新穎基於預訓練語言表示模型於語音辨識重新排序之研究

Innovative Pretrained-based Reranking Language Models for *N*-best Speech Recognition Lists

邱世弦 Shih-Hsuan Chiu, 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
{shchiu, berlin}@ntnu.edu.tw

摘要

本論文提出兩種基於 BERT 的語言排序模型，用於準確地重新排序自動語音辨識之候選結果(通常以 *N*-best 列表的形式來表示)。在過往的研究中，已經證明對於從聲學模型解碼出的 *N*-best 列表進行重新排序可以明顯改善兩階段式的語音辨識系統。然而在另一方面，隨著眾多從巨量文本預訓練的上下文語言模型的興起，它們在自然語言處理的領域中都達到了最先進的效能，像是問答系統或是機器翻譯，但是卻缺少有學者探討此種預訓練語言模型對於語音辨識的有效性。因此在本論文中，我們採用了 BERT 以開發簡單而有效的方法，對 *N*-best 列表進行重新排序。具體而言，我們可以將 *N*-best 重新排序問題視為 BERT 模型的下游任務，並提出了兩種基於 BERT 的語言排序模型，分別稱為(1) uniBERT: 給定一個 *N*-best 列表，輸出最理想的一連詞(Ideal Unigram)，(2) classBERT: 給定一個 *N*-best 列表，視為一道選擇題，輸出最好的候選排名(Oracle 當前名次)。這些模型試圖撼動 BERT 之強大之力僅僅透過一層附加輸出層，來重新排序第一階段語音辨識產生的 *N*-best 列表。我們評估模型於 AMI 會議語料庫，並實驗出比廣泛使用且堅強的基準 LSTMLM 改進了多達 3.14%的 WER 相對下降率。

關鍵詞：自動語音辨識，語言模型，BERT，*N*-best 列表重新排序

Abstract

This paper proposes two BERT-based models for accurately rescoring (reranking) *N*-best speech recognition hypothesis lists. Reranking the *N*-best hypothesis lists decoded from the acoustic model has been proven to improve the performance in a two-stage automatic speech

recognition (ASR) systems. However, with the rise of pre-trained contextualized language models, they have achieved state-of-the-art performance in many NLP applications, but there is a dearth of work on investigating its effectiveness in ASR. In this paper, we develop simple yet effective methods for improving ASR by reranking the *N*-best hypothesis lists leveraging BERT (bidirectional encoder representations from Transformers). Specifically, we treat reranking *N*-best hypotheses as a downstream task by simply fine-tuning the pre-trained BERT. We proposed two BERT-based reranking language models: (1) uniBERT: ideal unigram elicited from a given *N*-best list taking advantage of BERT to assist a LSTMLM, (2) classBERT: treating the *N*-best lists reranking as a multi-class classification problem. These models attempt to harness the power of BERT to reranking the *N*-best hypothesis lists generated in the ASR initial pass. Experiments on the benchmark AMI dataset show that the proposed reranking methods outperform the baseline LSTMLM which is a strong and widely-used competitor with 3.14% improvement in word error rate (WER).

Keywords: Automatic Speech Recognition, Language Models, BERT, *N*-best Lists Reranking

一、緒論

近年來，在眾多新穎精緻的神經網路引入下，自動語音辨識(Automatic Speech Recognition, ASR, 亦簡稱語音辨識)得到了快速而活躍的進展，基於語音辨識的各種應用(包括語音搜尋或是口語對話系統)因而取得了巨大的進步[1, 2]。儘管他們取得了很大的進步，但是在某些情境下，例如在嘈雜的環境中或在隨性風格(Casual-style)口語中執行語音辨識，普遍 ASR 系統的準確性仍然無法令人滿意[3, 4]。

在某些任務或應用中需要高準確率的 ASR，就採用了多個語音辨識候選假設(Hypotheses) (詞序列)，這些後選詞序列會以某種形式表示，例如詞圖(Lattice or Word Graph)、*N* 最佳列表(*N*-best List)或是詞混淆網路(Word Confusion Network, WCN) [5]來顯現。進行候選假設是因為第一階段的語音辨識結果(1-best)可能會包含許多錯誤在上述的嚴重情境中，但是如果經過重新排序(Reranking or Rescoring)，則從多個候選句中，都可以找到詞錯誤率(Word Error Rates, WERs)明顯低於 1-best 的其他候選句。舉例來說，在噪音環境的語音辨識任務 CHiME-4 [6]，在 ASR 的最後階段，就使用了遞迴神經網路語言模型(Recurrent Nerual Network Language Models, RNNLMs)，執行 *N*-best 或是詞圖的重新排序。*N*-best 列表也在口語對話系統被採用[7, 8]。

在本論文，我們專注於**語音辨識候選 *N*-best 之重新排序。**目前，最廣為使用進行

N-best 重新排序的模型為 RNNLMs [9, 10] (之後都內涵了 LSTM cell [11], 亦可稱為 LSTMLMs) [12]，此模型在近幾年達到了最先進的效能，比起稱霸多年的基於頻率計數的傳統回退 *n* 連詞模型(Back-off *n*-gram) [13, 14, 15]有更大的改善，這是因為 RNNLMs 能夠考慮到更長的上下文資訊(Long-term Context)。隨後，許多研究專注於探索 LSTMLMs 的調適(Adaptation)方法，以進行更準確的 *N*-best 重新排序。但是要注意的是，即使 LSTMLMs 在 *N*-best 重新排序表現傑出，但它最初是為了預測下一個單詞而開發的，而不是為了 *N*-best 重新排序任務而開發的。

在另一方面，隨著自然語言處理(Natural Language Processing, NLP)的技術大量發展，當前 ASR 系統用於評估 *N*-best 候選句的語言與語意合法性的資訊還是相當有限。在自然語言處理的領域中，許多膾炙人口的預訓練語言表示法模型 (Pre-trained Language Representation Models)，在近幾年如雨後春筍般的湧出，像是 ELMO (Embeddings from Language Models) [16]，GPT (Generative Pre-Training Transformer) [17]，BERT (Bidirectional Encoder Representations from Transformers) [18]…等等預訓練模型，來提取上下文相關(Context-dependent or Contextualized)的詞嵌入(Word Embedding)，此種 Contextualized 詞嵌入已經被證實在眾多下游 NLP 的任務下達到了最先進的效能，像是口語語言理解[19]、文本分類[20]和問答任務[21]…等等。然而，據我們所知，鮮少有相關研究探討將上述的預訓練語言模型，應用於 ASR 系統中並探討其有效性。因此在本論文，我們嘗試利用 Google 近來提出的 BERT 來對從 ASR 第一階段產生的 *N* 最佳候選列表(*N*-best List)，執行重新排序，希望提高 ASR 的效能。

我們提出了兩種基於 BERT 的語言排序模型，都是將 *N*-best 重新排序視為 BERT 的下游任務，都是基於在 BERT 之上，僅僅疊加一層全聯接層(Fully Connected Layer, FC)，分別稱為(1)uniBERT: 給定一個 *N*-best 列表，輸出理想的一連詞(Ideal Unigram)和(2)classBERT: 給定一個 *N*-best 列表，輸出最好的候選排名(Oracle 的排名，Oracle 代表的是與該正確文句做計算，WER 最小的那條候選句)，這兩種模型將會在第四章做詳細的介紹。在 BERT 的預訓練階段中，主要對模型進行訓練以從上下文，來預測被遮蔽的單詞，以使模型能夠“融合”左和右的表示，與以前的 bi-LMs (包括 bi-RNNLMs) [22, 23]不同，後者使用各方向的獨立編碼表示來淺層連接(Shallow Concatenation)，因此可能會限制 bi-RNNLMs 的潛力。有鑑於此，我們認為 BERT 對於 *N*-best 列表重新排序是有前途的，因而提出了兩種基於 BERT 的語言排序模型，這些模型試圖借助 BERT 之力僅通過微調(Fine-tuning)一層附加的輸出層。我們在基準語料庫 AMI 上評估我們的模型，並

表明所提出的模型比強大且廣泛使用的 LSTMLM 獲得了更好的性能。

二、文獻回顧

在本節中，我們將簡要回顧有關 ASR 系統中 *N*-best 重新排序方法的先前研究。隨著近年來深度神經網絡的興起，RNNLM (LSTMLM) [12]直接稱霸語言模型界成為流行且廣泛使用於 *N*-best 重新排序，遠勝過傳統的統計式 *n*-gram 模型[13, 14, 15]，因為前者能考慮更長距離的資訊。因此有許多研究都集中在探索 LSTMLM 的調適方法，以進行更準確的 *N*-best 重新排序，像是有一些研究利用**歷史資訊**(History Information)對 RNNLM 作語言模型調適[24, 25]。而有更多研究專注於對**主題資訊**(Topic Information)作語言模型調適，例如 Mikolov [26]使用上下文感知向量(Context-aware Vectors)作為 RNNLM 的額外輸入，以適應大範圍的主題資訊。同樣地，Chen [27]探究主題建模方法，以提取主題特徵作為 RNNLM 的附加輸入，用於多類型廣播轉錄任務中的類型和主題的調適。Ma [28]探索了基於 LSTMLM 的三種微調策略。Lam [29]對 LSTMLM 的激活函數(Activation Function)作高斯分佈處理(Gaussian Process)，得到了些微的進步。Irie [30]提出了一種基於 LSTM 的動態加權之混合器(Mixer)，各個主題模型在特定領域(Specific Domain)上分別進行訓練，並擁有動態的權重，可以勝過簡單的線性插值。之後，Li [31]使用上述前者的方法，但他改成使用基於 Transformer 的 LM 與加權混合器。

但要注意的是，即使 LSTMLM 在 *N*-best 重新排序方面表現出色，但它最初的設計是為了預測下一個單詞而開發的，而不是為進行 *N*-best 重新排序任務而開發的。所以有研究者直接提出專為 *N*-best 重新排序任務而設計的模型。像是鑑別式語言模型(Discriminative Language Models, DLM) [32-35]最初就是為 *N*-best 重新排序而開發的，它利用 ASR 的錯誤資訊來訓練鑑別式語言模型。Ogawa [36]受 DLM 啟發，但認為其損失函數(Loss Function)的設計很複雜，因此他們開發了一個簡單的編碼器-分類器模型（Encoder-Classifier Model），該模型訓練一個分類器進行一對一的候選句比較(氣泡排序(Bubble Sort))來執行 *N*-best 重新排序。Tanaka [37]提出了一種將端到端(End-to-End)ASR 系統視為一個神經語音到文本語言模型(Neural Speech-to-Text LMs, NS2TLM)的想法，該模型以輸入的聲學特徵為條件，並將其用於對 DNN-HMM hybrid ASR 系統中生成的 *N*-best 進行重新排序。Song [38]受資訊檢索(Information Retrieval, IR)中的核心問題，即排名學習(Learning-to-Rank, L2R)的啟發，提出了重新計分學習（Learning-to-

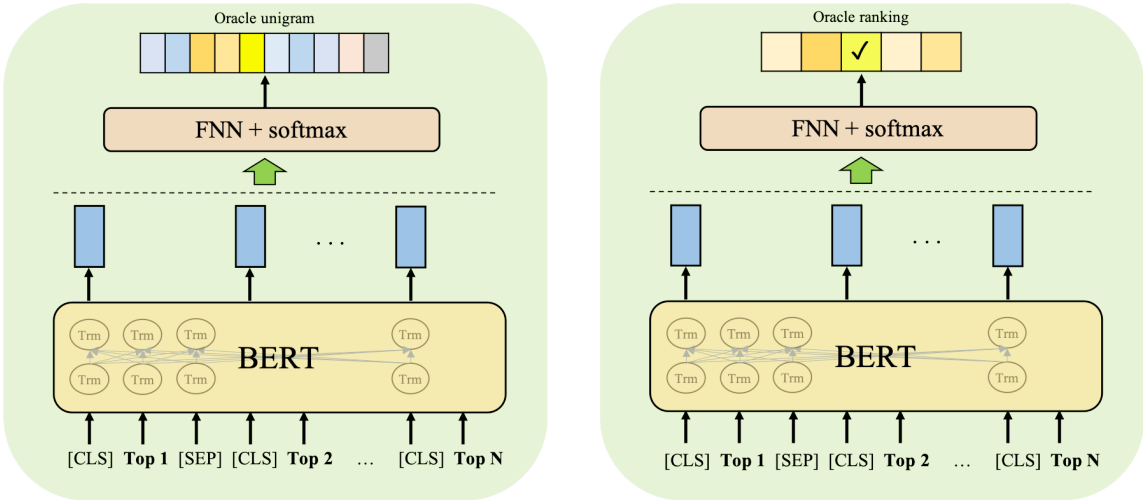
Rescore, L2RS）機制，這是第一個研究將 *N*-best 重新排序視為一種學習問題(Learning Problem)。亞馬遜 Gandhe [39]在 2020 年提出一種基於注意力機制(Attention-based)的鑑別式語言模型，該模型在訓練鑑別式 LSTM 時會同時考慮到詞與聲音的特徵，後者是借助端對端系統，把當前詞與聲音片段藉由注意力機制所獲取，該方法也得到了顯著的效果。

近年來，隨著預訓練語言表示法模型的興起且蔚為風潮，像是 BERT [18]，在 2018 年一問世直接打破 12 項最先進效能的 NLP 任務。因此在 2019 年，Wang [40]首度提及可以將 BERT 拿來作句子評分，但是他們並沒有以實驗證明。不久後，首爾大學 Shin 等人[41]，以預測每個位置遮蔽處([MASK])的機率加總作為句子的分數，雖然相較傳統的 LM 它不是真正的句子機率，但仍然可以作為 *N*-best 重新排序。隨後，亞馬遜 Salazar 等人[42]，也作[MASK]處的機率加總，把句子分數定義為 PLL (Pseudo-log-likelihood)，並以數學形式證明該方法的有效性，並提出為該方法執行加速，藉由知識蒸餾(Knowledge Distillation)訓練一個不用預先遮蔽([MASK])某處單詞的加速版句子評分模型，雖然效果變差了，但把計算複雜度從$O(|W| \cdot V)$ (其中$|W|$是句子長度，V是詞典大小)降到$O(1)$，且比較了多種預訓練模型的效能。首爾大學 Shin 等人[43]，受到 Salazar 的啟發，不同於原始 BERT 要重複性的遮蔽再預測(Mask-and-Predict Repetition)，在 BERT 內部 self-attention 處利用對角遮蔽(Diagonal Masking)就能達到原先功能且加快速度，而實驗證明，該方法不僅為 *N*-best 重新排序加快了 6 倍時間甚至連效能也得到提升。有鑑於此，我們也提出了兩種基於 BERT 的語言排序模型，這些模型希望撼動 BERT 強大之力透過僅僅一層附加輸出層，都將會在第四章節作詳細介紹。下一章節將會詳細介紹 BERT。

三、BERT

BERT (Bidirectional Encoder Representations from Transformers)是一個近期發表且廣為使用的語言表示模型，內部由多層雙向的 Transformer [44]的編碼器(Encoder)所組合起來。而 Transformer 是一個基於注意力機制(Attention-based)的結構，且能夠考慮全域的輸入與輸出的相依性。訓練 BERT 分成兩個階段，分別是預訓練(Pre-training)和微調(Fine-tuning)。在預訓練階段，有兩個訓練準則且同時訓練於大量且廣泛主題的無標注文本資料，一個是遮蔽語言建模(Masked Language Modeling, MLM)，另一個則是下句預測(Next Sentence Prediction , NSP)，前者是一種填空任務，學習如何從過去和未來的上下

文(Context)預測出被遮蔽的詞是什麼詞彙，後者學習兩兩之間的句子是否有連貫性(Contiguous)。BERT 特別受益於 MLM 的訓練方式，因為它能夠“融合”(Fuse)歷史與未來詞的資訊，不像傳統的 LM 只考慮先前詞，或是 bi-RNNLM 只能淺層連接(Shallow Concatenation)兩個方向的上下文資訊。在微調(Fine-tuning)階段，預訓練完成的 BERT 僅僅只要附加一層輸出層(Output Layer)針對特定任務視為下游任務，就能撼動強大的 BERT。輸出層從頭訓練(Training from Scratch)，而 BERT 本體的參數會被”微調”。BERT 在多項 NLP 的領域得到了最新進的效能，包括問答系統(Question Answering, QA), 自然語言推論(Natural Language Inference, NLI), 神經機器翻譯(Neural Machine translation, NMT) 和語音文件檢索(Spoken Document Retrieval, SDR)…等等。因此我們也將 *N*-best 重新排序視為一種 NLP 任務，認為 BERT 是一個有潛力的雙向語言模型(bi-LMs)，將 *N*-best 重新排序作為 BERT 的下游任務。



圖一：uniBERT 與 classBERT 的架構

四、基於 BERT 之語言排序模型(BERT-based Reranking Language Models)

本論文著重於嘗試借助 BERT (本論文採用“bert-base-uncased”版本)之力應用於第一階段之語音辨識結果 *N*-best 重新排序。在本節中，我們將提出兩種基於 BERT 的 *N*-best 重新排序模型，分別稱為 uniBERT 和 classBERT。簡單來說，首先使用預訓練的 BERT 參數初始化該模型，然後使用標記好的訓練資料，僅附加一層額外的輸出層即可對預訓練的 BERT 進行微調(Fine-tuning or Adaptation)。具體而言，輸出層將從頭訓練(Training From Scratch)，而預訓練的 BERT 將進行微調。

4.1 uniBERT

我們在原始(Vanilla)預訓練的 BERT 之上疊加了一層前饋式全連接層(Feed-forward Neural Network, FNN)，直接輸出 *V* 維(詞典大小)最理想的一連詞(Ideal Unigram)。具體來說，給定一組 *N*-best 列表，模型能夠在 *N*-best 列表中輸出“最好” (Oracle)的候選句的 unigram。我們希望利用 BERT 來萃取出 *N*-best 列表中，多個候選句中的詞與詞甚至是句與句之間的關係，輸出到一個理想的 unigram，我們稱此模型為 uniBERT。模型架構如圖一之左圖所示。在微調(Fine-tuning)階段，一次輸入 *N* 句(在本論文實驗設定為 *N* = 10)候選句，並在每個候選句的頭跟尾分別加入特殊符號[CLS]和[SEP])，uniBERT 要學習如何輸出最理想的 unigram。而模型輸入 *N*-best 後的流程如下所述，每一候選句會先藉由[CLS] token，BERT 自動編碼成句子表示法h_k(如圖 1 藍色長方形所示)，而全部的表示法h_k會互相逐項(Element-wise)的取平均，或是相連接起來(在實驗中會比較該兩種方法)，再經過一層線性分類層(FC layer)和 softmax 讓此 unigram 正規化(滿足機率公設，總和為 1)，就能得到理想的 unigram。uniBERT 的演算過程以數學式表示如下:

$$\begin{gathered}[h_1, h_2, \dots, h_{10}] = \text{BERT}\left(\left[hyp^1_{[CLS]}, hyp^2_{[CLS]}, \dots, hyp^{10}_{[CLS]}\right]\right) \\ h^{nb} = \text{Average}([h_1, h_2, \dots, h_{10}])\ or\ \text{Concat}([h_1, h_2, \dots, h_{10}]) \\ z^{nb} = \text{linear}(h^{nb}) \\ P_{bert_{uni}}(\cdot\,|h^{nb}) = \text{Softmax}(z^{nb})\end{gathered} \tag{1}$$

其中$P_{bert_{uni}} = uniBERT(nb) \in R^V$是模型輸出最理想的 unigram，$nb$ 為一組 *N*-best 列表。而訓練模型的資料收集於每個訓練文本(語句人工轉錄)解碼出的 *N*-best 列表中 WER 最低的那條候選句(與正確文本做計算)，並且創造它的 unigram 表示法，舉例來說，例如 Oracle 候選句是:“我 愛 你 你 愛 我 媽”，unigram 表示法為$P_{ora_{uni}} = [\dots, 0, \frac{2}{7}, \frac{2}{7}, 0, 0, \frac{2}{7}, 0, \frac{1}{7}, 0, \dots]$。訓練準則(Training Criterion)使用 Kullback-Leibler (KL)散度:

$$L = D_{KL}\left(P_{ora_{uni}} \middle| P_{bert_{uni}}\right) = \sum_{w \in Vocab} P_{ora_{uni}}(w) \log\left(\frac{P_{ora_{uni}}(w)}{P_{bert_{uni}}(w)}\right) \tag{2}$$

此模型對於每筆訓練資料去做最小化 KL 散度的訓練，找到模型最佳化參數。在測試階段時，我們可以使用$P_{bert_{uni}}$去替代或是插值於原本的語言模型分數，去為第一階段的語

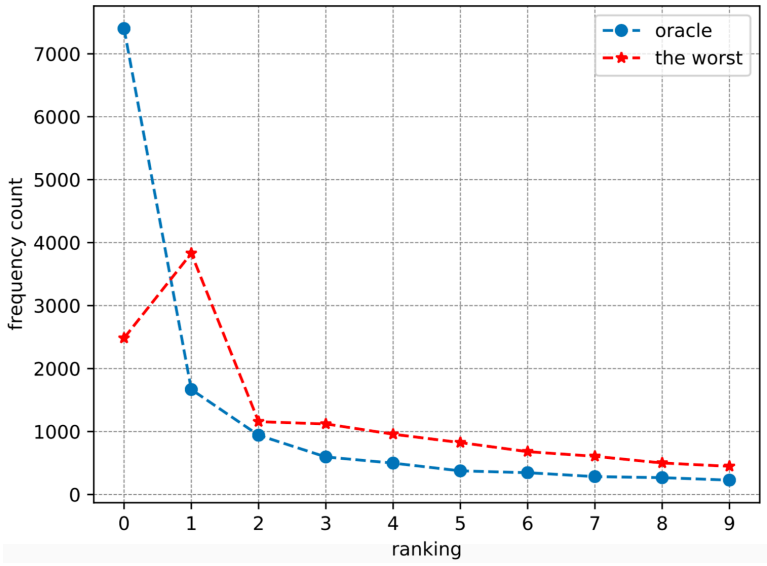
音辨識產生的 *N*-best 列表之每一候選句進行重新打分(Rescoring)，希望改善語音辨識效能。

4.2 classBERT

classBERT 模型結構與 uniBERT 非常相似，如圖一之右圖所示，唯一的區別是，給定一組 *N*-best 列表，它會直接輸出 Oracle 候選的目前所在位置(也就是名次)。在微調階段，classBERT 要學習如何輸出 Oracle 候選的排名。形式上，我們創建訓練資料$\mathcal{T} = \{(nb_1, y_1), \ldots, (nb_t, y_t), \ldots, (nb_T, y_T)\}$來訓練 classBERT，其中$nb_t$是某句訓練語句(Utterance) U_t的 *N*-best 列表，而y_t是在nb_t中 Oracle 候選句的當前排名位置，其中y_t會以 one-hot 的形式$P_{y_t} \in R^N$表示，N是 *N*-best 列表的大小(本論文採用 $N = 10$)。模型嘗試學習如何輸出 Oracle 候選的排名，訓練準則是交叉熵(Cross Entropy)損失函數，廣泛應用於許多多類別分類(Multi-class Classification)問題中：

$$L = H\left(P_y, P_{bert_{cls}}\right) = -\sum_{i}^{N} P_y(i)\log P_{bert_{cls}}(i) = -\log P_{bert_{cls}}(k) \tag{3}$$

其中$P_{bert_{cls}} = classBERT(nb) \in R^N$是模型輸出，$k$是預期的 Oracle 排名位置。classBERT 學習在每筆訓練資料去最小化 Cross Entropy。在測試階段，給定一組 *N*-best 列表，我們



圖二：表示 Oracle 排名的頻率分佈和 Worst 排名的頻率分佈，分別由藍線和紅線顯示

就可以直接輸出哪一句為“最佳”的候選句，希望藉此改善第一階段語音辨識效能。

表一：AMI 的評估集於第一階段的語音辨識結果(1-pass)、Oracle、隨機選擇、最差的 WERs，第二行顯示 3-Gram 困惑度(PPL)

	1-pass	Oracle	Random	Worst
WER	22.79	14.39	29.28	40.65
3G PPL		154.62		

五、語音辨識實驗

我們評估模型於 AMI 會議語料庫[45]上，這是一個眾所周知的基準(Benchmark)語音辨識語料庫，內含 100 個小時的會議對話記錄。100 小時的音檔用來訓練 DNN-HMM 結構的聲學模型$p(X|W)$，以及相應的轉錄文本(總共 108221 條語句)以訓練基於 Kneser-Ney (KN) [13]平滑技巧的 3-Gram 語言模型$P(W)$。這兩個分開訓練的模型構成了我們基礎的第一階段 ASR 系統。在本論文中，基礎 ASR 系統是使用語音辨識實驗的著名工具包 Kaldi [46]搭建的。本論文中，我們致力於將我們提出的方法（基於 BERT 的兩種新穎的重排模型）應用於 *N*-best 重排任務來改善第二階段的 ASR。首先，我們會使用維特比動態規劃搜尋(Viterbi Dynamic Programming Search)對第一階段 ASR 系統在評估集(Evaluation Set, 12612 條待測語句(音檔))建立的詞圖(Lattice or Word Graph)進行解碼，從而獲得每個音檔的前 *N* 個最佳候選列表 (*N*-best list，本論文採用 $N = 10$)。因此，我們能夠利用更高階的語言模型，例如: NNLMs (在本論文是使用提出的兩種模型)來替換或內插語言模型分數$P(W^{Hyp})$，並與相應的聲學模型分數$p(X|W^{Hyp})$結合以重新排列 *N*-best 列表：

$$W^* = \underset{W^{Hyp} \in N-best}{\operatorname{argmax}} \; p(X|W^{Hyp})P(W^{Hyp}) \tag{4}$$

期望獲得更好的辨識結果(詞序列) W^*。表一表示 ASR 系統在評估集第一階段的（使用 3-Gram LM）WER，而 Oracle WER 是 10-best 的理論上限(Ceiling Performance)，表示每個測試語句都選擇 WER 最低的候選句，Random 表示每個測試語句都隨機選擇一條候選句，Worst 表示每個測試語句都選擇 WER 最高的候選句，第二行顯示 3-Gram 在第一

階段語音辨識的困惑度(Perplexity, PPL)。圖二表示 Oracle 排名的頻率分佈和 Worst 排名的頻率分佈，分別由藍線和紅線顯示。

表二：uniBERT 應用於 AMI 語音辨識的結果(WERs)

AM = 27.55	LM	AM + 10 * LM
1-pass	26.80	22.79
LSTM	25.00	21.33
uniBERT	26.84	22.86
LSTM + uniBERT	25.12	**21.24**

5.1 uniBERT 之 *N*-best 重新排序實驗

在第一個提出的語言重排模型 uniBERT 中，當我們向模型輸入一組 *N*-best 列表，它會輸出理想的 unigram 語言模型$P_{bert_{uni}}(W)$。 此 unigram LM 可用於重新計分語言模型$P(W)$分數：

$$P(W^{Hyp}) = \alpha P_{rnn}(W^{Hyp}) + (1-\alpha)\left(\beta P_{bert_{uni}}(W^{Hyp}) + (1-\beta)P_{tri}(W^{Hyp})\right) \quad (5)$$

其中$P_{tri}(W)$是第一階段語音辨識的 3-Gram 語言模型，然後與模型輸出$P_{bert_{uni}}(W)$用係數β進行線性插值，分配兩者模型的相對貢獻，在本研究中我們在發展集(Developing set)中調配出最好的效能為$\beta = 0.2$或是 0.1。此外，我們還使用 RNNLM (LSTMLM)的分數$P_{rnn}(W)$與上述組合後的分數做線性插值，用自由超參數α來分配彼此的貢獻，並根據經驗法則將其設置為$\alpha = 0.7$或是 0.8。在這部分的實驗，我們主要是期望利用 BERT 萃取出的理想 unigram 來輔助最先進的基準 LSTMLM，並提供額外的資訊，例如詞頻。如

表三：classBERT 應用於 AMI 語音辨識的結果(WERs)

AM = 27.55	LM	Consider AM and LM	
1-pass	26.80	22.79	
LSTM	25.00	21.33	
classBERT	**23.18**	+2-dim	+1-dim
classBERT+3G	–	21.69	**21.27**
classBERT+LSTM	–	**20.66**	21.61

表二所示，雖然單獨使用 uniBERT 輸出的 unigram 不會直接改善 ASR 性能，但可以輔助 LSTMLM 並使其(LSTMLM)改善 0.2% 的 WER 相對下降率。

5.2 classBERT 之 *N*-best 重新排序實驗

在第二種提出的語言重排模型 classBERT 中，給定一組 *N*-best 列表，模型能直接選擇出哪一條是最佳(Oracle)的候選句。如表三所示，有三種實驗設定，第一種是 classBERT 僅考慮文本(候選句)，並且勝過基準 LSTMLM 相對減少了 7.28%的 WER。第二種是我們考慮了 ASR 的兩種分數(聲學分數和語言模型(3G 或是 LSTMLM)分數)，在 BERT 編碼出的候選句嵌入的頂端連接(Concatenating)該二種分數(即[CLS]的 768-dim + 2-dim)作為特徵，該方法在加入聲學和 LSTMLM 分數時，比基準的 LSTMLM 進步了 **3.14%**的 WER 相對下降率。第三種方法是將 AM 和 LM 得分利用我們的先備知識(即 AM + 10 * LM)事先結合起來，成為了單個分數，然後我們如同前者的方法，把該分數連接在候選句嵌入的頂端(也就是[CLS]的 768-dim + 1-dim)，該方法在加入 3G 分數時比原先沒有先結合(AM + 10 * LM)獲得了改善，也比 LSTMLM 進步了 0.3%的 WER 相對下降率。

六、結論與未來展望

在本文中，我們提出了兩種基於 BERT 之 *N*-best 重新排序模型，分別是 uniBERT 與 classBERT。uniBERT 給定一組 *N*-best 列表，輸出最理想的 unigram，而 classBERT 將 *N*-best 重新排序視為一道選擇題。我們已經通過實驗證實了兩者優異的 *N*-best 重新排序效能。這些方法都是通用框架，可以應用於使用 *N*-best 列表形式作為候選假設的其他研究領域，例如:機器翻譯(Machine Translation, MT)和資訊檢索(Information Retrieval, IR)。

在未來的研究中，我們計畫像以往的研究[32-35、47-48]一樣，通過使用鑑別式訓練(Discriminative Training)，主要是利用 ASR 的錯誤當作特徵，來提高語言排序模型的效能。我們還希望考慮語者之前所說過的內容(歷史資訊)，來幫助預測當前的話語。

參考文獻

[1] G. Hinton, L. Deng, D. Yu, G.E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, and B. Kingsbury, “Deep neural networks for acoustic modeling

in speech recognition: The shared views of four research groups," IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[2] D. Yu and L. Deng, Automatic speech recognition: A deep learning approach, Springer-Verlag London, 2015.

[3] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 4, pp. 745–777, Apr. 2014.

[4] T. Hori, S. Araki, T. Yoshioka, M. Fujimoto, S. Watanabe, T. Oba, A. Ogawa, K. Otsuka, D. Mikami, K. Kinoshita, T. Nakatani, A. Nakamura, and J. Yamato, "Low-latency realtime meeting recognition and understanding using distant microphones and omni-directional camera," IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 2, pp. 499–513, Feb. 2012.

[5] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: Word error minimization and other appli- cations of confusion networks," Computer Speech and Language, vol. 14, no. 4, pp. 373–400, Oct. 2000.

[6] E. Vincent, S. Watanabe, A.A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," Com- puter Speech and Language, vol. 46, pp. 535–557, Nov. 2016.

[7] J.D. Williams, "Exploiting the ASR N-Best by tracking mul- tiple dialog state hypotheses," in Proc. Interspeech, 2008, pp. 191–194.

[8] S. Young, M. Gasˇic´, B. Thomson, and J.D. Williams, "POMDP-based statistical spoken dialogue systems: A re- view," Proc. IEEE, vol. 101, no. 5, pp. 1160–1179, Nov. 2016.

[9] T. Mikolov, M. Karafiat, L. Burget, F.´ Cernock ˇ y, and S. Khu- ` danpur, "Recurrent neural network based language model," in Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech), 2010.

[10] Toma´s Mikolov, Stefan Kombrink, Luk ˇ a´s Burget, Jan ˇ Cernock ˇ y, and ` Sanjeev Khudanpur, "Extensions of recurrent neural network language model," in 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2011, pp. 5528–5531.

[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.

[12] M. Sundermeyer, R. Schluter, and H. Ney, "Lstm neural networks for language modeling," in Thirteenth annual conference of the international speech communication association, 2012.

[13] R. Kneser and H. Ney, “Improved backing-off for m-gram language modeling,” in ICASSP, 1995, vol. 1, p. 181e4.

[14] S. F. Chen and J. Goodman, “An empirical study of smoothing techniques for language modeling,” Computer Speech & Language, vol. 13, no. 4, pp. 359–394, 1999.

[15] J. T. Goodman, “A bit of progress in language modeling,” Computer Speech & Language, vol. 15, no. 4, pp. 403–434, 2001

[16] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. NAACL, 2018.

[17] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving language understanding by generative pre-training,” Technical Report, OpenAI, 2018.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In NAACL-HLT.

[19] Chao-Wei Huang and Yun-Nung Chen, “Learning ASR-Robust Contextualized Embeddings for Spoken Language Understanding,” In Proceedings of The 45th IEEE ICASSP, 2020.

[20] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 conference on empirical methods in natural language processing, pages 1631–1642, 2013.

[21] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250, 2016.

[22] Ebru Arisoy, Abhinav Sethy, Bhuvana Ramabhadran, and Stanley Chen, “Bidirectional recurrent neural network language models for automatic speech recognition,” in Proc. ICASSP, 2015, pp. 5421–5425.

[23] Xie Chen, Anton Ragni, Xunying Liu, and Mark Gales, “Investigating bidirectional recurrent neural network language models for speech recognition.,” in Proc. ICSA INTERSPEECH, 2017.

[24] Mittul Singh, Youssef Oualil, and Dietrich Klakow, “Approximated and domain-adapted lstm language models for first-pass decoding in speech recognition.,” in Proc. Interspeech, 2017.

[25] Ke Li, Hainan Xu, Yiming Wang, Daniel Povey, and Sanjeev Khudanpur, “Recurrent neural network language model adaptation for conversational speech recognition.,” in Proc. Interspeech, 2018.

[26] Tomas Mikolov and Geoffrey Zweig, “Context dependent recurrent neural network language model,” in Proc. SLT, 2012.

[27] Xie Chen, Tian Tan, Xunying Liu, Pierre Lanchantin, Moquan Wan, Mark JF Gales, and Philip C Woodland, “Recurrent neural network language model adaptation for multi-genre broadcast speech recognition,” in Proc. Interspeech, 2015.

[28] Min Ma, Michael Nirschl, Fadi Biadsy, and Shankar Kumar, “Approaches for neural-network language model adaptation.,” in Proc. Interspeech, 2017.

[29] M. W. Y. Lam, X. Chen, S. Hu, J. Yu, X. Liu, and H. Meng, “Gaussian process lstm recurrent neural network language models for speech recognition,” in ICASSP 2019, pp. 7235–7239, May 2019.

[30] Kazuki Irie, Shankar Kumar, Michael Nirschl, and Hank Liao, “Radmm: recurrent adaptive mixture model with applications to domain robust language modeling,” in Proc. ICASSP, 2018

[31] Ke Li, Zhe Liu, Tianxing He, Hongzhao Huang, Fuchun Peng, Daniel Povey, Sanjeev Khudanpur, “An Empirical Study of Transformer-Based Neural Language Model Adaptation” in Proc. ICASSP, 2020

[32] B. Roark, M. Saraclar, and M. Collins, “Discriminative n-gram language modeling,” Computer Speech and Language, vol. 21, no. 2, pp. 373–392, Apr. 2007.

[33] F.J. Och, “Minimum error rate training in statistical machine translation,” in Proc. ACL, 2003, pp. 160–167.

[34] M. Collins and T. Koo, “Discriminative reranking for natural language parsing,” Computational Linguistics, vol. 31, no. 1, pp. 25–70, Mar. 2005.

[35] T. Oba, T. Hori, A. Nakamura, and A. Ito, “Round-robin duel discriminative language models,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 4, pp. 1244– 1255, May 2012.

[36] Atsunori Ogawa, Marc Delcroix, Shigeki Karita, and Tomohiro Nakatani, “Rescoring n-best speech recognition list based on one-on-one hypothesis comparison using encoder-classifier model,” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 6099–6103.

[37] Tomohiro Tanaka, Ryo Masumura, Takafumi Moriya, and Yushi Aono, “Neural speech-to-text language models for rescoring hypotheses of dnn-hmm hybrid automatic speech recognition systems,” in 2018 AsiaPacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2018, pp. 196–200.

[38] Yuanfeng Song, Di Jiang, Xuefang Zhao, Qian Xu, Raymond Chi-Wing Wong, Lixin Fan, Qiang Yang, "L2RS: A Learning-to-Rescore Mechanism for Automatic Speech Recognition," arXiv preprint arXiv:1910.11496, 2019.

[39] Ankur Gandhe, Ariya Rastrow, "Audio-attention discriminative language model for ASR rescoring," In Proceedings of The 45th IEEE ICASSP, 2020.

[40] Alex Wang and Kyunghyun Cho. 2019. BERT has a mouth, and it must speak: BERT as a Markov random field language model. In NeuralGen.

[41] Joongbo Shin, Yoonhyung Lee, and Kyomin Jung. 2019. Effective sentence scoring method using BERT for speech recognition. In ACML.

[42] Julian Salazar, Davis Liang, Toan Q Nguyen, and Katrin Kirchhoff. 2020. Masked Language Model Scoring. In ACL.

[43] Joongbo Shin, Yoonhyung Lee, Seunghyun Yoon, Kyomin Jung. 2020. Fast and Accurate Deep Bidirectional Language Representations for Unsupervised Learning. In ACL.

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems, pages 6000–6010.

[45] J. Carletta et al., "The AMI meeting corpus: A pre-announcement," The International Workshop on Machine Learning for Multimodal Interaction, 2005.

[46] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely. The Kaldi Speech Recognition Toolkit. In IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society, 2011.

[47] Y. Tachioka and S. Watanabe, "A discriminative method for recurrent neural network language models," in Proc. ICASSP, 2015, pp. 5386–5389.

[48] T. Hori, C. Hori, S. Watanabe, and J.R. Hershey, "Minimum word error training of long short-term memory recurrent neural network language models for speech recognition," in Proc. ICASSP, 2016, pp. 5990–5994.