元件，各式機器閱讀理解模型多半著墨於此層中，並提出各式不同的模型架構來改善機器閱讀理解模型之成效。另外，傳統的機器閱讀理解模型多是以循環神經網路(Recurrent Neural Network, RNN)為主要架構，但由於循環神經網路較難以實現並行運算，因此多數模型皆易遭受訓練時間過久的問題。為解決此一問題，近年來有研究提出一套新穎的模型方法 QANet (Yu *et al.*, 2018)，將機器閱讀理解模型中常見的循環神經網路架構全部捨棄，只採用卷積神經網路(Convolution Neural Network, CNN)與可並行的注意力機制來建立機器閱讀理解模型，不僅大幅降低訓練所耗費的時間，其成效依然相當傑出。

在各式自然語言處理的任務中，未登錄詞是一個基礎且重要的問題。在機器閱讀理解任務裡，各式模型為了減緩此一問題造成的影響，多半在嵌入層裡除了每一個詞的詞向量外，會再額外利用字向量(Character Embedding)來豐富每一個詞的語彙資訊。QANet、BiDAF (Seo, Kembhavi, Farhadi & Hajishirzi, 2016)、jNet (Zhang *et al.*, 2017)、MEMEN (Pan *et al.*, 2017)與 ReasoNet (Shen, Huang, Gao & Chen, 2017)等，皆是利用卷積神經網路對一連串的字向量提取字與字之間相連的結構資訊(Kim, Jernite, Sontag & Rush, 2016)，再轉換成固定維度的向量表示法；R-NET (Wang, Yang, Wei, Chang & Zhou, 2017)與 S-NET (Tan *et al.*, 2017)則是利用雙向循環神經網路提取字向量的前後規則資訊；RMR (Reinforced Mnemonic Reader) (Hu, Wei, Mao & Chikina, 2017)、Conductor-net (Liu *et al.*, 2017)、V-Net (Wang *et al.*, 2018)、FastQA (Weissenborn, Wiese & Seiffe, 2017)與 Smartnet (Chen *et al.*, 2017)是將字向量直接與詞向量進行串接，形成新的向量表示法。綜觀上述各式模型方法，皆是以英文為處理目標，並且以機器閱讀理解任務為導向，在嵌入層中，訓練一組字向量表示法模型，用以彌補未登錄詞在機器閱讀理解任務中所造成的影響。然而，這樣的方式並非實際地為每一個未登錄詞產一個合理且適當的向量表示法，且使其可以應用於各式任務之中。有鑑於此，本論文旨於提出一套新穎的詞向量表示法學習技術，為每一個未登錄詞產生一個較為可靠的詞向量表示法，並且，本論文進一步地將此一技術運用於中文機器閱讀理解任務之中，實驗結果顯示，結合本論文提出之詞向量表示法學習技術，可以有效提升中文機器閱讀理解任務之成效。

2. 相關方法 (Related Methods)

2.1 詞向量表示法 (Word Embedding)

詞向量表示法(Mikolov, Chen, Corrado & Dean, 2013)是深度學習於自然語言處理中相當成功的應用之一，其目的是將每一個詞以一個低維度的向量表示之。常見的詞向量表示法模型有連續型詞袋模型(Continuous Bag-of-Words, CBOW)、略詞模型(Skip-gram)與全局向量模型(GloVe) (Pennington, Socher & Manning, 2014)。

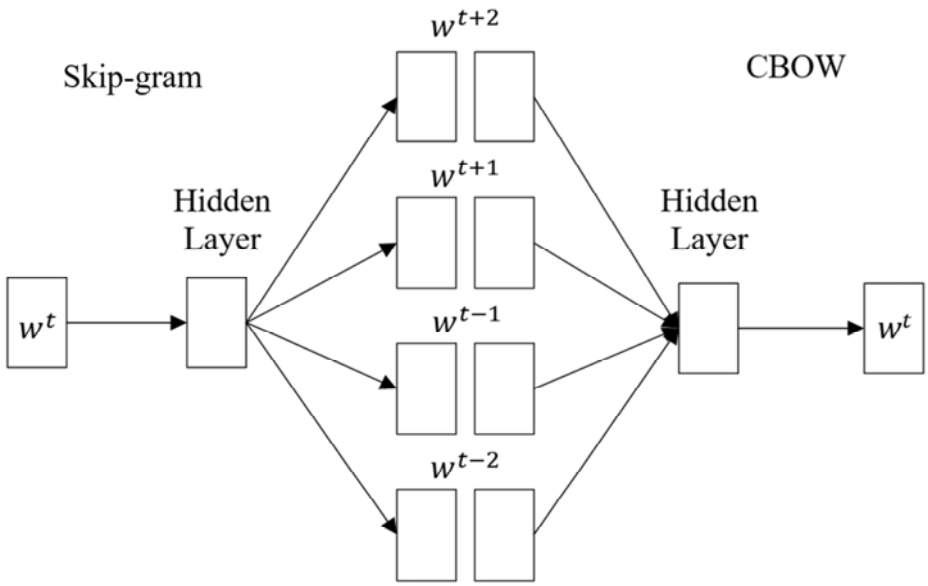

圖 1. 略詞模型(Skip-gram)與連續型詞袋模型(CBOW)示意圖
[Figure 1. Illustrations of Skip-gram and CBOW models.]

詞向量表示法學習的起源可以追溯至 2003 年被提出的神經網路語言模型(Bengio, Ducharme, Vincent & Jauvin, 2003)，其使用前饋式神經網路(Feed-forward Neural Network)來建立 N 連語言模型，在這個模型架構下，自然地獲得了每一個詞的向量表示法，通常稱之為詞向量。衍生至今，目前多數的研究著眼於將每一個字、詞用一個連續數值的向量(Distributed Vector)來表示。不同於神經網路語言模型以建立一個 N 連語言模型為目標，而每一個詞的向量表示法僅是模型建立過程的副產物，連續型詞袋模型的目標即是為每一個詞建立專屬的詞向量。連續型詞袋模型的架構類似於前饋式神經網路，但為了節省運算時間與參數量，省略了前饋式神經網路中的隱藏層，不僅打破了過去各式神經網路模型訓練耗時的缺點，並大幅提升此表示法學習的實用性，其架構如圖 1 所示。當給定一連串的文字序列：$w^1, w^2, \dots, w^T$，連續性詞袋模型的目標函數(Objective Function)是最大化每一個詞出現的可能性：

$$\sum_{t=1}^{T} logP(w^t | w^{t-c}, \dots, w^{t-1}, w^{t+1}, \dots w^{t+c}) \quad (1)$$

其中，c 表示相對於詞w^t的左右窗函數(Window Function)，T 表示文字序列的總長度，而條件機率$P(w^t|w^{t-c},\dots,w^{t-1},w^{t+1},\dots w^{t+c})$則為：

$$P(w^t|w^{t-c},\dots,w^{t-1},w^{t+1},\dots w^{t+c}) = \frac{\exp(v_{\bar{w}^t}\cdot v_{w^t})}{\sum_{i=1}^{V}\exp(v_{\bar{w}^t}\cdot v_{w_i})} \quad (2)$$

其中，V 為辭典的總詞數，v_{w^t}是詞w^t的詞向量，$v_{\bar{w}^t}$則是出現在詞w^t左右相鄰詞的詞向量之加權平均向量：

$$v_{\bar{w}^t} = \sum_{\substack{j=-c \\ j\neq 0}}^{c} \alpha_j v_{w^{t+j}} \quad (3)$$

其中，α_j為與距離有關的權重值。連續型詞袋模型的概念源起於分散式假說(Distributional Hypothesis)，也就是語意相近的詞通常其左右相鄰的詞不會差異很大，因此，建立詞向量時左右相鄰詞的詞向量是很重要的參考。

略詞模型的訓練目標函數則與連續型詞袋模型相反。當給定一連串的文字序列：$w^1, w^2, \dots, w^T$，略詞模型的目標函數為：

$$\sum_{t=1}^{T}\sum_{\substack{j=-c \\ j\neq 0}}^{c} log P(w^{t+j}|w^t) \quad (4)$$

其中，條件機率$P(w^{t+j}|w^t)$可表示為：

$$P(w^{t+j}|w^t) = \frac{\exp\left(v_{w^{t+j}}\cdot v_{w^t}\right)}{\sum_{i=1}^{V}\exp\left(v_{w_i}\cdot v_{w^t}\right)} \quad (5)$$

V 為辭典的總詞數，v_{w^t}與$v_{w^{t+j}}$分別表示詞w^t與w^{t+j}的詞向量，其模型架構如圖 1 所示。雖然連續型詞袋模型以及略詞模型在模型架構已相當簡化，在實作的過程中，前人的研究更提出了階層式軟性最大化演算法(Hierarchical Soft-max Algorithm) (Mnih & Hinton, 2009)以及負取樣演算法(Negative Sampling Algorithm) (Mikolov, Sutskever, Chen, Corrado & Dean, 2013)來加速模型參數（即詞向量）的估算過程。

由於連續型詞袋模型以及略詞模型在訓練的過程中，僅考慮短距離的詞彙規則關係，全局向量模型認為在求取詞向量時，應當考慮的是整個訓練語料中，詞與詞之間的相互關係，並且詞與詞之間的關係不應該以最大化預測機率來描述，應該考慮詞對(Word Pair)與詞對之間的比例關係，綜合以上考量，全局向量模型的目標函數為：

$$\sum_{i=1}^{V}\sum_{j=1}^{V} f\left(X_{w_iw_j}\right)(v_{w_i}\cdot v_{w_j} + b_{w_i} + b_{w_j} - log X_{w_iw_j})^2 \quad (6)$$

同樣地，V 為辭典的總詞數，v_{w_i}與v_{w_j}分別表示詞w_i與w_j的詞向量，$X_{w_iw_j}$代表詞w_i與w_j在訓練語料中共同出現的次數，$f(\cdot)$為一個單調的平滑函數，用來調整每一個詞對在訓練過程中的影響（重要）性，而b_{w_i}則為詞w_i的基數(Bias)。

除了經典的詞向量表示法模型外，上下文向量表示法(Context Vectors, CoVe) (McCann, Bradbury, Xiong & Socher, 2017)借鑑於遷移學習(Transfer Learning)的概念，將

各式詞向量表示法模型所求得的詞向量做為預訓練(Pre-trained)的模型參數，接著利用序列至序列(Sequence-to-Sequence) (Sutskever, Vinyals & Le, 2014)的方式進行機器翻譯(Machine Translation)模型的訓練，藉由最佳化機器翻譯為目標，調適出一組新的詞向量表示法。最後，上下文向量表示法是將新的詞向量與原始的詞向量表示法串接，作為一個新的詞向量表示法。在許多任務上已證實，上下文向量表示法確實可以獲得更好的任務成效。快文向量模型(FastText) (Bojanowski, Grave, Joulin & Mikolov, 2016)則為略詞模型之延伸，不同之處在於，略詞模型是以詞為單位，通過目標單詞來預測其上下文中之其他詞彙的出現機率，而快文向量模型則是以字符為單位，因此，每一個詞彙的向量表示法是詞彙中所有字符向量的平均。

2.2 問答模型 (Question Answering)

有鑑於傳統的機器閱讀理解模型多半採用循環神經網路為基礎，容易遭受訓練時間過長的問題，QANet 改以卷積神經網路為基礎，提出一套嶄新的機器閱讀理解模型，不僅有效地縮短訓練所需的時間，也在許多實驗中，被驗證可獲得相當不錯的任務成效。

QANet 包含五個主要元件：嵌入層(Embedding Layer)、嵌入編碼層(Embedding Encoder Layer)、語境查詢注意力層(Context-query Attention Layer)、模型編碼層(Model Encoder Layer)以及輸出層(Output Layer)。除了在嵌入編碼層與模型編碼層捨棄傳統主流的循環神經網路，改採卷積神經網路外，QANet 亦引入自我注意力機制(Self-attention Mechanism) (Vaswani *et al*., 2017)，用以擷取每一個詞與整個文本中每一個詞之間的關係，並且可以採用平行化的訓練方式，使得訓練速度與推論(Reasoning)的速度更快。QANet 的模型架構如圖 2 所示，值得注意的是，模型中有許多堆疊架構，架構裡的神經網路結構皆是相同的，並且每一層之間皆使用層正規化(Layer Normalization) (Ba, Kiros & Hinton, 2016)與殘差(Residual Network) (He, Zhang, Ren & Sun, 2016)技術來穩定訓練過程。除此之外，為了增加泛化能力，QANet 的文本與問題編碼器是共享權重的。

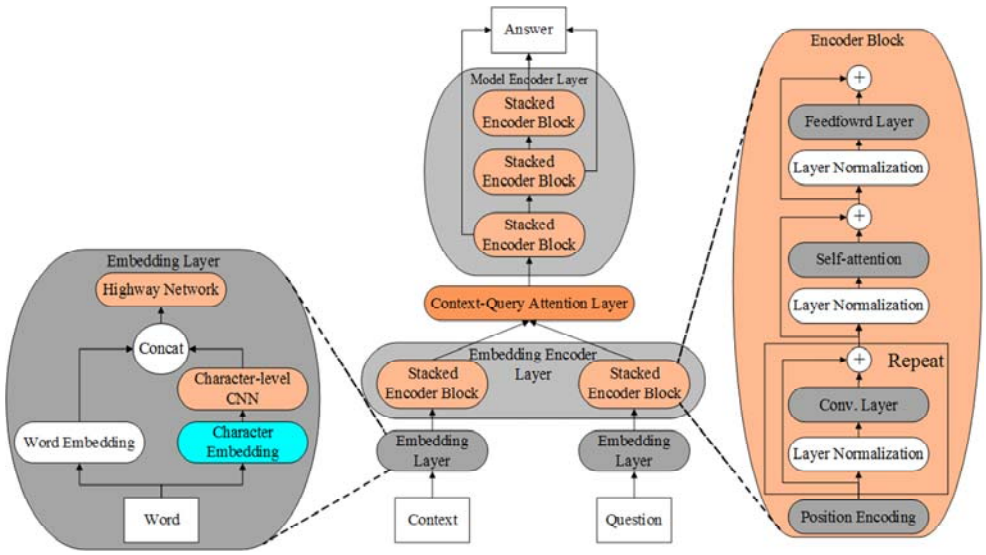

圖 2. QANet 模型示意圖
[Figure 2. Illustration of QANet Model]

3. 新穎的詞向量表示法模型技術 (Novel Word Embedding Model Technique)

在使用深度學習方法於自然語言處理的問題時，我們通常會先將每一個詞以一個相對應的詞向量進行表示，作為各式神經網路模型之輸入。當遭遇未登錄詞的問題，最常見的處理方式是略去未登錄詞、以一個零向量表示之或是用一個隨機產生的向量表示這個未登錄詞。為此，本論文提出一套新穎的詞向量表示法學習技術，為未登錄詞產生一個較為可靠的詞向量表示法，並進一步地將此一技術運用於中文機器閱讀理解任務之中，探究未登錄詞對於中文機器閱讀理解任務之影響，並驗證本論文提出的詞向量表示法學習技術之成效。

中文是字符語言(Character-based Language)，通常每一個字都有其獨特的意義，或是字符的形狀與其所對應的事物有關聯。詞(Word)通常是由兩個以上的字(Character)所組成，用以表達某一種事、物或現象，而只用一個字所成的詞通常稱之為單字詞。由於新字的生成是複雜且繁瑣的，所以我們假設所有的中文字符是固定且已知的$V_C = \{c_1, c_2, \dots, c_{|V_C|}\}$。藉由一份文字語料，傳統的各式詞向量表示法模型可用來為字典 V 中的每一個詞 $\{w_1, w_2, \cdots, w_{|V|}\}$產生一個相對應的低維度向量表示法$\{\{v_{w_1}, v_{w_2}, \cdots, v_{w_{|V|}}\}$。由於每一個詞所欲表達的事、物或現象常與詞中字的意義或字與字之間的排列順序有關，因此我們可以利用每一個詞的詞向量為學習目標，為每一個字求取一個相對應的字向量表示法(Chen, Wang & Chen, 2015)。之後，當遇到未登錄詞時，就可以透過字向量表示法取得此一未登錄詞的詞向量表示法。

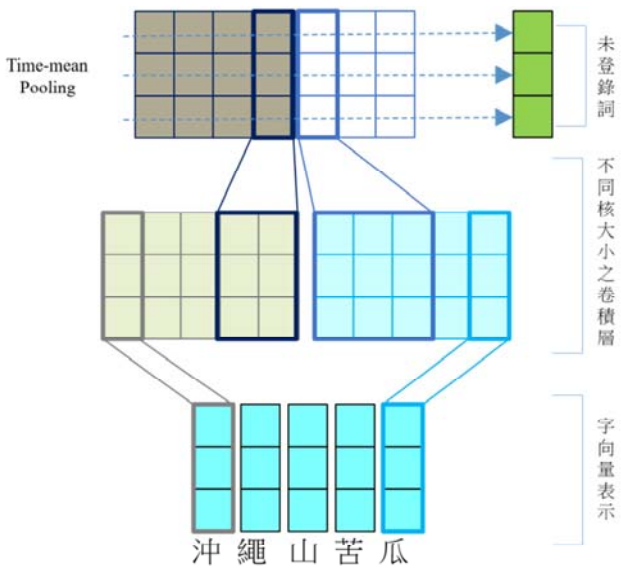

圖 3. 基於卷積神經網路的未登錄詞之向量表示法模型
[Figure 3. Convolutional Neural Network-based Out-of-Vocabulary Embedding Model, COEM]

為了達成此一目的，我們嘗試提出兩種模型架構：基於卷積神經網路的未登錄詞之向量表示法模型(Convolutional Neural Network-based Out-of-Vocabulary Embedding Model, COEM)，如圖 3 所示；基於循環神經網路的未登錄詞之向量表示法模型(Recurrent Neural Network-based Out-of-Vocabulary Embedding Model, ROEM)，如圖 4 所示。在基於卷積神經網路的未登錄詞之向量表示法模型中，我們首先將每一個詞以數個相對應的字向量表示之，這些字向量先經由全連接層進行線性轉換後，接著透過不同核大小(Kernel Size)的卷積神經網路，探究字符間特定距離的排列順序資訊，最後利用池化層(Pooling Layer)，將各式特徵取平均後作為輸出。在基於循環神經網路的未登錄詞之向量表示法模型裡，每一個詞同樣以一連串的字向量表示之，接著藉由雙向循環神經網路抽取字與字之間長距離的意義關係，再將雙向循環神經網路最後一個時間點所產生的隱藏層輸出進行串接，最後經過全連接層產生輸出結果。

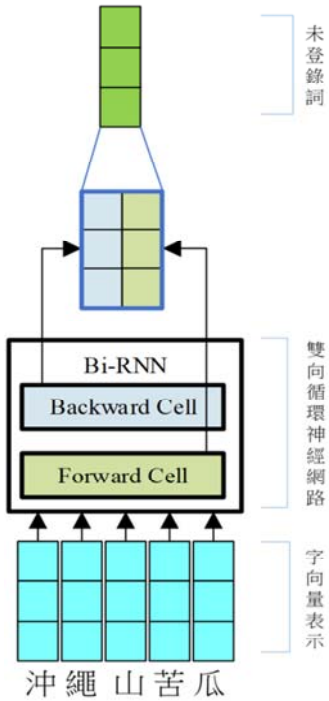

圖 4. 基於循環神經網路的未登錄詞之向量表示法模型
[Figure 4. Recurrent Neural Network-based Out-of-Vocabulary Embedding Model, ROEM]

總言來說，本論文提出的未登錄詞之向量表示法模型訓練可分為兩階段：第一階段為利用經典的各式詞向量表示法模型求取一組詞向量；在第二階段中，我們以第一階段所獲得的詞向量為目標，訓練本論文所提出的未登錄詞之向量表示法模型，其中包含神經網路的模型參數以及一組字向量表示法。更明確地，我們將目標詞向量與模型之輸出向量進行目標函示定義為：

$$Loss = \sum_{i=1}^{V}(v_{w_i} - \varphi(v_{c^{i,1}}, \cdots, v_{c^{i,|w_i|}}))^2 \tag{7}$$

其中，v_{w_i}是詞w_i的在第一階段所獲得的詞向量表示法，$|w_i|$表示詞w_i內所含字的個數，$v_{c^{i,j}}$表示詞w_i中第j個字的字向量，$\varphi(\cdot)$為本論文所提出的未登錄詞之向量表示法模型。本論文藉由此目標函數來訓練未登錄詞之模型，並且訓練流程架構如圖 5 所示，其中輸入為該詞之字向量表達，輸出為未登錄詞之詞向量。

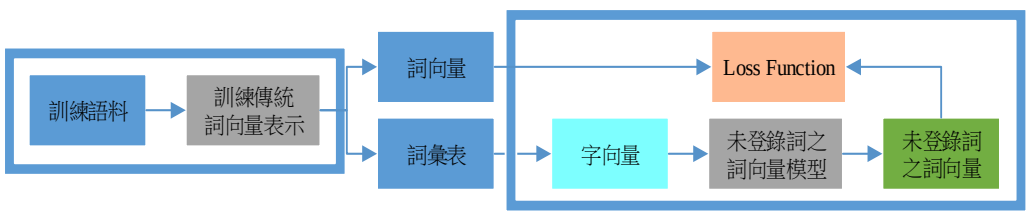

圖 5. 未登錄詞之詞向量模型整體訓練示意圖
[Figure 5. Training Flowchart for the Proposed Framework]

4. 實驗設定與結果 (Experimental Settings and Results)

4.1 實驗設定 (Experimental Settings)

本論文蒐集批踢踢實業坊(PTT)與中央社新聞(CNA)作為文本資料，用於訓練各式詞向量表示法模型(即連續型詞袋模型(CBOW)、略詞模型(Skip-gram) 與全局向量模型(GloVe))，詞向量的維度大小設定為 300 維，並將詞頻小於 5 的詞彙捨去。在本論文所提出之新穎的詞向量表示法模型（即 COEM 與 ROEM）裡，同樣將字向量表示法的維度設定為 300 維，因此，每個詞都會重新以 16 個 300 維的字向量表示之，若某一個詞長度超過 16 個字，則將過長的字捨去，相反地，長度不足 16 個字的詞，則進行貼補(Padding)。在基於 COEM 中，全連接層輸入與輸出皆為 300 維，卷積層共包含四組分別核大小為 2、4、6 與 8，跨步(Stride)大小為 1，過濾器個數(Filter Size)皆設定為 300。在 ROEM 中，雙向循環神經網路的隱藏層大小設定為 300，全連接層輸出大小為 300。此外，除了循環神經網路激活函數採用雙曲函數(Tanh)外，其它激活函數皆採用 TanhShrink。中文文本斷詞是採用結巴(Jieba) (Sun, 2012)作為斷詞器。在訓練新穎的詞向量表示法模型時，我們採用 Adam (Kingma & Ba, 2014)演算法做為求取參數的優化器。所有實驗皆以 python 3.5.2 與 Tensorflow1.6.0 (Abadi *et al*., 2016)套件實現。

為了比較傳統詞向量表示法與本論文提出的未登錄詞詞向量表示法模型於中文機器閱讀理解任務之成效，我們使用台達電閱讀理解資料集(Delta Reading Comprehension Dataset, DRCD) (Shao, Liu, Lai, Tseng & Tsai, 2018)進行各式實驗與觀察，並將此資料集切分為訓練集、發展集與測試集，其統計資訊如表 1 所示，其中 DRCD 在我們蒐集之文本資料上，未登錄詞在訓練集、發展集與測試集各佔 65.50%、46.19%與 46.20%。由於本論文著眼於不同詞向量表示法在中文機器閱讀理解任務的成效，因此我們將基於前人在 SQuAD 上所提出之機器閱讀理解模型架構 QANet，把各式詞向量或其結合作為輸入，驗證其成效。在 QANet 中，隱藏層大小設定為 96，多重注意力機制(Multi-attention head number)數量為 2，訓練資料批次大小為 32，總期次數為 75000。值得一提的是，各式詞向量（即連續型詞袋模型(CBOW)、略詞模型(Skip-gram)與全局向量模型(GloVe）與本論文所提出之新穎的詞向量表示法模型皆不會在訓練閱讀理解模型時再被調整、更新，但 QANet 架構中的字向量表示法，則會隨著模型訓練而改變。

表 1. 台達電閱讀理解資料集之統計資訊
[Table 1. Statistics on Delta Reading Comprehension Dataset]

訓練集 Training Set			發展集 Development Set			測試集 Test Set		
文本數	問題數	OOV Ratio	文本數	問題數	OOV Ratio	文本數	問題數	OOV Ratio
1,960	26,936	65.50%	383	3,524	46.19%	383	3,524	46.20%

在中文機器閱讀理解的實驗中，我們採用標準的 F1 與 EM(Exact Match)分數作為評估指標。EM 指標是當模型預測之答案與正確答案完成一致，才會獲得分數；F1 指標是

資訊檢索領域常用的評分方法，是基於精確度(Precision)與召回率(Recall)計算而得。

4.2 實驗結果 (Experimental Results)

在第一組實驗中，我們首先比較各式基於 QANet 與傳統詞向量表示法模型之基準系統(Baseline Systems)，發展集與測試集實驗結果分別詳列於表 2 與表 3。在基準系統中，我們將傳統詞向量作為 QANet 的輸入，而未登錄詞可分別以零向量或隨機向量來表示，分別標示為 Baseline(Zero)與 Baseline (Rand)。除了以詞向量表示法作為輸入外，我們進一步地將字向量亦加入模型之中，其實驗結果則標註為 Baseline(Zero)＋Char 與 Baseline (Rand)＋Char。除了標準的基準系統外，我們亦嘗試將訓練文本資料切割成一個個的字，接著分別利用連續型詞帶模型、略詞模型以及全局向量模型訓練一組字向量表示模型，因此未登錄詞的詞向量表示法則為該詞中所有字向量表示法的平均。此一系統可以視為一個強健性基礎系統，我們標示為 Strong Baseline。當然字向量亦可加入模型之中，其實驗結果則標註為 Strong Baseline＋Char。首先，結合傳統詞向量表示法模型與 QANet（即 Baseline(Zero)與 Baseline(Rand)），不論是使用連續型詞袋模型、略詞模型或全局向量模型，都可獲得超過 5%與 7%的 F1 與 EM 分數。當我們進一步地將字向量表示法也加入模型中後，不論使用何種 F1 與 EM 分數作為評估指標，都可以進一步地獲得效能的提升。另外，我們發現在 F1 上，略詞模型比連續型詞袋模型和全局模型有更好的表現；然而在 EM 評估標準上，則是連續型詞袋模型與略詞向量有更好之結果。我們還發現在於為每個未登錄詞產生隨機向量的做法，在不考慮字向量表示的時候，是有實質提升的作用，但在加入字向量表示與零向量的作法可以發現，結果都會略低於零向量的效能。

表 2. 運用各式詞向量表示法模型於基礎系統中之機器閱讀理解任務成效(發展集)
[Table 2. Experimental Results on Development Set with Respect to Various Word Embedding Methods and Baseline Systems]

Development Set	CBOW		Skip-gram		GloVe	
	F1	EM	F1	EM	F1	EM
Baseline(Zero)	71.34%	55.34%	70.82%	55.59%	70.44%	53.43%
Baseline(Zero)+Char	**80.24%**	**68.00%**	**80.73%**	**68.00%**	**80.58%**	**67.25%**
Baseline(Rand)	76.92%	63.49%	76.97%	62.88%	75.86%	61.83%
Baseline(Rand)+Char	79.46%	66.62%	79.71%	66.20%	78.44%	64.56%
Strong Baseline	78.72%	64.65%	79.20%	65.43%	79.44%	65.37%
Strong Baseline+Char	**79.84%**	**67.25%**	**79.77%**	**66.03%**	**80.60%**	**66.09%**

表 3. 運用各式詞向量表示法模型於基礎系統中之機器閱讀理解任務成效(測試集)
[Table 3. Experimental Results on Test Set with Respect to Various Word Embedding Methods and Baseline Systems]

Test Set	CBOW		Skip-gram		GloVe	
	F1	EM	F1	EM	F1	EM
Baseline(Zero)	70.72%	54.85%	70.26%	54.94%	69.42%	53.21%
Baseline(Zero)+Char	**79.53%**	**67.22%**	**80.10%**	**67.54%**	**80.07%**	**66.69%**
Baseline(Rand)	76.14%	62.60%	77.29%	63.12%	77.79%	63.73%
Baseline(Rand)+Char	79.32%	66.46%	79.84%	66.41%	79.32%	66.37%
Strong Baseline	78.26%	64.36%	79.76%	66.29%	78.92%	64.59%
Strong Baseline+Char	**79.88%**	**66.40%**	**80.79%**	**67.85%**	**79.94%**	**66.92%**

接著，我們探討本論文所提出之未登錄詞詞向量表示法模型之成效。這組實驗共分為三種輸入方式來進行實驗，如圖 6 所示。首先，我們使用本論文提出之 COEM 與 ROEM 為每個詞（包括登錄詞與未登錄詞）產生對應的詞向量，做為機器閱讀理解模型的輸入（實驗結果標註為 COEM 與 ROEM）；其二，基於本論文提出之 COEM 與 ROEM 為每個詞(包括登錄詞與未登錄詞)產生的對應詞向量外，再加上字向量，一起做為機器閱讀理解模型的輸入（實驗結果標註為 COEM+Char 與 ROEM+Char）；接著，我們將傳統詞向量表示法與本論文提出之未登錄詞向量表示法結合，其作法是將未登錄詞使用 COEM 與 ROEM 來產生詞向量，而登錄詞則使用原本詞向量表示法所得的向量（實驗結果標註為 WE+COEM 與 WE+ROEM）；最後，我們進一步地將傳統詞向量表示法與本論文提出之未登錄詞向量表示法結合，並且加上字向量來進行實驗（實驗結果標註為 WE+COEM+Char 與 WE+ROEM+Char）。發展集與測試集的實驗結果如表 4 與表 5 所示。

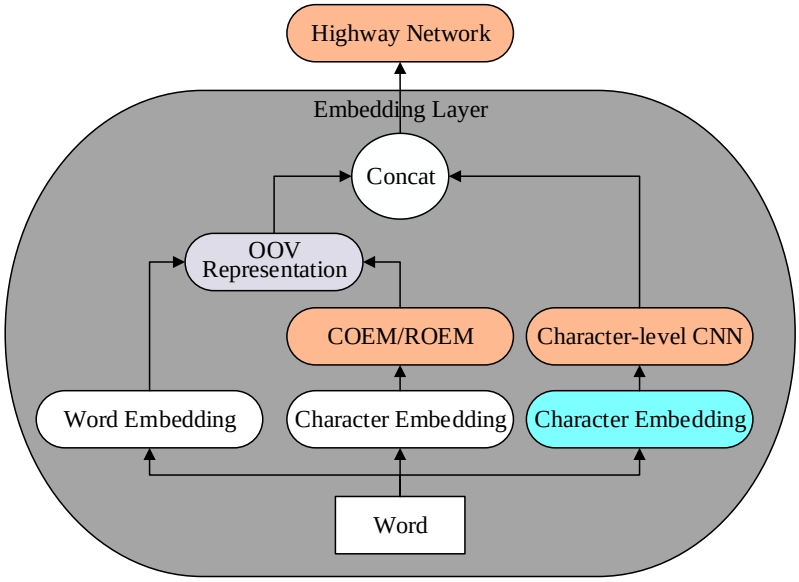

圖 6. QANet 中詞向量、字向量與未登錄詞向量之關係圖
[Figure 6. Relationship Among Word, Character and Out-of-Vocabulary Embeddings in the QANet]

表 4. 結合各式詞向量表示法模型於本論文提出之未登錄詞向量表示法模型在機器閱讀理解任務的成效(發展集)
[Table 4. Experimental Results on Development Set with Respect to Various Word Embedding Methods and the Proposed Framework]

Development Set	CBOW		Skip-gram		GloVe	
	F1	EM	F1	EM	F1	EM
COEM	80.29%	67.19%	78.37%	64.74%	77.89%	63.71%
COEM+Char	80.60%	67.16%	79.91%	66.72%	**80.90%**	67.75%
WE+COEM	**80.82%**	67.94%	80.50%	66.94%	80.31%	67.22%
WE+COEM+Char	80.69%	**68.32%**	**81.28%**	**68.25%**	80.86%	**67.88%**
ROEM	75.44%	60.76%	73.48%	63.59%	74.25%	59.10%
ROEM+Char	79.90%	66.97%	79.89%	67.00%	79.49%	66.31%
WE+ROEM	79.00%	65.72%	78.70%	65.12%	77.78%	64.02%
WE+ROEM+Char	**80.16%**	**67.35%**	**80.82%**	**68.29%**	**81.06%**	**68.00%**

表 5. 結合各式詞向量表示法模型於本論文提出之未登錄詞向量表示法模型在機器閱讀理解任務的成效(測試集)
[Table 5. Experimental Results on Test Set with Respect to Various Word Embedding Methods and the Proposed Framework]

Development Set	CBOW		Skip-gram		GloVe	
	F1	EM	F1	EM	F1	EM
COEM	79.86%	66.79%	77.83%	64.25%	77.36%	63.59%
COEM+Char	80.23%	66.80%	79.33%	66.17%	80.25%	67.13%
WE+COEM	80.12%	67.20%	80.01%	66.32%	79.84%	66.69%
WE+COEM+Char	**80.28%**	**67.93%**	**80.38%**	**67.34%**	**80.45%**	**67.40%**
ROEM	74.65%	59.96%	73.07%	58.46%	73.74%	58.83%
ROEM+Char	79.44%	66.52%	79.26%	66.43%	78.72%	65.66%
WE+ROEM	78.35%	65.07%	78.07%	64.47%	77.00%	63.14%
WE+ROEM+Char	**79.79%**	**66.91%**	**80.19%**	**67.57%**	**80.57%**	**67.76%**

首先，當基於卷積神經網路的未登錄詞模型(COEM)來取代所有詞向量與未登錄詞來進行訓練時，在每個詞向量表示模型上，都可以比基礎系統單用詞向量提升 7%至 8%之效果（請參考表 2 與表 3）；當我們採用基於循環神經網路的未登錄詞模型(ROEM)時，相較於基礎系統只能獲得 3%至 4%之效能提升（請參考表 2 與表 3），但其他實驗結果成效

皆不顯著，探究可能的原因是循環神經網路並不像卷積神經網路的未登錄詞模型，利用多種不同核大小，抽取字與字之間相鄰的特徵，而是直接一次性的在字與字之間進行雙向掃描，細緻的相鄰資訊較不容易保留。再來，我們進一步地討論將字向量表示法加入模型中（即 COEM+Char 與 ROEM+Char），在各種實驗結果上，皆可獲得一定程度之效能提升。至此，我們可以歸納出，不論是傳統的詞向量表示法模型或本論文提出之未登錄詞詞向量表示法模型，皆屬於學習全域性資訊的特徵，而 QANet 中的字向量是利用訓練閱讀理解模型時一併獲得，可看作是任務導向之特徵向量，因此當我們將這兩種資訊結合，通常可以進一步的提升任務成效。並且可以發現除了基於卷積神經網路的未登錄詞向量模型在連續型詞袋模型可以獲得相差不遠之效果，其餘甚至循環神經網路皆低於基礎模型詞向量加字向量之結果。我們認為這是因為我們將已有的詞向量也以未登錄詞來進行取代，並且模型的學習分布還不夠強健，導致雖然有字向量的幫助下，還不能贏過基礎系統詞向量加字向量之結果。接著，我們探討結合傳統詞向量表示法模型以及未登錄詞詞向量表示法模型（即 WE+COEM 與 WE+ROEM）相較於基礎系統之成效。實驗結果顯示，結合傳統詞向量表示法可獲得更進一步的效能提升，並且，基於卷積神經網路的未登錄詞模型在連續型詞袋模型、略詞模型與全局向量模型，都可以達到甚至超越基礎系統加上字向量之成果。值得一提的是，綜觀實驗結果，我們可歸納出，在多數情況下，基於卷積神經網路的未登錄詞模型會較基於循環神經網路的未登錄詞模型獲得較好的成效，推測其原因，應是由於在中文裡，每一個詞所欲表達的事、物或現象，經常與詞彙中字與字之間的排列順序有關，因此基於卷積神經網路的未登錄詞模型，長於擷取短距離的字與字的排列關係，並將此資訊用於生成未登錄詞的向量表示法，是較合理且有效的。當與強基礎系統（即 Strong Baseline 與 Strong Baseline+Char）進行比較，透過額外訓練字表示的文本資料為每一個未登錄詞產生詞向量之結果，基本上都與基礎系統詞向量加上字向量的效能差不多；甚至 F1 評估上，連續型詞袋模型反而得到更低之分數。最後，我們結合字向量表示法、傳統詞向量表示法模型以及未登錄詞詞向量表示法模型（即 WE+COEM+Char 與 WE+ROEM+Char）相較於基礎系統之結果，多數情況可再進一步獲得效能提升，並且超越基礎系統之最好結果，其中又以略詞模型最為突出。因此，我們可以歸納出，基於循環神經網路的未登錄詞詞向量模型，當使用全局向量模型為訓練目標時，可以獲得比基於卷積神經網路更好的任務成效；而略詞向量模型則是搭配基於卷積神經網路的未登錄詞向量模型可以獲得較好的任務成效。值得一提的是，本論文所採用的台達電閱讀理解資料集，未登錄詞在訓練集中的比例高達 65.50%，在發展集與測試集中則分別為 46.19%與 46.20%，而綜觀上述實驗結果，我們可以發現，未登錄詞的詞向量表示法確實會影響機器閱讀理解任務的成效，因此當利用本論文提出之方法時，可以大幅的改善任務的成效，甚至只需使用 COEM 或 ROEM 產生的詞向量表示法為輸入，就可以與傳統結合詞向量與自向量表示法為輸入的基礎系統擁有相近（甚至更好）的任務成效。

5. 結論與未來展望 (Conclusions and Future Work)

有鑑於自然語言處理中，未登錄詞一直是一個亟待解決的研究議題，本論文提出一套簡單又有效的未登錄詞詞量表示法模型，並藉由中文閱讀理解任務來驗證未登錄詞詞向量表示法模型之成效。實驗結果顯示，未登錄詞的詞向量表示法確實會影響任務的成效，因此，當使用本論文所提出的方法時，在未登錄詞的表達上更加合理與可靠，相較於基礎系統，可以獲得更好的結果。在未來，我們將繼續延伸本論文提出的未登錄詞詞向量表示法模型，並進一步地運用於英文的實驗語料中，也將探討再不同比例之未登錄詞比率之成效。再者，我們亦希望將未登錄詞詞向量表示法模型運用於其他任務之中。

參考文獻 (References)

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., … & Zheng, X. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. In *Proceeding of OSDI'16 Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation*, 265-283.

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. Retrived from arXiv:1607.06450.

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, *3*, 1137-1155.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. Retrieved from arXiv:1607.04606.

Chen, K.-Y., Wang, H.-M., & Chen, H.-H. (2015). A Probabilistic Framework for Chinese Spelling Check. *Transactions on Asian and Low-Resource Language Information Processing (Special Issue on Chinese Spell Checking)*, *14*(4), 15. doi: 10.1145/2826234

Chen, Z., Yang, R., Cao, B., Zhao, Z., Cai, D., & He, X. (2017). Smarnet: Teaching Machines to Read and Comprehend Like Human. Retrived from arXiv:1710.02772.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition(CVPR)*, 770-778. doi: 10.1109/CVPR.2016.90

Hu, M., Peng, Y., Huang, Z., Qiu, X., Wei, F., & Zhou, M. (2017). Reinforced Mnemonic Reader for Machine Reading Comprehension. Retrived from arXiv:1705.02798

Kim, Y., Jernite, Y., Sontag, D., & Rush, A. M. (2016). Character-Aware Neural Language Models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI)*, 2741-2749.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference for Learning Representations*, Retrieved from arXiv:1412.6980.

Liu, R., Wei, W., Mao, W., & Chikina, M. (2017). Phase conductor on multi-layered attentions for machine comprehension. Retrived from arXiv:1710.10504.

McCann, B., Bradbury, J., Xiong, C., & Socher, R. (2017). Learned in translation: Contextualized word vectors. In *Proceedings of Advances in Neural Information Processing Systems*, 6294-6305.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. Retrived from arXiv:1301.3781

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of Advances in neural information processing systems*, 3111-3119.

Mnih, A., & Hinton, G. E. (2009). A scalable hierarchical distributed language model. In *Proceedings of Advances in neural information processing systems*, 1081-1088.

Pan, B., Li, H., Zhao, Z., Cao, B., Cai, D., & He, X. (2017). MEMEN: multi-layer embedding with memory networks for machine comprehension. Retrived from arXiv:1707.09098.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532-1543.

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of empirical methods in natural language processing (EMNLP)*, 2383-2392. doi: 10.18653/v1/D16-1264

Seo, M., Kembhavi, A., Farhadi, A., & Hajishirzi, H. (2016). Bidirectional attention flow for machine comprehension. Retrived from arXiv:1611.01603.

Shao, C. C., Liu, T., Lai, Y., Tseng, Y., & Tsai, S. (2018). DRCD: a Chinese Machine Reading Comprehension Dataset. Retrieved from arXiv:1806.00920.

Shen, Y., Huang, P. S., Gao, J., & Chen, W. (2017). Reasonet: Learning to stop reading in machine comprehension. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1047-1055. doi: 10.1145/3097983.3098177

Sun, J. (2012). 'Jieba'Chinese word segmentation tool.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Proceedings of Advances in neural information processing systems*, 3104-3112.

Tan, C., Wei, F., Yang, N., Du, B., Lv, W., & Zhou, M. (2017). S-net: From answer extraction to answer generation for machine reading comprehension. Retrived from arXiv:1706.04815.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of Advances in Neural Information Processing Systems*, 5998-6008.

Wang, Y., Liu, K., Liu, J., He, W., Lyu, Y., Wu, H., ... & Wang, H. (2018). Multi-Passage Machine Reading Comprehension with Cross-Passage Answer Verification. Retrived from arXiv:1805.02220.

Wang, W., Yang, N., Wei, F., Chang, B., & Zhou, M. (2017). Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, *1*, 189-198. doi: 10.18653/v1/P17-1018

Weissenborn, D., Wiese, G., & Seiffe, L. (2017). Making neural QA as simple as possible but not simpler. Retrived from arXiv:1703.04816.

Yu, A. W., Dohan, D., Luong, M. T., Zhao, R., Chen, K., Norouzi, M., & Le, Q. V. (2018). QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension. Retrived from arXiv:1804.09541.

Zhang, J., Zhu, X., Chen, Q., Ling, Z., Dai, L., Wei, S., & Jiang, H. (2017). Exploring question representation and adaptation with neural networks. In *Proceedings of 2017 3rd IEEE International Conference on Computer and Communications (ICCC)*, 1975-1984. doi: 10.1109/CompComm.2017.8322883