The 2018 Conference on Computational Linguistics and Speech Processing
ROCLING 2018, pp. 78-80
©The Association for Computational Linguistics and Chinese Language Processing

探討聲學模型的合併技術與半監督鑑別式訓練於會議語音辨識之研究

Investigating acoustic model combination and semi-supervised discriminative training for meeting speech recognition

羅天宏 Tien-Hong Lo, 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
{teinhonglo, berlin}@ntnu.edu.tw

摘要

近年來鑑別式訓練(Discriminative training)的目標函數 Lattice-free Maximum mutual information (LF-MMI)在自動語音辨識(Automatic speech recognition, ASR)上取得了重大的突破[1]，有別於傳統交互熵訓練(Cross-Entropy training, CE)和鑑別式訓練(Discriminative training)的二階段訓練，LF-MMI 提供更快的訓練與解碼。儘管 LF-MMI 在監督式環境下斬獲最好的成果，然而在半監督式環境的表現仍有待研究。在半監督式環境最常見的訓練方法是自我學習(Self-training)[2][3][4]中，由於種子模型(Seed model)常因語料有限而效果不佳。且 LF-MMI 屬於鑑別式訓練之故，更易受到標記錯誤的影響。為了減緩上述的問題，過往常加入置信度過濾器(Confidence-based filter)[4][5][6]對訓練語料做挑選。過濾語料可在不同層級上進行，分為音框層級[7]、詞層級[8]、句子層級[3][8][9]。

本論文利用兩種思路於半監督式訓練。其一，引入負條件熵(Negative conditional entropy, NCE)權重與詞圖(Lattice)，前者是最小化詞圖路徑的條件熵(Conditional entropy)，等同對 MMI 的參考轉錄(Reference transcript)做權重平均，權重的改變能自然地加入 MMI 訓練中，並同時對不確定性建模。其目的希望無置信度過濾器(Confidence-based filter)也可訓練模型。後者加入詞圖，比起過往的 one-best，可保留更多假說空間，提升找到參考轉錄(Reference transcript)的可能性；其二，我們借鑒整體學習(Ensemble learning)

的概念[10]，使用弱學習器(Weak learner)修正彼此的錯誤，分為音框層級合併(Frame-level combination)[11]和假說層級合併(Hypothesis-level combination)[12]。

本論文的實作目的便是在語料缺乏的半監督式環境下，利用負條件熵與詞圖輔助LF-MMI的訓練，並利用模型合併技術，進一步提升模型的辨識結果。我們希望即使在語料不足的情況下，仍能達到不錯的辨識效果，甚至媲美原先有標記語料的訓練結果。實驗結果顯示，加入NCE與詞圖皆能降低詞錯誤率(Word error rate, WER)，而模型合併(Model combination)則能在各個階段顯著提升效能，且兩者結合可使詞修復率(Word recovery rate, WRR)達到60.8%。

關鍵詞：自動語音辨識、鑑別式訓練、半監督式訓練、模型合併

參考文獻

[1] D. Povey et al., “Purely sequence-trained neural networks for ASR Based on Lattice-Free MMI,” in Proc. *INTERSPEECH*, 2016.

[2] K. Vesely et al., “Semi-supervised training of deep neural networks,” in *ASRU*, 2013.

[3] F. Grezl et al., “Semi-supervised bootstrapping approach for neural network feature extractor training,” in *ASRU*, 2013.

[4] P. Zhang et al., “Semi-supervised dnn training in meeting recognition,” in Proceedings of. Sheffield, 2014.

[5] L. Lamel et al., “Lightly supervised and unsupervised acoustic model training,” *Computer Speech & Language* , 2002.

[6] H. Y. Chan et al., “Improving broadcast news transcription by lightly supervised discriminative training,” in *ICASSP*, 2004.

[7] S.-H. Liu et al., “Investigating data selection for minimum phone error training of acoustic models,”in *Multimedia and Expo*, 2007.

[8] K. Vesely et al., “Semisupervised training of Deep Neural Networks,” in ASRU, 2013.

[9] S. Thomas et al., “Deep neural network features and semisupervised training for low resource speech recognition,” in Proc. *ICASSP*, 2013.

[10] P. Zhang et al., “Semisupervised DNN training in meeting recognition,” in *SLT*, 2014.

[11] L. Deng et al., “Ensemble deep learning for speech recognition,” in *INTERSPEECH*, 2014.

[12] H. Xu et al., “Minimum bayes risk decoding and system combination based on a recursion for edit distance,” *Computer Speech and Language*, 2011