Computational Linguistics and Chinese Language Processing
Vol. 24, No. 2, December 2019, pp. 1-18

基於特徵粒度之訓練策略於中文口語問答系統之應用

A Feature-granularity Training Strategy for Chinese Spoken Question Answering

羅上堡*、陳冠宇*

Shang-Bao Luo and Kuan-Yu Chen

摘要

在口語問答系統(Spoken Question Answering, SQA)中，一個簡單且直覺的作法，是先將一段音訊透過自動語音辨識(Automatic Speech Recognition, ASR)轉換成一連串的辨識文字結果，再輸入給現有各式基於文字的問答系統模型來完成任務需求。然而，這樣的做法通常會遭遇自動語音辨識錯誤(Recognition Errors)的影響，導致問答系統模型的效果不如預期。為了解決此一問題，本論文提出一種基於輸入特徵粒度的訓練策略，其目標是改善自動語音辨識錯誤所造成的效能損失，而且不需要額外模型的需求即可完成。我們將本論文所提出之訓練策略運用於中文口語機器閱讀理解(Machine Reading Comprehension, MRC)任務之中，驗證此一方法對於自動語音辨識錯誤的影響與改善。

Abstract

In spoken question answering, a segment of audio is usually converted into a textual representation through an automatic speech recognition (ASR) system, and then input to a text-based question answering model to generate the answer. However, based on the ASR transcriptions, which usually contain lots of recognition errors, text-based question answering system may produce imperfect results. In order to mitigate the performance gap, in this study, a featured-granularity training strategy is proposed. Accordingly, we evaluate the proposed training strategy on spoken Chinese machine reading comprehension task,

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

which not only demonstrates the capability and ability of the proposed strategy, but several valuable observations can be drawn from the experimental results.

關鍵詞：口語問答系統，語音辨識，特徵粒度，訓練策略。
Keywords: Spoken Question Answering, Speech Recognition, Featured-granularity, Training Strategy

1. 緒論 (Introduction)

機器閱讀理解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，進而回答相關之問題。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，例如問答系統(Question Answering, QA)、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著極高的研究價值。

近年來問答系統已有大量研究與發展，問答系統主要又分為多種形式：基於圖像的問答系統、基於文字的問答系統以及口語問答系統等等。目前基於圖像的問答系統主要的形式為場景理解(Scene Understanding)，其目標是給定一張圖像，讓系統進行物件檢索(Scene Object Retrieval)或場景分割(Scene Segmentation)等任務，已有許多經典的模型(Kingma & Dhariwal, 2018; Karras, Laine & Aila, 2019; Wang, Shen, Guo, Cheng & Borji, 2018)被提出並驗證於 LSUN 數據集(Construction of a Large-scale Image Dataset using Deep Learning with Humans in the Loop) (Yu *et al*., 2015)。基於文字的問答系統，除了傳統完形填空(Cloze Style)與文本段(Text Span)預測外，許多研究紛紛提出各式選擇題(Multiple Choices)和簡答題(Short Answer Questions)的問答模型。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應用於實際生活中。為了彌補完型填空的不足，2016年時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，此一數據集包含十萬多個問題答案組，文本皆為維基百科的文章。為了簡化問題，現今的文本段預測多半是在給定文本與問題後，機器需由文本中找出一個連續片段作為答案輸出，也就是問題的答案指定為文本中的一個片段。除此之外，此類問題漸漸地從單輪(Single-turn)問答的形式至多輪(Multi-turn)問答的方向發展，就形成了對話式的問答系統(Conversational Question Answering)，其中最具代表性的就是 CoQA (A Conversational Question Answering Challenge) (Reddy, Chen & Manning, 2019)與 QuAC (Question Answering in Context) (Choi *et al*., 2018)數據集。另外，選擇題型式的問答系統則是給予機器文章、問題以及多個選項，機器須從這些選項中選擇一個做為答案輸出，RACE (Large-scale ReAding Comprehension Dataset From Examinations) (Lai, Xie, Liu, Tang & Hovy, 2017)是選擇題式問答系統極具代表性的數據集，它是從國中與高中的考題上進行蒐集而成的大型數據集。

綜觀問答系統的發展，基於文本的問答模型，在各種題型上，都有愈來愈強健的模型陸續在近年被提出(Tang, Cai & Zhuo, 2019; Wang, Yu, Jiang & Chang, 2018; Zhang *et al*., 2019; Ran, Li, Hu & Zhou, 2019)，但是口語問答系統的研究則是較少被探討的領域。

在口語問答系統中，一個簡單且合理的作法是將口述內容(Spoken Content)透過自動語音辨識轉寫成文本後，輸入給基於文字的各式問答系統模型(Shiang, Lee & Lee, 2014)，完成口語問答系統的任務要求。然而，透過自動語音辨識轉寫而得的文本，通常會伴隨著自動語音辨識錯誤，產生諸如斷詞錯誤、辨識正確率低或關鍵字不存在辨識系統辭典當中等問題，因此這種作法合理且省時省力，但往往發現其效能不如預期。雖然後續有研究提出以次字單元(Subword Unit) (Szöke, 2010; Heerden, Karakos, Narasimhan, Davel & Schwartz, 2017)或是領域調適(Domain Adaptation) (Lee, Chen, Lee, 2019)等各式技術，但其成效亦相當有限。有鑑於口語問答系統對於未來的重要，本論文針對中文口語問答系統，提出一套基於資料不同粒度單元(Granularity Unit)的訓練策略，簡單地利用辨識出來的文字從粗到細的數據粒度特徵，訓練一套基於現有的文字問答系統，就可以有效地提升口語問答系統之成效。因此，我們並不需要大費周章地從頭建立一套複雜的口語問答模型，就可以有效地提升含有錯誤辨識文字的問答任務之成效。

2. 相關方法 (Related Methods)

傳統基於文字的問答系統裡，QACNN (Query-based Attention CNN) (Liu, Wu, Lee, 2017)與 Co-Matching **(**Zhang *et al*., 2019**)**為選擇題形式的經典模型，QANet (Yu *et al*., 2018)則為文本段預測的經典模型。QACNN 是應用於英文選擇題的問答模型，此模型共包含三個主要部分：相似映射層(Similarity Mapping Layer)、QACNN 層與預測層。在相似映射層中，將輸入文字透過詞向量(Mikolov, Chen, Corrado & Dean, 2013; Pennington, Socher & Manning, 2014; Bengio, Ducharme, Vincent & Jauvin, 2003; Bojanowski, Grave, Joulin & Mikilov, 2017)表示後，透過餘弦相似度計算每個文章與問題或選項之間的相似度取得相對應的矩陣PQ與PC來表示位置關聯(Location Relationship)的資訊：

$$P_nQ = \left\{Cos\left(P_n^i, Q^j\right)\right\}_{i=1,j=1}^{I,J} \tag{1}$$

其中n為文章共有幾個句子、I與J則是分別代表文章與問題每句的長度。接著，經由 QACNN 層透過兩階段的注意力機制(Attention Mechanism)將位置相關資訊視為一種圖形(Pattern)去產生字級(Word-level)到句級(Sentence-level)的特徵r_{m}來表示第 m 個選項對於文章與問題的資訊：

$$r_m = QACNN_{Layer}(PQ, PC_m) \tag{2}$$

最後透過預測層來蒐集每組選項的資訊r_{m}來預測最有可能是答案的選項。QACNN 的特色是將文章、問題與選項透過文字相似度的方式，將此任務變成一種類似於圖形學習(Pattern Learning)的方式來呈現。因此，QACNN 層中，利用多個核大小(Kernel Size)來取得不同尺度的特徵，最後利用這些特徵來進行答案的預測。QACNN 模型示意圖如圖 1

所示。

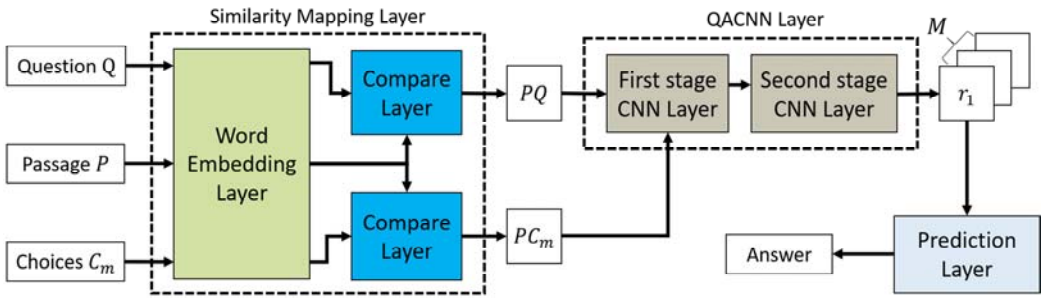

圖 1. QACNN 模型示意圖
[Fiqure 1. Illustration of QACNN Model.]

除了 QACNN 外，Co-Matching 同樣為英文選擇題的經典模型，其模型示意圖如圖 2 所示。此模型利用匹配特徵(Match Feature)來產生文章、問題與選項的字級特徵。匹配特徵通常會有兩種輸入，分別以A與B來表示，並且透過各自所定義的注意力機制算法來產生$\bar{A}$之後來進行後續的運算需求，如式(3)所示。

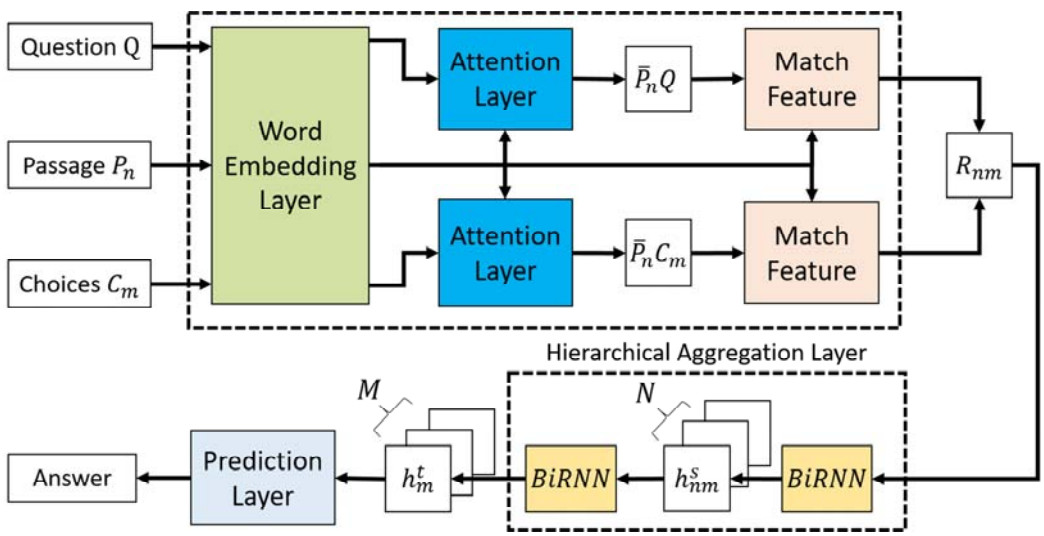

圖 2. Co-Matching 模型示意圖
[Fiqure 2. Illustration of Co-Matching Model.]

透過匹配特徵提取相關(Correlation)資訊$Match^{AB}$重新表達A與B的特徵關係，如式(4)所示。Co-Matching 將文章中的句子P_n、問題Q與選項C_m，來產生R_{nm}特徵以供後續模型需求使用，如式(5)所示，並且其中n與 m分別為文章第幾個句子以及第幾個選項。

$$\bar{A} = Attention(A, B) \tag{3}$$

$$Match^{AB} = ReLU\left(W\begin{bmatrix}\bar{A} \ominus B \\ \bar{A} \otimes B\end{bmatrix} + b\right) \tag{4}$$

$$R_{nm} = \begin{bmatrix} Match^{P_n Q} \\ Match^{P_n C_m} \end{bmatrix} \quad (5)$$

接著透過雙向循環神經網路(Recurrent Neural Network, RNN)來彙整每個字級特徵R_{nm}來產生句級特徵h_{nm}^s：

$$h_{nm}^s = MaxPooling(BiRNN(R_{nm})) \quad (6)$$

將文本每個句子的句級特徵h_{nm}^s串接起來後，重新表達為H_m^s：

$$H_m^s = [h_{1m}^s; h_{2m}^s; \ldots; h_{Nm}^s] \quad (7)$$

再經由階級式彙集(Hierarchical Aggregation)對於每個句級的結構整合成文檔級(Document-level)的特徵h_m^t，並用此特徵來預測最有可能是答案的選項：

$$h_m^t = MaxPooling(BiRNN(H_m^s)) \quad (8)$$

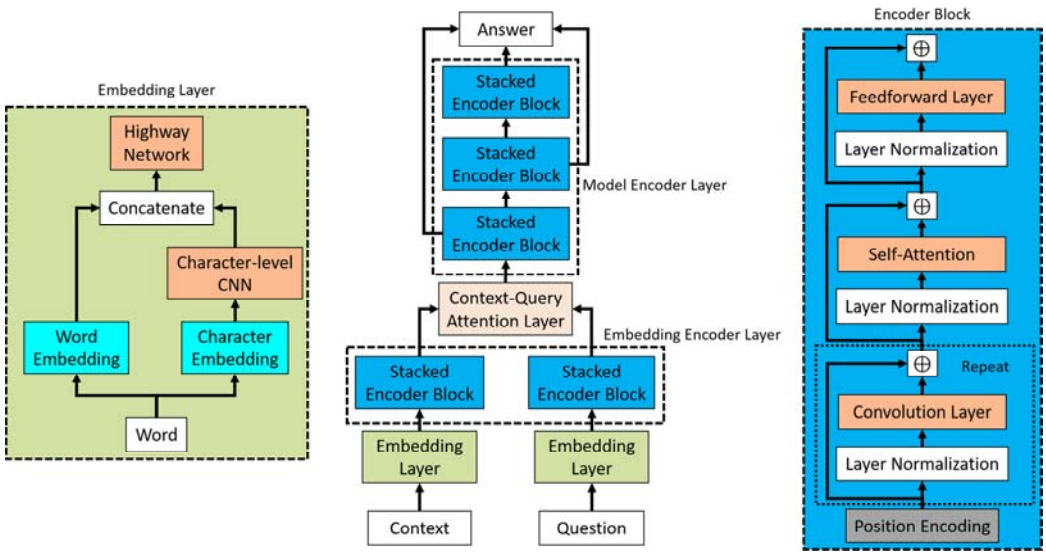

圖 3. QANet 模型示意圖
[Figure 3. Illustration of QANet Model.]

相較於 QACNN 與 Co-Matching，QANET 是預測文本段的問答模型，此模型包含五個主要部分：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)與輸出層(Output Layer)。目前大多數的機器閱讀理解模型主要皆以注意力機制(Attention Mechanism)與循環神經網路為主要架構，但 QANET 整體的網路設計皆捨棄循環神經網路，僅使用卷積神經網路(Convolutional Neural Networks, CNN)和自我注意力機制(Self-attention Mechanism) (Vaswani *et al*., 2017)來建構，由於 QANET 並未使用循環神經網路，因此它可以採用平行化的訓練方式，使得訓練速度與推論(Reasoning)速度更快。更明確地，QANET 採用卷積神經網路獲取文本的局部結構，而自我注意力機制可以學

習全文中單詞與單詞之間的關係，由於沒有時間上的遞迴關係，可以把模型建的更深，也讓訓練速度提升 3-13 倍、推論速度提升 4-9 倍。QANET 所使用的編碼器都是基於一定規格來建構，如圖 3 所示，並對每一層卷積層的數量進行修改，而且每一層之間皆使用層正規化(Layer Normalization) (Ba, Kiros & Hinton, 2016)與殘差網路(Residual Network) (He, Zhang, Ren & Sun, 2016)來穩定訓練過程。除此之外，QANET 共享文章、問題與模型編碼器之間的部分權重，以達到更加泛化之效果。

近期 BERT(Bidirectional Encoder Representation from Transformers) (Devling, Chang, Lee & Toutanova, 2018)的問世以及後續 XLNet (Yang *et al*., 2019)對於 BERT 的改進，讓各項自然語言處理任務創下最新紀錄。BERT 透過多層雙向轉換編碼器來訓練兩個無監督的預測任務，分別為遮掩式語言模型(Masked Language Model)以及下一句預測(Next Sentence Prediction)，不僅讓模型成為高強健性的表徵學習法，並且此一模型僅需藉由簡易的微調(Fine-tune)機制，即可在各式自然語言處理的任務上取得相當亮眼的效果(Zhang *et al*., 2019; Ran *et al*., 2019)。XLNet 則是基於 BERT 的缺點進行改進，提出一種泛化自回歸的訓練方法(Generalized Autoregressive Pretraining Method)，針對 BERT 對於遮掩位置與其他的依賴關係進行突破，藉此達到刷新 BERT 紀錄的模型。

3. 基於特徵粒度之訓練策略 (A Feature-granularity Training Strategy)

3.1 語音辨識與口語問答系統 (ASR and Spoken Question Answering)

自動語音辨識往往由於環境噪音、說話者口音或新興的詞彙不存在於辭典中而無法被辨識等因素，導致語音辨識錯誤的發生，當辨識錯誤發生時，後續又會造成諸如斷詞錯誤、語意不清或是關鍵字錯誤等問題。這些問題又都將影響後續各式基於文本的問答模型，造成任務成效不彰的狀況。更明確地，我們以一個實際的口語問答系統資料為例，如圖 4 所示。在以詞(Word)為單位的狀況下，選項跟文章是發生完全不匹配(Mismatch)的情況。值得注意的是，因為選項沒有前後文資訊，而且通常都是簡短的關鍵字所組成，因此對於語音辨識而言，很難利用語言模型等技術讓輸出的結果變得更好，也因此在這個例子中，辨識出『**廢**』與『**剛才**』這兩個在單字詞語言模型(Unigram)中機率相對較高的單詞。若我們將辨識的結果以字符(Character)呈現，則選項與文章會匹配到『**廢**』這個字符，『**剛才**』與文章中所辨識出的『**鋼 材**』同樣是完全不匹配的狀況，並且明顯地，這些字符在語意上幾乎是不相關的。最後，儘管辨識結果在詞與字符的表現上相當不理想，但若我們進一步地將辨識結果以音節(Syllable)呈現，反而可以出現完全匹配的情況！也就是說理想情況上，以詞為單位的表示時，是可以明確且清楚的表達語意的資訊，在以字或音節為單位的表示時，語意是較不清楚的；但是在語音辨識結果裡，雖然以詞為單位看似有較明確的語意資訊，但往往受到語音辨識錯誤的影響，可能獲得的是錯誤的字詞資訊，而當我們將辨識的結果轉換成字符或音節的表示時，雖然語意資訊較為薄弱，甚至是沒有語意資訊的，卻似乎可以克服一些語音辨識所造成的錯誤問題。有鑑於此，本論文提出一種同時使用不同特徵粒度(即詞、字與音符)的訓練策略，以達到改善自動語音辨識

錯誤所造成的效能減損，讓原本問答模型達到更具強健性的效果。

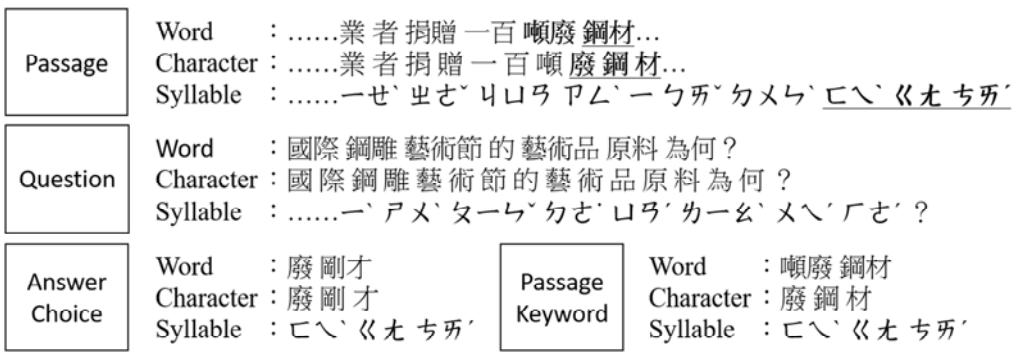

圖 4. 自動語音辨識文本與特徵差異之示意圖
[Fiqure 4. Schematic diagram of automatic speech recognition text and feature differences.]

3.2 基於數據粒度之訓練策略 (A Data Granularity Training Strategy)

為了改善自動語音辨識錯誤在口語問答系統中的影響，本研究提出一套利用訓練資料中不同粒度的特徵之訓練策略。更明確地，我們將利用自動語音辨識系統輸出的文本T，進一步地轉換成三種表示法，分別是以詞為單位$T_W = \{w_1, w_2, \cdots, w_N\}$、以字符為單位$T_C = \{c_1, c_2, \cdots, c_M\}$和以音節為單位$T_S = \{s_1, s_2, \cdots, s_L\}$，其中$N$、$M$與$L$分別代表文本$T$被表示成詞、字符與音節時的長度，並且通常$N \geq M = L$，如圖 5 所示。在選擇題形式的口語問答中，文本$T$則包含文章$T^P$、問題$T^Q$與選項$T^C$，因此我們可以進一步的將文章以詞、字符以及音節分別表示為T_W^P、T_C^P與T_S^P；同樣的，問題與選項也可以分別表示為T_W^Q、T_C^Q、T_S^Q與T_W^C、T_C^C、T_S^C。接著，本論文提出的訓練策略是輪流使用不同粒度的資料來進行同一個問答模型的參數更新與訓練，如圖 6 所示。

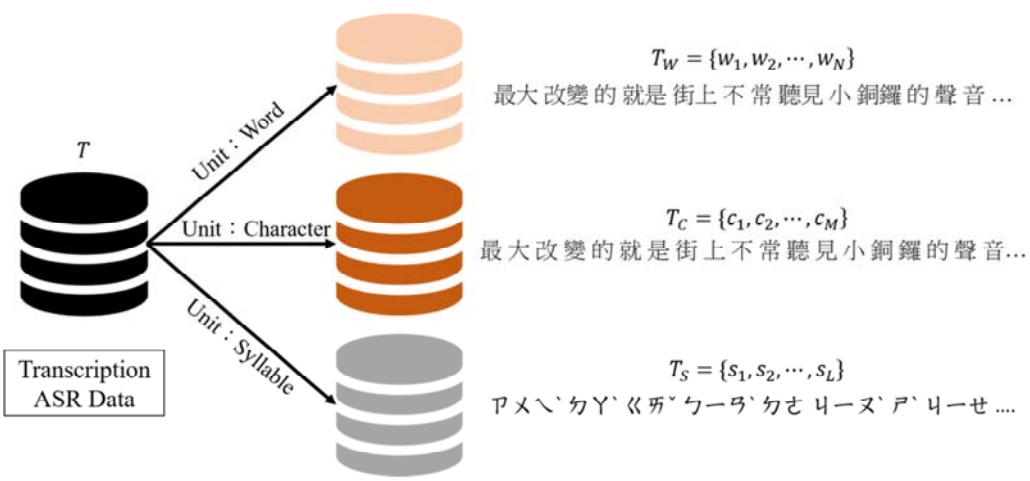

圖 5. 數據粒度示意圖
[Fiqure 5. Illustration of Data Granularity.]

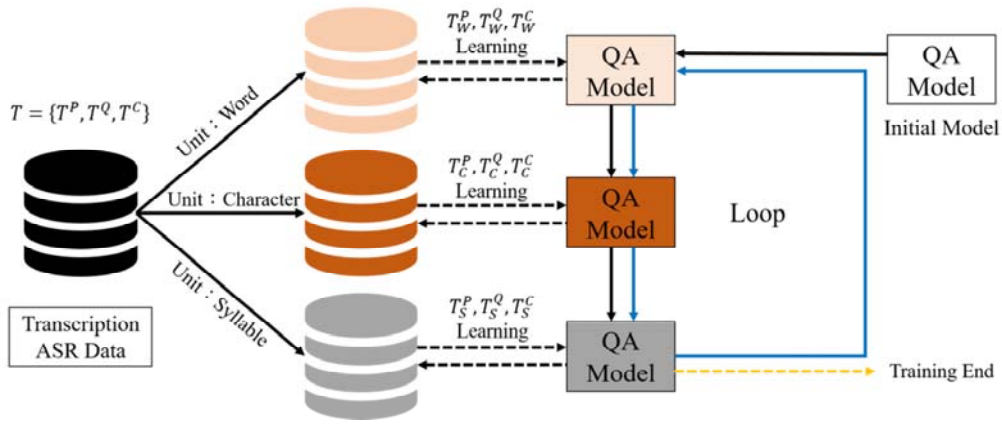

圖 6. 先詞再字符後音節（W-C-S）之訓練策略於選擇題問答模型示意圖
[Fiqure 6. Schematic Diagram of the Training Strategy of the Pre-character and Post-character Syllable (W-C-S) in the Multiple Choice Question Answering Model.]

更明確地，以圖 6 為例，我們欲訓練一個選擇題形式的口語問答系統，因此我們首先將所有訓練資料（包含文章、問題與選項）轉換成以詞為單位$\{T_W^P, T_W^Q, T_W^C\}$、以字符為單位$\{T_C^P, T_C^Q, T_C^C\}$和以音節為單位$\{T_S^P, T_S^Q, T_S^C\}$的表示法。接著，這個問答模型將首先以詞為單位的資料（即$\{T_W^P, T_W^Q, T_W^C\}$）進行訓練，再以字符為單位的數據（即$\{T_C^P, T_C^Q, T_C^C\}$）來給予模型進行訓練；最後依照同樣步驟，將以音節為單位的表示法（即$\{T_S^P, T_S^Q, T_S^C\}$）給予問答模型進行學習。換句話說，我們提出依照先詞再字符後音節（即 W-C-S）的順序，循環地訓練同一組問答模型的參數，直到模型收斂為止。必須注意的是，在訓練不同粒度的資料時，是對整個數據集進行模型更新後，再換下一個粒度的資料進行學習，而不是一筆訓練資料以三種不同的表示法分別學一次後，才換下一筆資料。因此，在使用此一訓練策略時，詞、字符與音節的向量表示法維度應設為相同，以利於模型共享參數進行學習。

總結來說，本論文提出的模型訓練策略是期望藉由學習不同粒度的表示法，例如有清楚的語意資訊的詞層次表示法或是有較多匹配的單元但語意資訊薄弱（甚至沒有）的字符或音節粒度，來模擬受到語音辨識錯誤影響的口語問答系統，以達到改善自動語音辨識錯誤所造成的效能損失，讓口語問答模型達到更具強健性的效果。

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

4.1 實驗設定 (Experimental Settings)

4.1.1 詞向量訓練 (Word Embeddings Training)

本論文進行口語問答模型實驗時，將首先透過詞向量技術來表達辨識出來的每個詞、字符與音節，並且皆採用批踢踢實業坊(PTT)以及中央新聞社(CNA[1])文本作為訓練語料，維度大小皆為 300 維，並將詞頻小於 5 的詞彙捨去。在訓練詞與字符向量時，是利用快文(Fasttext)向量模型來進行訓練，由於快文向量本身演算法的優勢，可以在訓練詞向量時，也可取得字符向量表達。音節向量訓練則是採用全局向量模型(GloVe)來進行訓練。

4.1.2 自動語音辨識系統 (Automatic Speech Recognition System)

在自動語音辨識系統上，我們基於 Kaldi (Povey *et al*., 2011)工具包，聲學模型是基於神經網路的 TDNN-F (Povey *et al*., 2018)系統，透過 Lattice-free MMI (Povey *et al*., 2016)和 sMBR (Vesely, Ghoshal, Burget & Povey, 2013)技術來進行模型訓練。我們的聲學模型訓練語料是採用 NER-Trs-Vol1~3 與 MATBN (Wang, Chen, Kuo & Cheng, 2005)。語言模型為三連語言模型(Trigram Language Model)並以 Kneser-Ney 平滑化技術來解決稀疏數據的問題。最後此自動語音辨識系統的字符錯誤率(Character Error Rate)為 7.79%。

4.1.3 口語選擇題的問答模型 (Spoken Multiple Choices Question Answering Model)

我們使用 2018 年科技大擂台-與 AI 對話(Formosa Grand Challenge – Talk to AI[2])的比賽資料來進行實驗。該比賽是專屬於中文口語選擇題的問答任務，其中每個題目都會包含文章、問題與選項之音檔，回答則是哪個選項是正確答案。關於此比賽資料所涉及的領域非常的廣泛，包含科學、新聞或文言文等等，我們選用其中 8 場比賽資料來進行實驗，訓練集、發展集與兩種測試集之統計資訊如表 1 所示。發展集是包含雜訊干擾後的音訊，測試集則分為兩種，第一種測試集是有文言文內容的音訊，第二種則是決賽的題目以中高階華語文測驗為主。關於每個音檔的資訊，其文章、問題與選項平均分別為 22,538、2,230 與 5,904 個幀(Frame)；在正確文本(Manual Transcriptions)上，每組題目平均分別為 235.3、16.7 與 4.1 個字元。此實驗採用正確率做為評估指標。在 QACNN 中，隱藏層大小設定分別為 32、64 與 128，卷積層的核大小也分成[1,2,3]、[1,3,5]、[1,3,7]與[1,4,7]來進行實驗。在 Co-matching 上，隱藏層大小則設定為 64 與 128。以上兩種模型丟失率(Dropout Ratio)皆為 0.2，並都採用 Adam 優化器(Kingma & Ba, 2014)且學習率為 0.001、批次(Batch)大小分別為 32 與 4，總期次數分別為 13 與 80。對於文章、問題與每個選項

[1] CNA: http://www.cna.com.tw/

[2] Formosa Grand Challenge - Talk to AI: https://fgc.stpi.narl.org.tw/activity/techai2018

的長度限制分別為 600、40 與 40 個字元。實驗皆以 python 3.7.2 與 PyTorch 0.4.1(Paszke *et al.*, 2017)套件實現。

表 1. 2018 科技大擂台數據集
[Table 1. 2018 Formosa Grand Challenge – Talk to AI Dataset.]

訓練集 Training Set	發展集 Development Set	測試集 Test Set	
7,050	1,000	1,500	1,000

表 2. 台達電閱讀理解資料集
[Table 2. Delta Reading Comprehension Dataset.]

訓練集 Training Set		測試集 Test Set	
文本數	問題數	文本數	問題數
5,014	26,936	1,000	3,524

4.1.4 文本段的問答模型 (Answer-span Prediction Question Answering Model)

除了選擇題式的問答系統外，本論文亦進行以文本段形式的問答模型實驗，以進一步地驗證本論文提出之方法在不同形式的問答系統中，皆有其效用。我們使用台達電閱讀理解數據集(Delta Reading Comprehension Dataset, DRCD) (Shao, Liu, Lai, Tseng & Tsai, 2018)進行中文文本段問答模型之實驗，其訓練集與測試集之統計資訊如表 2 所示。其中每個題目都是由一段文章與問題所組成，其答案為此文章的一小段落作為答案。我們採用 F1 與 EM 分數作為評估指標。在 QANet 中，隱藏層大小設定為 96，並且批次大小與總訓練次數分別為 10 與 20，其餘皆參照原論文之訓練方式來進行學習。對於文章、問題的長度限制分別為 800 與 50 個字。

4.1.5 訓練策略於問答系統 (Training Strategy for Question Answering)

本論文對於訓練策略進行實驗時，基礎系統(Baseline System)是以下訓練代號：W(詞)、C(字符)與 S(音節)來呈現，分別代表輸入詞向量、字符向量或音節向量來訓練模型。除了基礎系統的設定外，驗證本論文提出的方法共有四種訓練策略方式，分別為 W-C、W-S、C-S 與 W-C-S 訓練代號來呈現，分別代表進行訓練時，會依照其所指定的訓練順序來給予模型進行學習；像是 C-S 則是代表在訓練策略中，會讓模型先訓練於字符向量後再給予音節向量來進行學習。順帶一提的是，本論文提出的訓練策略，是應用不同的數據粒度的資料呈現來訓練模型，以達到更好的模型效能，並不是專注於如何讓不同的數據特徵同時輸入給網路，因此在實際預測時，可以透過不同數據粒度的詞向量來輸入給模型，並於實驗結果中呈現並探討。

4.2 實驗結果 (Experimental Results)

4.2.1 口語選擇題的實驗結果 (Experimental Results of Spoken MCQA)

此實驗結果中，我們先呈現 QACNN 以及 Co-Matching 在於中文口語選擇題的問答任務之結果，此組實驗也是本實驗的基礎系統，實驗結果如表 3 與表 4 所示。由於我們訓練每個詞、字符以及音節向量時，其維度皆相同，故此問答模型可以丟入不同的詞向量來進行預測。我們這邊是以成對的方式來進行數據呈現；例如 C 就是訓練與預測模型時皆只使用字符向量來進行評估。有鑑於在每個不同的參數設定並沒有很明顯的模型效能差異，故採用每個系統平均精確度的方式來呈現實驗結果。可以發現在於兩種模型在驗證與測試集 1 上，都呈現 S > C > W 的狀況。我們認為這是當有自動語音辨識錯誤產生的時候，導致單用詞向量會有語意不清的情況；當透過字符向量來訓練模型，雖也有語意不清的狀況，但是相較於詞向量會有機會找出重要的字符，而不是僅依靠詞來進行判斷。單訓練在音節向量時，其實驗結果皆較詞向量與字符向量佳，可能的原因為對於 QACNN 來說，不需要太過專注於詞向量的表徵上，而是看每個詞與詞之間的相似程度來進行學習，因此利用音節這種一對多的關係來給表示相似程度或許是有助益的；反過來說，Co-Matching 是對輸入的文字進行編碼器後，再取得匹配特徵進行所需之流程，在實驗上，在驗證集與測試集 2 中並無明顯效益，而測試集 1 是含有噪音的干擾問題，所以當採用音節特徵給予模型學習時，會有比較好的結果。在正確轉寫文本的實驗結果上，相較於自動語音辨識的結果有相當顯著的提升，並且在不同粒度的詞向量的差距反而變小。在中高階華語文難度的測試集，因為需要明確語意來回答問題，故在相較於使用字符與音節向量上，詞向量有比較好的預測效果。

表 3. 各模型於中文口語選擇題的問答任務基礎系統之自動語音辨識實驗結果
[Table 3. Experimental Results of the Baseline System in Chinese Spoken Question Answering Task with ASR Transcription.]

	Validation			Test 1			Test 2		
	W	C	S	W	C	S	W	C	S
QACNN	56.67	63.59	64.78	68.68	70.73	74.02	39.07	39.45	39.53
Co-Matching	54.00	64.02	65.05	58.56	72.32	74.86	38.14	39.62	39.20

表 4. 各模型於中文口語選擇題的問答任務基礎系統之正確文本實驗結果
[Table 4. Experimental Results of the Baseline System in Chinese Spoken Question Answering Task with Manual Transcription.]

	Validation			Test 1			Test 2		
	W	C	S	W	C	S	W	C	S
QACNN	64.76	70.42	70.54	80.14	82.58	82.26	41.67	42.25	42.53
Co-Matching	59.88	73.26	72.90	62.56	71.83	73.22	39.17	42.01	40.64

接下來，我們呈現本論文提供之訓練策略於各式模型的效果，如表 5 至 7 所示。有別於表 3 與表 4 的表示法，W 至 W-C-S 是指各種不同的訓練策略時，所使用的詞向量順序；在於表格中 QACNN[W/C/S]，則是 QACNN 在這訓練策略下，我們使用[W/C/S]的[詞/字符/音節]向量作為輸入時，其模型預測的準確度。其中粗體代表的是對應於訓練決策時，有所使用的向量表示；粗體加底線的數據則是由於訓練策略有利用多個向量表示來進行訓練，並透過這種標記方式來呈現，這些都屬於相對有所使用的向量。我們發現藉由這種訓練策略，相較於單用單一向量訓練 W、C 與 S 的狀況，當利用多個向量進行訓練時，是有助於模型進行預測時的效果。在驗證集、測試集 1 與 2 之中，分別達到 1.28%~2.07%、0.52%~2.35%與 0.38%~1.84%之效能改進。上述的實驗結果中，可以發現 QACNN 與 Co-Matching 兩個模型，在於不同的向量輸入上，並沒有很嚴重的不匹配情況。就像是只訓練於詞向量的問答模型，對於以字符與音節向量作為預測資料的輸入，也能有一定之效能。從另外一個角度來觀察，可以發現在大部分的數據中，不管是採用何種訓練策略下，採用音節向量進行輸入時，通常會取得最好的結果，我們認為這是因為在自動語音辨識錯誤的文本上，音節對於 QACNN 與 Co-Matching 來說，是比較有利於預測的。反之在正確文本上，則是採用字符向量或音節向量時，會得到大部分最好的結果。

表 5. 各模型於中文口語選擇題的問答任務驗證集之自動語音辨識實驗結果
[Table 5. Experimental Results of the Baseline System in Chinese Spoken Question Answering Validation Task with ASR Transcription.]

	Validation						
	W	C	S	W-C	W-S	C-S	W-C-S
QACNN [W]	**56.67**	55.97	56.38	**56.69**	**56.97**	56.47	**56.72**
QACNN [C]	62.56	**63.59**	63.49	**63.45**	63.78	**63.61**	**63.94**
QACNN [S]	64.19	64.98	**64.78**	65.99	**65.91**	**66.07**	**66.26**
Co-Matching [W]	**54.00**	44.95	39.66	**55.31**	**55.16**	44.00	**55.48**
Co-Matching [C]	59.84	**64.02**	57.66	**64.19**	61.80	**64.72**	**65.11**
Co-Matching [S]	59.51	61.43	**65.05**	61.37	**66.23**	**66.60**	**66.44**

表 6. 各模型於中文口語選擇題的問答任務測試集 1 之自動語音辨識實驗結果
[Table 6. Experimental Results of the Baseline System in Chinese Spoken Question Answering Test 1 Task with ASR Transcription.]

	Test 1						
	W	C	S	W-C	W-S	C-S	W-C-S
QACNN [W]	**68.68**	68.28	67.57	**69.33**	**68.57**	68.16	**69.07**
QACNN [C]	68.58	**70.73**	71.16	**71.22**	71.30	**71.39**	**71.30**
QACNN [S]	72.19	73.36	**74.02**	74.05	**74.33**	**74.09**	**74.54**
Co-Matching [W]	**58.56**	49.97	42.97	**62.02**	**60.48**	50.15	**62.38**
Co-Matching [C]	63.35	**72.32**	61.24	**73.38**	65.31	**73.36**	**73.21**
Co-Matching [S]	66.43	68.65	**74.86**	66.92	**74.88**	**75.26**	**74.98**

表 7. 各模型於中文口語選擇題的問答任務測試集 2 之自動語音辨識實驗結果
[Table 7. Experimental Results of the Baseline System in Chinese Spoken Question Answering Test 2 Task with ASR Transcription.]

	Test 2						
	W	C	S	W-C	W-S	C-S	W-C-S
QACNN [W]	**39.07**	40.19	39.35	**40.24**	**40.31**	39.35	**39.65**
QACNN [C]	38.97	**39.45**	39.10	**39.59**	40.16	**39.43**	**40.01**
QACNN [S]	38.09	39.18	**39.53**	40.08	**39.08**	**39.48**	**39.58**
Co-Matching [W]	**38.14**	36.46	32.52	**38.79**	**38.18**	36.59	**39.09**
Co-Matching [C]	38.43	**39.62**	35.50	**40.64**	38.07	**39.89**	**40.97**
Co-Matching [S]	36.37	36.68	**39.20**	36.94	**38.71**	**39.73**	**39.76**

4.2.2 文本段的實驗結果 (Experimental Results of Answer-span Question Answering)

我們呈現 QANet 在 DRCD 資料集之成效，並包含各式訓練策略與不同詞向量輸入之 F1 與 EM 評估結果，實驗結果如表 8 與表 9 所示。同樣地，W、C、S 中粗體部分，即為文本段問答模型的基礎系統，並其餘表格設定皆與表五至表七相關描述相同。可以發現在沒學習過的向量表示上，有別於選擇題形式的結果，會有嚴重的**不匹配問題**產生，但在 W-C 與 W-S 上，卻發現了不同的情況：在 W-C 的訓練策略上，是符合未匹配的情況；在 W-S 的訓練策略上，其效能比其他訓練策略都低之外，卻可以在輸入字符向量時，有一定之效能。因此，QANet 在本論文所提出的訓練策略中，最適合之組合為透過 W-C 的訓練策略來訓練，並採用字符向量作為輸入，可得到最好之效果。值得注意的是，DRCD 的資料都以正確文本來進行實驗，理應不包含自動語音辨識錯誤的問題，所以本論文實

驗於這組測試上，是為了證明本訓練策略在於非自動語音辨識錯誤的文本下，也可以讓模型得到一定程度之改善。

表 8. QANet 於 DRCD 問答任務測試集之 F1 實驗結果
[Table 8. F1 Experimental Results of the QANet Model in DRCD Test Task.]

	DRCD						
	W	C	S	W-C	W-S	C-S	W-C-S
QANet [W]	**78.04**	24.31	12.76	**79.02**	**75.27**	30.11	**76.37**
QANet [C]	57.99	**81.61**	10.63	**83.40**	65.33	**80.77**	**82.23**
QANet [S]	11.56	22.81	**76.52**	25.58	**69.20**	**74.78**	**74.00**

表 9. QANet 於 DRCD 問答任務測試集之 EM 實驗結果
[Table 9. EM Experimental Results of the QANet Model in DRCD Test Task.]

	DRCD						
	W	C	S	W-C	W-S	C-S	W-C-S
QANet [W]	**64.21**	10.12	2.31	**64.89**	**60.60**	13.16	**60.13**
QANet [C]	35.79	**69.83**	1.89	**74.40**	43.36	**68.97**	**70.30**
QANet [S]	1.06	3.82	**59.18**	5.80	**48.80**	**57.79**	**55.43**

5. 結論 (Conclusions)

本論文提出一種簡單的訓練策略，透過數據粒度的概念，來有效改善自動語音辨識錯誤所導致的效能問題，並透過實驗證明，不需額外模型的支援下，僅利用輸入不同向量的方式讓模型達至更好的結果。我們所提出之方法，相較於基礎系統，在口語選擇題的實驗結果，分別在 QACNN 與 Co-Matching 上，得到 2%至 4%的進步；在 QANet 於中文數據集文本段的實驗結果，F1 與 EM 分別得到 1.79%與 4.57%的進步。未來會透過將詞向量再細分為二元、三元等方式來進行延伸擴充或是將此訓練策略應用於英文數據集上來驗證成效，並且將會與 BERT 與 XLNET 等基於語言模型的神經網路系統進行相結合。

參考文獻 (Reference)

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer Normalization. In arXiv preprint arXiv:1607.06450

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A Neural Probabilistic Language Model. *JMLR*, *3*, 1137-1155.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. In *Proceedings of TACL*, *5*, 135-146. doi: 10.1162/tacl_a_00051

Choi, E., He, H., Iyyer, M., Yatskar, M., Yih, W.-t., Choi, Y., …Zettlemoyer, L. (2018). QuAC : Question Answering in Context. In arXiv preprint arXiv:1808.07036

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In arXiv preprint arXiv:1810.04805

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In *Proceedings of CVPR 2016*,770-778. doi: 10.1109/CVPR.2016.90

van Heerden, C., Karakos, D., Narasimhan, K., Davel, M., & Schwartz, R. (2017). Constructing Sub-word Units for Spoken Term Detection. In *Proceedings of ICASSP 2017*, 5780-5784. doi: 10.1109/ICASSP.2017.7953264

Karras, T., Laine, S., & Aila, T. (2019). A Style-based Generator Architecture for Generative Adversarial Networks. In *Proceedings of CVPR 2019*, 4401-4410. doi: 10.1109/CVPR.2019.00453

Kingma, D. P. & Ba, J. (2014). Adam: A Method for Stochastic Optimization. In arXiv preprint arXiv:1412.6980

Kingma, D. P. & Dhariwal, P. (2018). Glow: Generative Flow with Invertible 1x1 Convolutions. In *Proceedings of NIPS 2018*, 10215-10224.

Lai, G., Xie, Q., Liu, H., Yang, Y., & Hovy, E. (2017). RACE: Large-scale ReAding Comprehension Dataset From Examinations. In arXiv preprint arXiv:1704.04683

Lee, C.-H., Chen, Y.-N., & Lee, H.-y. (2019). Mitigating the Impact of Speech Recognition Errors on Spoken Question Answering by Adversarial Domain Adaptation. In *Proceedings of ICASSP 2019*, 7300-7304. doi: 10.1109/ICASSP.2019.8683377

Liu, T.-C., Wu, Y.-H., & Lee, H.-y. (2017). Query-based Attention CNN for Text Similarity Map. In arXiv preprint arXiv:1709.05036

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In arXiv preprin arXiv:1301.3781

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ...Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems 32*, 8024-8035

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. In *Proceedings of EMNLP 2014*, 1532-1543. doi: 10.3115/v1/D14-1162

Povey, D., Cheng, G., Wang, Y., Li, K., Xu, H., Yarmohamadi, M., …Khudanpur, S. (2018). Semi-orthogonal Low-rank Matrix Factorization for Deep Neural Networks. In *Proceedings of INTERSPEECH 2018*, 3743-3747. doi: 10.21437/Interspeech.2018-1417

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., …Veselý, K. (2011). The Kaldi Speech Recognition Toolkit. In *Proceedings of ASRU 2011*.

Povey, D., Peddinti, V., Galvez, D., Ghahremani, P., Manohar, V., Na, X., Wang, Y., …Khudanpur, S. (2016). Purely Sequence-trained Neural Networks for ASR Based on Lattice-free MMI. In *Proceedings of INTERSPEECH 2016*, 2751-2755.

Ran, Q., Li, P., Hu, W., & Zhou, J. (2019). Option Comparison Network for Multiple-choice Reading Comprehension. In arXiv preprint arXiv:1903.03033

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of EMNLP 2016*, 2383-2302. doi: 10.18653/v1/D16-1264

Reddy, S., Chen, D., & Manning, C. D. (2019). Coqa: A Conversational Question Answering Challenge. In *Proceedings of TACL*, *7*, 249-266. doi: 10.1162/tacl_a_00266

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

Shiang, S.-R., Lee, H.-y., & Lee, L.-s. (2014). Spoken Question Answering Using Tree-structured Conditional Random Fields and Two-layer Random Walk. In *Proceedings of ISCA 2014*, 263-267.

SZŐKE, I. (2010). *Hybrid word-subword spoken term detection*. (Doctoral thesis, Brno University of Technology, Brno, Czech Republic). Retrieved from: https://www.fit.vut.cz/study/phd-thesis/150/

Tang, M., Cai, J., & Zhuo, H. H. (2019). Multi-Matching Network for Multiple Choice Reading Comprehension. In *Proceedings of AAAI 2019*. doi:10.1609/aaai.v33i01.33017088

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is All You Need. In *Proceedings of NIPS 2017*, 5998-6008.

Veselý, K., Ghoshal, A., Burget, L., & Povey, D. (2013). Sequence-discriminative Training of Deep Neural Networks. In *Proceedings of INTERSPEECH 2013*, 2345-2349.

Wang, H.-M., Chen, B., Kuo, J.-W., & Cheng, S.-S. (2005). MATBN: A Mandarin Chinese Broadcast News Corpus. *IJCLCLP*, *10*(2), 219-236.

Wang, W., Shen, J., Guo, F., Cheng, M.-M., & Borji, A. (2018). Revisiting Video Saliency: A Large-Scale Benchmark and a New Model. In *Proceedings of CVPR 2018*, 4894-4903. doi: 10.1109/CVPR.2018.00514

Wang, W., Shen, J., Guo, F., Cheng, M.-M., & Borji, A. (2018). Revisiting Video Saliency: A Large-Scale Benchmark and a New Model. In *Proceedings of CVPR 2018*, 4894-4903. doi: 10.1109/CVPR.2018.00514

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. In arXiv preprint arXiv:1906.08237.

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. In arXiv preprint arXiv:1804.09541

Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., & Xiao, J. (2015). LSUN: Construction of a Large-scale Image Dataset Using Deep Learning with Humans in the Loop. In arXiv preprint arXiv:1506.03365.

Zhang, S., Zhao, H., Wu, Y., Zhang, Z., Zhou, X., & Zhou, X. (2019). Dual Co-Matching Network for Multi-choice Reading Comprehension. In arXiv preprint arXiv:1901.09381