The 2017 Conference on Computational Linguistics and Speech Processing
ROCLING 2017, pp. 354-369

序列標記與配對方法用於語音辨識錯誤偵測及修正

On the Use of Sequence Labeling and Matching Methods for ASR Error Detection and Correction

吳佳樺 Chia-Hua Wu, 蔡淳伊 Chun-I Tsai, 洪孝宗 Hsiao-Tsung Hung, 高予真 Yu-Chen Kao, 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
{chiahua, joy.tsai, alexhung, cybelia, berlin}@ntnu.edu.tw

摘要

本論文著重在研究語音辨識錯誤相關的幾個重要面向，尤其是當一般的語音辨識系統應用於特殊領域下所產生的未知詞問題。為此目的，我們提出一個兩階段的方法，包括了語音錯誤偵測和錯誤內容修補。在錯誤偵測階段，我們嘗試比較多種序列標記方法去偵測不同型態的錯誤。更進一步，在錯誤修正階段，藉由上一階段所偵測的結果作為依據，利用音素比對方法以特殊領域的關鍵詞表來修正錯誤。在四種應用領域，包括教育議題、工業技術相關訪談、語音記事及會議錄音，所進行的一系列實驗。由實驗結果顯示，我們提出的方法可以使得一般語音辨識系統在上述應用領域中有某種程度上的提升。

Abstract

This paper sets out to study several important aspects pertaining to speech recognition errors, especially the out-of-vocabulary (OOV) word problem that is caused by using generic speech recognition systems for a specific application domain. To this end, a two-stage processing method, involving error detection and error correction, is proposed. For error detection, we explore and compare disparate sequence labeling methods to detect possible errors of different types. Further, in the error correction stage, an effective phone-level matching mechanism along with a domain-specific keyword list is exploited to correct errors of different types detected by the previous stage. Extensive experiments conducted on four application domains, including educational issues, industrial technology-related interviews and speech memos and meeting recordings, show that our proposed methods can boot the performance of a given

general speech recognition system on the aforementioned application domains to some extent.

關鍵詞：語音辨識，辨識錯誤，錯誤偵測，錯誤修正，未知詞

Keywords: Speech Recognition, Recognition Errors, Error Detection, Error Correction, Out of Vocabulary Words.

一、 緒論

由於機器學習及深度學習的迅速發展[1]，許多領域的性能表現都有大幅度的提升及突破，而語音辨識也不例外。許多大型企業相繼投入語音方面的研究及應用上，並且提供使用者語音相關服務，包含雲端計算與終端裝置的語音辨識的應用程式介面(API)。因為上述平台提供的便利性，使得大量語音互動的智慧型裝置被廣泛地應用，例如車載電腦的語音對話介面和語音客服等，這類的應用通常是依附在語音辨識器之後。因為語料的收集便利性及成本的差異，使得一般的使用者日常對話或熱門話題都能達到良好的辨識正確率，但在特定領域，例如工業應用，則會遭遇到特殊詞彙或該領域的專業術語，可能造成重要字詞無法辨認的問題。最直接的解決方法是重新收集語料，包含語音以及文字內容，再應用模型調適等方法解決問題。然而，新式的語音辨識技術依賴深層類神經網路與大數據資料，需要花費更長的收集資料的時間成本，才能達到和一般情境相近的辨識率，而真實的應用情境可能無法先收集語音再使用服務。因此，如何快速地將辨識器應用至各項領域是個重要的問題。

近年來，大數據及電腦運算能力的大幅提升，以至於語音辨識技術已經進展到更具挑戰的應用，甚至被實踐於現實環境中[2]。而語音辨識系統中的聲學模型已由深層類神經網路(Deep Neural Network, DNN)技術取代傳統高斯混合模型(Gaussian Mixture Model, GMM)，並且在語音辨識任務上獲得更好的效能[1]。而在過去三十多年來，已有數以百計的強健性(noise-robust)語音辨識方法被提出，並且證明其中有許多方法在研究及商業用途上具有重大影響及效用[2]。而本論文主要討論在現實環境中，大規模運用的語音辨識技術，將其應用於特定領域的情境下，導致辨識率大幅降低，針對這樣不匹配(mismatch)的問題去探討其修復錯誤的可能性。本論文根據人工收集關鍵詞清單，用來

改善關鍵詞辨識錯誤所導致的問題。假設這些字詞若能被正確轉寫，則能幫助語音辨識應用於更多領域及情境之下。通用的語音辨識器是由大數據訓練而得的複雜辨識器，且需要 GPU 等運算資源，而每個終端應用都從此辨識器得到第一階段的轉寫文字，再搭配輕量的演算法，進行第二次的轉寫內容修正。我們嘗試兩階段的解決方法：首先設計一個自動分析辨識器錯誤的分類器，再根據錯誤類型資訊，搭配少量的關鍵詞清單來修正內容。

在特殊領域知識的語音辨識任務中，由於語音內容包含大量的特殊名詞，使得辭典外的未知詞(out of vocabulary words, OOV words)會嚴重影響辨識正確率。經由語音辨識流程後的結果仍有轉寫錯誤，在此我們根據是否破壞對話任務的理解，將錯誤分為兩類。第一種輕微的錯誤，通常是一般字詞，但發音不清楚導致發生語音辨識錯誤。這樣錯誤常發生在於自然對話上，例如不流利的重複贅詞或語助詞等。第二種會影響理解的錯誤，大多是特殊字詞且不存在訓練語料的辭典中，導致語音辨識器無法去正確轉寫。例如專有名詞、人名、地名、數字及中英夾雜的字詞等。在實際應用中，不流利語句造成的錯誤適合在第一階段的通用語音辨識器中解決，在此並不討論。而影響理解的特殊詞彙不容易收集大數據並加入訓練語料中，需要獨立解決此問題。重要詞彙清單較容易人工定義，而影響理解的詞可以視為關鍵詞，而我們可以從關鍵詞的精確率和召回率評估轉寫文字否能滿足對話內容的理解。

在本論文當中，我們嘗試解決在特定領域上語音辨識率的不足，將針對這個任務的缺失提出兩步驟的改善模型；第一步驟先偵測語音錯誤之區塊，第二步驟則以關鍵詞回復語音錯誤，並且提升語音辨識率及可讀性。我們提出的方法能比基本的音素對照法更可靠，並且更有效去改善文本錯誤。本論文在第二節將介紹語音錯誤偵測及未知詞改錯相關研究的發展近況；第三節介紹監督式學習的語音錯誤偵測和改錯方法；第四節則是本次改錯任務上實驗結果及討論；最後，第五節提出結論及探討未來可以嘗試的方向。

二、 相關研究

未知詞是一個出現在測試語料，但並且不存在於辨識辭典中的字詞。然而，大多數語音辨識系統都是屬於封閉詞彙(closed-vocabulary)的辨識器，即只能辨識固定且有限的詞彙。當這些未知詞出現在測試語料中，系統將無法識別，導致它被誤認成已知詞。此外，發生未知詞的同時，更可能連帶影響周遭其他的已知詞[3]。而平均來說，一個未知詞可能產生 1.2 個字錯誤[4]。為了改善未知詞的問題，許多研究提出了以模型調適(model adaptation)或是開放詞彙(open-vocabulary)方法來做改善。一般而言，需要收集自然語句才能建立語言模型供辨識器使用，但使用專有名詞的語句不容易收集。以下我們針對語音識別錯誤的改善所使用的特徵及模型方法做更進一步的探討。

近二十年來，已有許多研究嘗試檢測和修復語音辨識錯誤。有幾個方法能夠偵測未知詞：1)以混合語言模型(hybrid language model)做解碼(decoding)，並且以音素、子詞等來表示未知詞；2)以信心分數(confidence score)和其他資訊來尋找可能的未知詞區域；3)結合混合語言模型及信心分數，進一步提升檢索性能[5]。

錯誤修復流程包含錯誤偵測及錯誤修正兩階段。錯誤偵測方法可分為基於設定門檻值(threshold-based)與分類器(classification-based)為基礎的兩種策略。兩者之間有些許差異，基於門檻值的方法是設定單一評估指標或分數來判定是否發生錯誤；而基於分類器的方法大多是整合多種特徵去訓練二元分類器。基於制定門檻值的作法可依據聲學模型的發音分數[6]或語言模型的機率當作信心分數。聲學模型所擷取的對數事後機率或對數相似值作為發音分數[7]。另一方面，利用語言模型計算詞序列機率也是常用的方法，可以作為辨識字詞的信心分數。在基於分類器的方法，主要是以統計模型、機器學習或類神經網路等的進行二元分類。例如以音長模型(duration model)、語音辨識模組中的聲學模型機率及辨識結果等作為輸入特徵，再搭配合適的標記方式，例如條件隨機域(conditional random field, CRF)[4], [8]、類神經網路(neural network, NN)[4]等都是常被採用的選項。

在錯誤修正方面，演算法可以分成簡易的字串搜尋比對，和基於語句擷取特徵再更正文字的兩類。基於語句特徵的方法是藉由上下文資訊來判斷修復內容，方法包含機率

模型、統計模型、機器學習、機器翻譯[7]以及音素對照法[9]。通常是以字詞(word)、音素(phone)、符號(symbol)等作為輸入特徵。例如：[10]提出了一個在對話系統中的語音到語音(speech-to-speech)轉換機制，是利用條件隨機域偵測錯誤標記達到修正文字的目的。[11]提出了一種基於藉由潛在語義分析(latent sematic analysis, LSA)提取上下文的向量表示方法，並利用支持向量機(support vector machine, SVM)分類器作人名辨認。由於上下文語意及主題模型需要大量資料訓練字詞表示法，並且不適用於文本結構較弱的會議語音轉寫中，所以在本文，我們將採用字串搜尋比對[9]作為基礎方法。

在特定領域的語音辨識中，罕見詞或未知詞的處理都是核心的問題[2]。而本論文探討的情境是在一個具有語音強健性的辨識器的情況下，嘗試利用該領域少量的語料資源解決罕見詞與未知詞造成的問題。

三、 語音辨識錯誤偵測和改錯

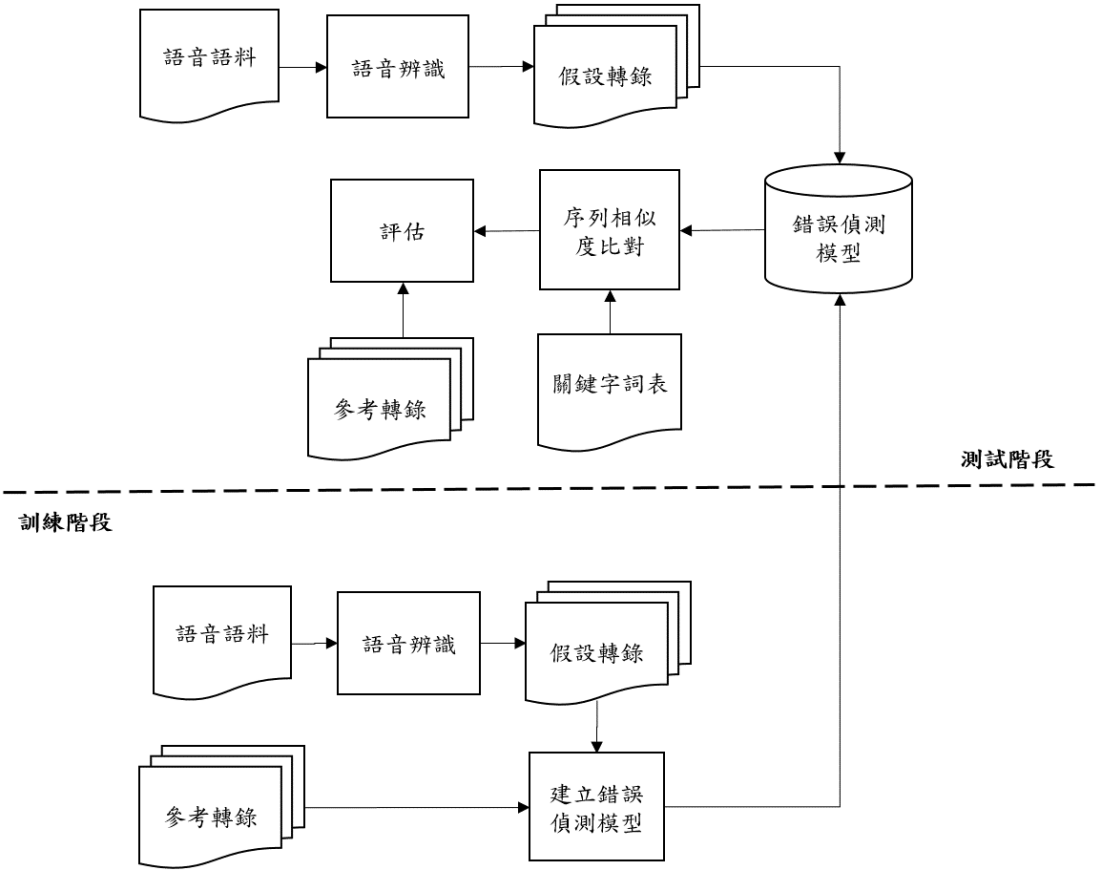
在本節，我們探討辨識錯誤修正的問題，並且提出了一個兩步驟錯誤修正架構(圖一)。第一步驟，尋找可能測試語料中，可能發生錯誤的位置。第二步驟，以音素比對法尋找可能發生錯誤的區塊。以下我們將實驗架構中的兩大主軸，錯誤偵測模型與辨識錯誤修正，做更進一步的模型及方法介紹。

(一)、 錯誤偵測模型

我們探討使用機器學習及類神經網路模型來捕捉辨識字的特性，在此架構下，網路輸入為辨識轉寫文件 D，其中 n 個詞構成的語句以$\{w_1, w_2, w_3 \dots w_n\}$ 表示。網路輸出為錯誤類別，我們使用 $p(k|w_i, \Theta)$ 來定義字詞 w_i 屬於錯誤類別k的事後機率，其中 Θ 表示模型中的參數。

1 詞表示法

詞嵌入(word embedding)是一個字詞的分布表示(distributed representation)。分布表示適用在類神經網路模型的輸入值，並且能與其一起調整參數，計算出一個更佳的任務字詞



圖一、修正及偵測錯誤之流程圖

表示法。傳統表示法中，例如一元表示法(one-hot representation)，可能因為辭典太大導致維度詛咒的問題[12]。因此在本論文中，我們提出同時考慮詞與詞性的新標記，再訓練新標記的詞向量。首先將辨識結果的每個語句詞序列做中文斷詞及標記詞性，並且將字詞及詞性存放在辭典中。文本中的字詞 w_i 與其詞性 p_i 的詞索引值為 $\widehat{w}_i$，可以表示為 $\widehat{w}_i = [w_i; p_j]$ 。經由結合詞以及其詞性得到的新索引，預期增強中文詞彙在不同用法間的鑑別性，再透過預訓練詞向量作為合適的表示法，新的詞向量以 $\mathbf{e}_1, \mathbf{e}_2, \boldsymbol{e}_3 \ldots \boldsymbol{e}_n$ 表示。

2 標記

語音辨識錯誤偵測任務是去評估辨識字和人工轉寫字的比對結果。而在本任務中，我們將語音辨識錯誤偵測任務歸類成三種類別，並且嘗試以機器學習及類神經網路的架構去探討偵測錯誤之效果。而在模型標記方面，我們以辨識結果之文本和人工轉寫文本計算

編輯距離，並標記為三種模式：

- 正確字及錯誤字分類模型標記：正確(C)及錯誤($\bar{C}$)之區塊
- 未刪除及刪除字分類模型標記：未發生刪除錯誤($\bar{D}$)及刪除錯誤(D)之區塊
- 錯誤類型分類模型標記：正確(C)、插入錯誤(I)及替換錯誤(S)之區塊

3 模型

在模型部分，我們探討在不同監督式學習方法中，對於偵測錯誤的效能。其中，我們使用了兩種機器學習方法::支持向量機(SVM)、決策樹(DT)，並以預訓練詞向量(pre-trained word vectors) 作為輸入。為了符合語音轉寫富含時間及序列特性，我們更深入探討了以下幾種方法，在本任務上的效能。如：深層類神經網路(DNN)、遞迴神經網路(RNN)、長短期記憶類神經網路(LSTM)、雙向遞迴神經網路(BRNN)。以詞向量作為輸入，並且與神經網路參數一同訓練。

(二)、 辨識錯誤修正

在本論文中，我們使用萊文斯坦距離(Levenshtein distance)[9]去比較自動語音辨識輸出的音素序列與假設的關鍵詞相似性，而這樣的方式也常被使用在字層級的比對。語音轉寫的錯誤主要分為三種，包含：代替、插入、刪除。當語音辨識中未知詞導致語音錯誤時，可能同時發生代替及刪除的連續錯誤。因此，為了解決連續錯誤導致字詞邊界模糊的問題，我們將使用音素層提升尋找關鍵字的可能性。並且經由我們初步實驗，音素對照法能比文字層級的比對找尋到更細部的差異，由於本論文使用之語料富含之較多領域詞，並且內容通常中英混雜，因此在這樣的情況下，以字層級來做比對是較難符合我們的期待。

萊文斯坦距離能夠簡單找到一組給定句子中最可能的全貌，或是用給定詞彙中最相似的詞來替換識別的單詞。而為了改善並且尋找到更多可能領域詞，我們將在本論文第四節中，我們所實驗的錯誤修正是以設定相似度門檻值為 0.8。(圖一)

表一、　華語會議語料內容介紹

錄製模式	編號代號	語音主題
實驗室錄音	Corpus01	課堂試驗對話
	Corpus02	業務拜訪對話
	Corpus03	語音記事情境
	Corpus04	技術會議對話

四、　實驗

本節中主要是介紹本論文中實驗語料庫與相關實驗設定，第一部分將介紹本論文所使用的實驗語料庫及語料庫分析；第二部分將說明本論文所使用的相關實驗設定；第三部分介紹實驗效能的評估方法；最後將呈現相關實驗結果及觀察。

(一)、　語料庫介紹

本論文使用華語對話及會議語料為台灣師範大學與國內企業的產學合作計畫語料庫，本語料部分語音為改善語音辨識錯誤而重新錄製的實驗室錄音語料。主要由四個不同領域主題內容及兩種不同的錄製模式，其中 Corpus01~Corpus04 為實驗室錄音，實驗室錄音之內容主要選取對話中關鍵詞彙片段錄製，並且由專業人員轉寫與標記。會議參與人數約 7 位語者，本實驗將語料庫分成訓練集、發展集及測試集，主要以語料之總句數比例為 8:1:1。會議語言主要為中文，夾雜少部分英文。

華語會議語料主要為對話或會議交談內容，語音內容是以真實對談或會議模式為主，對話內容無經過特殊設計，所以此語料內容相較於其他語料是較貼進一般領域知識對話或是實際開會內容，然而這樣的內容對於一般語音辨識系統也相對是一個困難的挑戰，其中本語料內容中可能會面臨到以下幾個問題，如：專有名詞、人名、中英文夾雜內容等，並且每位語者的說話模式可能也非常不同，如：發音準確性、語速及音量等，再加

表二、　華語語料庫及自動語音辨識結果

	Corpus01	Corpus02	Corpus03	Corpus04
字數	3878	2593	1267	1665
句數	204	176	323	85
語者數	7	8	8	7
準確率	93.4%	87.5%	77.7%	75.9%
正確字	94.1%	87.8%	79.2%	83.3%
替換錯誤字	4.8%	11.1%	19.3%	15.0%
插入錯誤字	0.7%	0.3%	1.5%	7.4%
刪除錯誤字	1.2%	1.1%	1.5%	1.7%
關鍵字	40	52	36	28

上會議錄音中會議廳的麥克風收音效果及會議室環境下的噪音干擾等等，其實這樣的語料庫是非常具有挑戰性，因此我們也特別為此內容，重新錄製對話，並且嘗試去改善此對話內容對於語音辨識性的困難度，並藉此進一步分析更有效改善辨識結果的方法。

(二)、　實驗設定

語音錯誤檢測的難易度與語音辨識系統的性能表現息息相關，然而語音辨識系統的表現和語者、語音的內容及錄製模式有很大關聯性，因此在語料庫方面，我們詳細分析華語語料的錄製內容及統計語音辨識後的結果。並針對辨識結果做分類，期望能夠預測辨結果的字類型，在語料庫方面我們也將其分成訓練集、發展集及測試集作訓練。

本實驗實作在華語實驗室錄音及會議語料上，並基於 Python 程式語言的函式庫 Scikit-learn[13]、Theano[14]及 Keras[15]等提供機器學習及類神經網路，在第三小節中詞表示法部分輸出值**e** 以 20 維表示。錯誤修正相似度門檻值設定為 0.8。

在本論文的分類問題中，我們將根據表三中的四項指標計算二種評估方式：召回率($Recall$)和準確率($Precision$)，並以 F1 分數($F1-score$)作為本實驗中主要評估。我們

表三、 ROC 分析中的四項指標在辨識偵測任務中的定義

	描述
正確接受 (true positives, TP)	實際上是正確字，並且被分類到正確
錯誤接受 (false negatives, FN)	實際上是錯誤字，並且被分類到錯誤
錯誤拒絕 (false positives, FP)	實際上是錯誤字，並且被分類到正確
正確拒絕 (true negatives, TN)	實際上是正確字，並且被分類到錯誤

將對於正確字及錯誤字偵測結果做評估，因此我們先定義正確字偵測的召回率($Recall_c$)、精準度($Precision_c$)及 F1 分數($F1-score_c$)的計算，反之，錯誤詞之評估方法亦可類推。

在本實驗中，我們以 F1 分數作為本論文研究討論的評估方法，由於 F1 分數能夠同時考慮召回率與精準度，將較於分類器的準確率(Accuracy)評估更能夠看見正反分類之精準度及分類的細節，故在本論文實驗中，我們將以 F1 分數作實驗結果討論的依據。

(三)、 偵測辨識錯誤之實驗結果

本段落主要呈現第三節方法的實驗結果，偵測錯誤類型分類模型中又劃分出兩個子模型，分別是正確字(C)及錯誤字($\bar{C}$)分類模型及未刪除($\bar{D}$)及刪除字(D)分類模型，而偵測錯誤類別主要分成正確字(C)、替換字(S)及插入字(I)，並且在本實驗中，利用模型及語料主題之差異去探討偵測辨識錯誤之議題。

我們比較了分類器在各主題領域之性能表現，並且針對其分類結果計算出表四~表七的 F1 分數，觀察正確及錯誤型態分類的情形，以下我們將以兩個傳統機器學習方法：SVM、Decision tree，和四個類神經網路方法：DNN、RNN、LSTM、BRNN，並更進一步分析及探討其性能表現。做實驗中，若召回率($Recall$)及精準度($Precision$)任一值為

0 將無法計算 F1 分數，因此在實驗表格中將以 -- 表示。

首先，在表四我們能夠觀察到，在實驗室錄音語料中，表現較好分類器為 Decision tree 及 RNN，而更為複雜的 BRNN 類神經網路架構反而在錯誤字偵測上不如前者，甚至 BRNN 的正確字偵測也相對表現較差，我認為可能原因如下：1)由於實驗室錄音是屬於選取重要語句重複錄音，後者較複雜架構可能相對看了比較長的前後文資訊，並非真的有利，因為其實挑選句子重新錄製的文本相對於實際對話文本結構較弱，上下文語句的關聯性也較低，導致其在分類上受到干擾，故無法有效去做正確分類。2)以我們的資料集而言，類神經網路本身需要大量資料來訓練才能夠有較好的效能，相對於小資料集而言，過度複雜的類神經網路反而導致其發生過度擬合的現象。

表四、　比較各模型在 Corpus01~ Corpus04 正確及錯誤區域偵測效能

	Type	SVM	Decision tree	DNN	RNN	LSTM	BRNN
CORPUS01	C	0.9	0.97	0.96	0.96	0.96	0.94
	$\bar{C}$	--	0.85	0.62	0.67	0.61	0.55
CORPUS02	C	0.9	0.97	0.97	0.94	0.95	0.91
	$\bar{C}$	--	0.88	0.84	0.69	0.73	0.59
CORPUS03	C	0.7	0.94	0.98	0.98	0.98	0.98
	$\bar{C}$	--	0.92	0.97	0.97	0.97	0.97
CORPUS04	C	1.0	0.95	0.96	0.86	0.91	0.86
	$\bar{C}$	--	--	0.62	0.70	0.79	0.71

除了探討正確字及錯誤字偵測之外，我們更進一步去討論在辨識器中所發生的刪除錯誤，並且去探討預測刪除錯誤字之議題，而由語料庫分析中我們可以觀察到由於表五，本語料刪除字平均發生機率約 1.4%，所以其實是相對非常罕見的錯誤型態，而在本語料上的刪除字偵測，我們也發現由於大部分字都歸類為未刪除，所以此任務上發現刪除

字是更為重要的效能評估方法，而我們以刪除字的效能表現來看，Decision tree 表現相較於類神經網路更為突出，針對此任務 SVM 及 BRNN 無法有較好的性能表現我認為可能原因為：1)由於大量資訊皆為未刪除字屬於分類類別數量較為極端，因此 SVM 較無法有效作分類。2)錯誤刪除資訊，其實不容易從前後文觀察到，故當我們嘗試使用較複雜的類神經網路架構時，極可能導致其反效果，而沒有良好的效能表現。

表五、　比較各模型在 Corpus01~ Corpus04 未發生刪除及發生刪除錯誤偵測效能

	Type	SVM	Decision tree	DNN	RNN	LSTM	BRNN
CORPUS01	$\overline{D}$	0.93	0.95	0.98	0.99	0.99	0.98
	D	--	0.50	0.22	0.44	0.44	0.5
CORPUS02	$\overline{D}$	0.97	0.97	0.99	0.99	0.99	0.99
	D	--	0.66	0.66	0.66	0.66	0.66
CORPUS03	$\overline{D}$	0.96	0.98	0.95	0.96	0.94	0.98
	D	--	0.66	0.33	0.57	0.44	0.8
CORPUS04	$\overline{D}$	0.95	1.0	0.99	0.99	0.99	0.98
	D	--	1.0	0.66	0.72	0.66	--

討論完本實驗之子模型之後，我們將由實驗更深入討論辨識錯誤相關問題，並且由不同分類方法及語料去探討去偵測之難易度，而我們由表六可觀察發現除了 Corpus04 之外，實驗室錄音語料平均插入率為 0.83%。在實驗室錄音語料上，RNN 的分類效果最為突出，而傳統的機器學習方法都表現較差之外，本任務在插入錯誤偵測上平均表現都較為普通，主要原因為：1)由於實驗室錄音品質相對較好，所以在語音辨識上較上出現插入錯誤的情形，而由語料庫探討中我們也能觀察到平均插入率為 0.83%，相較於其他錯誤是較為少見的錯誤類型。在替換錯誤偵測上，我們也觀察到一個有趣的現象，經常被辨識錯誤且被替換的字似乎可以從一些規則中看見，例如：某字詞常被替換成其他幾

種字詞，而藉此發現我們也觀察到，以時間序列且長記憶性的神經網路在偵測替換字時能夠有很不錯的表現，這對於我們在第二步驟的錯誤修正是非常有利的一種現象。

表六、　比較 Corpus01~ Corpus04 錯誤型態偵測效能

	Type	SVM	Decision tree	DNN	RNN	LSTM	BRNN
CORPUS01	C	0.54	0.88	0.96	0.98	0.97	0.97
	S	--	0.23	0.38	0.75	0.5	0.64
	I	--	--	--	0.33	--	--
CORPUS02	C	1.0	0.87	0.97	0.97	0.96	0.96
	S	--	0.29	0.77	0.79	0.71	0.72
	I	--	--	--	0.33	--	--
CORPUS03	C	1.0	0.87	0.97	0.97	0.96	0.96
	S	--	0.29	0.77	0.79	0.71	0.72
	I	--	--	--	0.33	--	--
CORPUS04	C	0.82	0.85	0.91	0.96	0.94	0.94
	S	--	--	0.62	0.84	0.76	0.79
	I	--	--	0.18	0.33	--	--

(四)、　辨識錯誤修正之實驗結果

在表七中，我們做了錯誤修正的基礎實驗稱為音素比對法(Phone Match)簡稱為 PM 以及改良方法簡稱為 IMP_PM，如同第三節所描述方法，我們使用音素比對法來去尋找與關鍵詞相似的位置，但由實驗中觀察到，此方法在某些語料上容易產生假警報(false alarm)。為了改善這個問題，我們將偵測辨識錯誤的結果作為此部分的參考值，若我們偵測此區域發生辨識錯誤，才以關鍵字詞表做為替換的候選詞，並以音素比對法找出最相似的關

鍵字詞。而由我們在基礎實驗中的關鍵字修正表現就能達到平均召回率約 78%、精確率約 87%，然而我們更進一步做修正改善，並且呈現出更好性能表現平均召回率約 78%、精確率約 90%，有效提升 3%領域詞精確率，並且改善語音辨識文本的錯誤。

表七、　比較以音素比對方法修正辨識錯誤之效能

Corpus Name	evaluation	PM	IMP_PM
Corpus01	Precision	75.20%	82.00%
	Recall	94.50%	94.50%
Corpus02	Precision	87.60%	90.00%
	Recall	87.90%	87.90%
Corpus03	Precision	94.80%	97.00%
	Recall	91.60%	91.60%
Corpus04	Precision	93.20%	93.20%
	Recall	39.90%	39.90%

五、　結論與未來展望

本論文探討一般的語音辨識系統應用於特定領域的對話中導致的辨識錯誤，並且提出了兩步驟改善措施，其中包含了辨認錯誤區域和修補毀損內容。在第一步驟中，我們探討了序列標記的方法應用於錯誤檢測的效能，在實驗中我們發現利用有時間序列及記憶的遞迴神經網路對於錯誤偵測是非常有幫助的；在第二步驟中，我們以第一步驟的標記結果作為依據，並以特殊領域的關鍵詞表與錯誤字做音素比對。經由我們的兩階段改錯方法，能夠有效提高關鍵字修正的精確率，並且降低原本音素對照法造成假警報所產生的問題。未來我們希望能夠針對辨識錯誤及未知詞做更進一步的探討及分析，並且加入語句及語意資訊強化偵測模型，讓修正錯誤字能夠有更穩定的效能表現。本論文期望提出一個改善架構，來解決未知詞所導致文本語意不清的問題。

參考文獻

[1] G.Hinton *et al.*, “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, 2012.

[2] J.Li, L.Deng, Y.Gong, andR.Haeb-Umbach, “An overview of noise-robust automatic speech recognition,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 22, no. 4. pp. 745–777, 2014.

[3] L.Qin, “Learning Out-of-Vocabulary Words in Automatic Speech Recognition,” 2013.

[4] A.Ogawa andT.Hori, “Error detection and accuracy estimation in automatic speech recognition using deep bidirectional recurrent neural networks,” *Speech Commun.*, vol. 89, pp. 70–83, 2017.

[5] L.Qin, M.Sun, andA.Rudnicky, “OOV Detection and Recovery using Hybrid Models with Different Fragments,” no. August, pp. 1913–1916, 2011.

[6] Y.Kim, H.Franco, andL.Neumeyer, “Automatic pronunciation scoring of specific phone segments for language instruction,” in *Proc. of EUROSPEECH*, 1997, vol. 97, pp. 649–652.

[7] L. F.D’Haro andR. E.Banchs, “Automatic correction of ASR outputs by using machine translation,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 08–12–Sept, pp. 3469–3473, 2016.

[8] P.Fayolle, J., Moreau, F., Raymond, C., Gravier, G., & Gros, “CRF-based combination of contextual features to improve a posteriori word-level confidence measures.,” *Elev. Annu. Conf. Int. Speech Commun. Assoc.*, 2010.

[9] J.Twiefel, T.Baumann, S.Heinrich, andS.Wermter, “Improving domain-independent cloud-based speech recognition with domain-dependent phonetic post-processing,” in *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI-14)*, 2014, pp. 1–7.

[10] F.Bechet andB.Favre, “ASR error segment localization for spoken recovery strategy,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2013, pp. 6837–6841.

[11] R.Bigot, B., Senay, G., Linares, G., Fredouille, C., & Dufour, “Person name recognition in ASR outputs using continuous context models.,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 8470–8474, 2013.

[12] R.Rosenfeld, “Optimizing lexical and n-gram coverage via judicious use of linguistic data,” in *Fourth European Conference on Speech Communication and Technology*, 1995, pp. 1763–1766.

[13] F.Pedregosa andG.Varoquaux, *Scikit-learn: Machine learning in Python*, vol. 12. 2011.

[14] J.Bergstra *et al.*, “Theano: a CPU and GPU Math Expression Compiler,” *Proc. Python Sci. Comput. Conf.*, pp. 1–7, 2010.

[15] F.Chollet, “Keras,” *GitHub*, 2015. [Online]. Available: https://github.com/fchollet/keras.