OPTIMIZATION IN SPEECH-CENTRIC INFORMATION PROCESSING: CRITERIA AND TECHNIQUES

Xiaodong He and Li Deng

{xiaohe,deng}@microsoft.com

Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA

ABSTRACT

Abstract— Automatic speech recognition (ASR) is an enabling technology for a wide range of information processing applications including speech translation, voice search (i.e., information retrieval with speech input), and conversational understanding. In these speech-centric applications, the output of ASR as "noisy" text is fed into down-stream processing systems to accomplish the designated tasks of translation, information retrieval, or natural language understanding, etc. In conventional applications, the ASR model as a sub-system is usually trained without considering the down-stream systems. This often leads to sub-optimal end-to-end performance. In this paper, we propose a unifying end-to-end optimization framework in which the model parameters in all sub-systems including ASR are learned by Extended Baum-Welch (EBW) algorithms via optimizing the criteria directly tied to the end-to-end performance measure. We demonstrate the effectiveness of the proposed approach on a speech translation task using the spoken language translation benchmark test of IWSLT. Our experimental results show that the proposed method leads to significant improvement of translation quality over the conventional techniques based on separate modular sub-system design. We also analyze the EBW-based optimization algorithms employed in our work and discuss its relationship with other popular optimization techniques.

Index Terms— Speech translation, speech recognition, EBW algorithm, end-to-end optimization criteria

1. INTRODUCTION

Automatic speech recognition (ASR) is central to voice-enabled information processing applications. For example, speech translation (ST) takes the source speech signal as input, then after speech recognition, the output is fed into a machine translation (MT) system and produces as output the translated text of that utterance in another language. That is, the full ST system can be viewed as ASR and MT sub-systems in tandem [3, 11]. Such tandem architecture also characterizes voice search and speech understanding (e.g., [15, 16]).

In all these applications, ASR works together with the down-stream components to deliver the end-to-end result. Different applications emphasize different errors in the ASR output. For example, information retrieval systems focus on the match of content words, while ignoring functional words. Therefore, it is important for the ASR component to have the content words correctly recognized, while the errors in functional words can be tolerated. On the other hand, functional words bear important contextual and syntactic information, which is critical to MT. Therefore, it is crucial to get functional words correctly recognized in MT applications. However, most of the current ASR models are optimized without considering the down-stream sub-systems. Instead, word error rate (WER) is widely accepted as the de facto metric for ASR, treating all types of word errors equally. Since WER only measures word errors at the surface level and takes no consideration of the roles of a word in the ultimate performance measure, this often leads to sub-optimal end-to-end performance [18].

In this paper, we address this critical optimization inconsistency problem, motivating our development of a unifying end-to-end optimization framework for speech-centric information processing applications. In this framework, the ASR models are estimated via optimizing an end-to-end performance metric. Further, we show that this framework can be extended so that the models of down-stream information systems can also be trained together with ASR models to optimize the final performance metric.

The complexity of the end-to-end optimization task requires careful optimization techniques [3,4,19,20]. We have developed over the past few years based on the specialized extended Baum-Welch (EBW) approach. To align with the theme of the special session, we also discussed our optimization techniques in relationship to other popular optimization techniques such as gradient descent and quasi-Newton methods.

2. PREVIOUS WORK

End-to-end optimization of speech-centric information systems involves optimization of complex objective criteria. Efforts have been made and reported in the literature in proposing better optimization criteria and methods. In [5], the margin concept is incorporated into conventional discriminative training criteria such as Minimum Phone Error (MPE) and Maximum Mutual Information (MMI) for string recognition problems. In [6], a fast EBW algorithm built on K-L divergence based regularization is proposed. In [8, 9], a line search A-function (LSAF) is introduced to generalize the EBW algorithm for optimization of discriminant objective functions. In [4], a unified discriminative training criterion is proposed for ASR and is later extended to speech translation based on the Bayesian framework [3] and with the similar growth-transformation (GT) or EBW-based optimization method. This body of earlier work sets up the background for the current work, aimed to exploit more advanced EBW-based optimization technique for improving global, end-to-end optimization for all types of speech-centric information systems with not only faster convergence but better overall performance.

3. EMBEDDING DISCRIMINATIVE TRAINING OF ASR IN END-TO-END SYSTEMS

In this section, we extend the general parameter learning criterion presented in [4] for ASR alone to more general speech-centric applications such as speech translation and voice search. Let us first briefly review the unified discriminative training criterion for ASR proposed in [4], which takes the form:

$$O(\Lambda) = \frac{\sum_{F_1 \dots F_R} p(X_1 \dots X_R, F_1 \dots F_R|\Lambda) \cdot C_{DT}(F_1 \dots F_R)}{\sum_{F_1 \dots F_R} p(X_1 \dots X_R, F_1 \dots F_R|\Lambda)} \quad (1)$$

where R is the number of sentences in the training set. We denote by Λ the parameter set of the ASR models, and denote by X_i and F_i the speech feature sequence and the recognition symbol sequence of i-th utterance, respectively. $C_{DT}(F_1 \dots F_R)$ is a classification quality measure of the concatenated hypotheses $F_1 \dots F_R$, which is independent of Λ. This model training objective function is a model-based expectation of the quality measure of the recognition hypotheses $F_1 \dots F_R$.

In [4], it has been shown that by taking different forms of $C_{DT}(F_1 \dots F_R)$, this general objective function covers a variety of commonly used discriminative training criteria such as Minimum Classification Error (MCE), MMI, and MPE. In Table 1, we further show that this objective function can cover the quality metrics of other speech-centric applications. For example, in voice search, a set of ASR hypotheses, F, are fed into the IR system each as a query. Here, we can measure the quality of F by the metric of retrieval results, e.g., the widely adopted normalized discounted cumulative gain (NDCG) [7]. This can be accomplished by feeding F into the IR system, and then computing the NDCG score on the returned list of ranked documents. In speech translation, the quality of F can be measured by the quality of the translation of F, e.g., the bi-lingual evaluation understudy (BLEU) score [13] for F.

Following similar derivations in [4], the ASR model in various speech-centric applications can be trained to optimize the end-to-end performance in each type of the full systems.

Table 1. By taking different forms of the classification quality measure $C_{DT}(F_1 \dots F_R)$, the unified criterion covers separate discriminative training criteria for distinct speech-centric information systems. F_r^* denotes the reference of the r-th sentence.

Training Criterion	**Performance Metric**	$\boldsymbol{C_{DT}(F_1 \dots F_R)}$
MPE/MWE	WER	$\sum_r \boldsymbol{A(F_r, F_r^*)}$ †
MAX-NDCG	NDCG	$\sum_r \boldsymbol{NDCG(IR(F_r))}$
MAX-BLEU	BLEU	$\sum_r \boldsymbol{BLEU(Tr(F_r), E_r^*)}$

† $A(F_r, F_r^*)$ denotes the raw phone or word accuracy count.

4. JOINT TRAINING ASR & MT MODELS IN SPEECH TRANSLATION

In the previous section, although the ASR models can be trained to optimize the end-to-end performance in speech centric applications, the down-stream components (e.g., MT, IR, NLU, etc.) are fixed, and they act merely as a quality measure component for the ASR hypothesis. It is desirable to optimize these down-stream components in a consistent way and with the same end-to-end performance measure, which we address in this section [20].

Here we propose a unifying end-to-end optimization method for joint training of ASR and down-stream components so as to optimize the final performance measure of a speech-centric information processing system. Particularly, using speech translation as an example, we illustrate how different sub-system components such as ASR and MT can be integrated into a log-linear model and how speech and translation models can be optimized jointly for end-to-end speech translation performance.

4.1. The Unified Log-Linear Model for ST

The optimal translation $\hat{E}$ given the input speech signal X is obtained via the decoding process according to $\hat{E} = \text{argmax}_E P(E|X)$, where

$$P(E|X) = \sum_F P(E, F|X) \quad (2)$$

Then we model the posterior probability of the (E, F) sentence pair given X through a log-linear model:

$$P(E, F|X) = \frac{1}{Z} exp\left\{\sum_i \lambda_i log\varphi_i(E, F, X)\right\} \quad (3)$$

where $Z = \sum_{E,F} exp\{\sum_i \lambda_i log\varphi_i(E, F, X)\}$ is the normalization denominator, and $\{\varphi_i(E, F, X)\}$ are the feature functions empirically constructed from E, F, and X.

The free parameters of the log-linear model, i.e., the weights $\boldsymbol{\lambda} = \{\lambda_i\}$ of these features, are usually trained on a validation set by algorithms such as minimum error rate training (MERT) [12]. However, in the past the features are usually derived from complicated models such as phrase-level and lexicon-level translation models, acoustic models, and language models, etc. Since there are millions of free parameters in these models and the training criterion is a non-linear function of these parameters, the MERT algorithm is no longer suitable. In the next two subsections, we present an end-to-end optimization method to train various types of generative component models from which the features are derived.

4.2. Training Criteria

Similar to (1), we denote by $X = X_1 \dots X_R$ the super-string formed by concatenating all R training utterances. Likewise, $E = E_1 \dots E_R$ denotes the super-string formed by concatenating all R translation hypotheses. And Λ denotes the full parameter set in both the ASR and MT models. Then we define the objective function in a succinct form of

$$O(\Lambda) = \frac{\sum_E p(X, E|\Lambda) \cdot C_{DT}(E)}{\sum_E p(X, E|\Lambda)} \quad (4)$$

where $C_{DT}(E) = \sum_r BLEU(E_r, E_r^*)$. This gives model-based expectation of the average of sentence level BLEU scores, scaled by a factor of $1/R$.

4.3. EBW Algorithm for Joint Training

After substitute (2) and (3) into (4), we obtain

$$O(\Lambda) = \frac{\sum_E \sum_F \prod_i \varphi_i^{\lambda_i}(E, F, X|\Lambda)\, C_{DT}(E)}{\sum_E \sum_F \prod_i \varphi_i^{\lambda_i}(E, F, X|\Lambda)} \quad (5)$$

where $\varphi_i(E, F, X|\Lambda) = \prod_{r=1}^{R} \varphi_i^{\lambda_i}(E_r, F_r, X_r|\Lambda)$ represent all elemental features used. We call this product form as *feature decomposable* at the sentence level. Similarly, we have $C_{DT}(E) =$

$\sum_r C_{DT}(E_r)$, where $C_{DT}(E_r) = BLEU(E_r, E_r^*)$ is the BLEU score of the r-th sentence, and we call this summation form as *measure decomposable* at the sentence level. Hereafter, we will omit the subscript of $C_{DT}(E)$ for simplification.

Using the super-string annotation, we can construct the primary auxiliary function [4]:

$$F(\Lambda;\Lambda') = \sum_E \sum_F \prod_i \varphi_i^{\lambda_i}(X,E,F|\Lambda)\,[C(E) - O(\Lambda')] \tag{6}$$

where Λ denotes the model to be estimated, and Λ' the model obtained from the immediately previous iteration.

Then, similar to [2, 4], EBW algorithm can be derived for estimating Λ based on the extended Baum-Eagon method [1]. Limited by the space, here we will only provide the parameter estimation formula for the lexicon-level translation model and mean vectors of the Gaussian-HMM based acoustic model to elaborate on the EBW-based discriminative approach to joint training of ST components.

First, we use the backward lexical weighting feature as a concrete example to illustrate the EBW approach. The lexical weighting feature function is based on the lexicon-level translation model:

$$P(F|E,\Lambda) = \prod_k \prod_m \sum_n p(f_{k,m}|e_{k,n},\Lambda) \tag{7}$$

Therefore, according to [1, 2], we have EBW formula for the lexicon-level translation model parameter $p(g|h,\Lambda)$ as follows:

$$p(g|h,\Lambda) = \frac{p(g|h,\Lambda')\left(\frac{\partial F(\Lambda;\Lambda')}{\partial p(g|h,\Lambda)}|_{\Lambda=\Lambda'} + D_h\right)}{\sum_g p(g|h,\Lambda')\left(\frac{\partial F(\Lambda;\Lambda')}{\partial p(g|h,\Lambda)}|_{\Lambda=\Lambda'} + D_h\right)} \tag{8}$$

This can be further simplified to

$$p(g|h,\Lambda) = \frac{\sum_{k,m:\ f_{k,m}=g} \sum_{E,F} p(E,F|X,\Lambda')\Delta_E\,\gamma_h(k,m) + D_h \cdot p(g|h,\Lambda)'}{\sum_{k,m}\sum_{E,F} p(E,F|X,\Lambda')\Delta_E\,\gamma_h(k,m) + D_h} \tag{9}$$

where $\Delta_E = [C(E) - O(\Lambda')]$, and

$$\gamma_h(k,m) = \frac{\sum_{n:e_{k,n}=h} p(f_{k,m}|e_{k,n},\Lambda')}{\sum_n p(f_{k,m}|e_{k,n},\Lambda')} \tag{10}$$

In our implementation, D_h is set by

$$D_{\tilde{e}} = \tau + \rho \cdot \sum_k \sum_{\substack{E,F:\\ e_k=\tilde{e}}} p(F,E|X,\Lambda')\max(-\Delta_E, 0) \tag{11}$$

where we set τ to a small positive value and $\rho \geq 1$, so that the denominator of (9) is guaranteed to be positive.

Next, let's use the mean vector of the Gaussian model as another example. We have the estimation formula

$$\mu_j = \frac{\sum_t \Delta\gamma(j,t)\cdot x_t + D_j\mu_j{}'}{\sum_t \Delta\gamma(j,t) + D_j} \tag{12}$$

where

$$\Delta\gamma(j,t) = \sum_{E,F} p(E,F|X,\Lambda')[C(E) - O(\Lambda')]\gamma_{j,F}(t) \tag{13}$$

and $\gamma_{j,F}(t)$ is the occupation probability of state j of recognition hypothesis F at time t. D_j is set in a way similar to (11).

Let's now compare the model re-estimation formula in ST with its counterpart in ASR. The main difference is at the computation of $\Delta\gamma(j,t)$. In ST, as shown in (13), the weighting factor of the raw occupation probability $\gamma_{j,F}(t)$ is$[C(E) - O(\Lambda')]$, which is based on the translation quality measure $C(E)$ for the translation hypothesis E. In ASR, $\Delta\gamma(j,t) = \sum_F p(F|X,\Lambda')[C(F) - O(\Lambda')]\gamma_{j,F}(t)$ [4], where the weighting factor of $\gamma_{j,F}(t)$, i.e., $[C(F) - O(\Lambda')]$, is based on the recognition quality measure $C(F)$ for the recognition hypothesis F. Also, in ASR, $\Delta\gamma(i,t)$ is the expectation of the weighted $\gamma_{i,F}(t)$ over all recognition hypotheses $\{F\}$, while in ST, $\Delta\gamma(i,t)$ is the expectation of the weighted $\gamma_{i,F}(t)$ over all recognition hypotheses $\{F\}$ and also over all the translation hypotheses $\{E\}$.

In order to study the relationship between the EBW-based updating formula of (12) and the gradient-based optimization, we first compute the first-order gradient of the objective function (5) w.r.t. the parameters to be estimated, using μ_j as an example:

$$\nabla_{\mu_j} O(\Lambda)|_{\Lambda=\Lambda'} = \Sigma_j'^{-1} \sum_t \Delta\gamma(j,t)(x_t - \mu_j') \tag{14}$$

Further, assuming that different dimensions of the mean vector are independent of each other, we can approximate the Hessian matrix by

$$H_j = \nabla_{\mu_j}^2 O(\Lambda)|_{\Lambda=\Lambda'} \approx -\Sigma_j'^{-1} \sum_t \Delta\gamma(j,t) \tag{15}$$

Then, (12) can be rewritten into

$$\mu_j \approx \mu_j' - \frac{\sum_t \Delta\gamma(j,t)}{\sum_t \Delta\gamma(j,t) + D_j} H_j^{-1} \nabla_{\mu_j} O(\Lambda)|_{\Lambda=\Lambda'} \tag{16}$$

This approximates the 2nd-order update and usually gives faster learning speed than the simple gradient-based search. On the other hand, the relationship between this method and other optimization methods in [8, 9, 17] is yet to be explored.

5. EXPERIMENTAL EVALUATION

In this section, we conduct evaluation on the standard IWSLT Chinese-to-English DIALOG task benchmark test [14]. The baseline is a phrase-based translation system as described in [10]. Each sentence in the development and test sets has seven reference translations, on which the BLEU score is measured with.

In our approach, we tune the feature weights and the translation models alternatively. At each iteration, the feature weights are optimized on the development set by MERT according to [12], then we decode the full training corpus using the current feature models and weights. After that, sufficient statistics are collected. Finally, the translation models are updated. These steps go through several iterations until convergence is reached.

We then perform the full iteration of training, i.e., both the feature weights and the translation models are updated at each iteration, by MERT and GT, respectively. Fig 1 shows the BLEU

score on the development set after different number of iterations. It is shown that after three iterations, the BLEU score is improved from 46.88% (the baseline) to 48.33%. Then, we apply this setting to the test set. The results on the test set are tabulated in Table 2. It shows that the end-to-end discriminative training of translation models significantly improves the BLEU score on the test set from 44.20% to 45.42%, a gain of absolute 1.22% BLEU points.

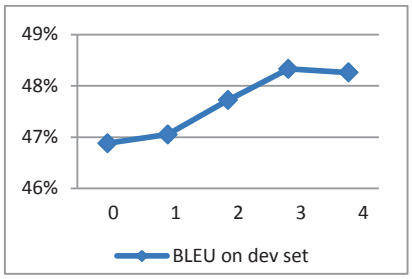

Fig. 1. BLEU scores on the development set over EBW iterations.

Table 2. Translation results in BLEU scores on the test set

Method	Test
Baseline (ML)	44.20%
E2E learning via EBW	45.42%

6. CONCLUSION

A unifying framework is presented for optimally constructing speech-centric information processing systems. The hallmark of such systems is the serial combination of ASR as the "front end" sub-system and the down-stream sub-systems such as MT, IR, and NLU. The key innovation in our unifying framework is the end-to-end optimization objectives and techniques, where the model parameters in all sub-systems are jointly learned via optimizing the objective function directly tied to the final performance measure defined by the full systems.

In addition to the theoretical and algorithmic contributions, in this paper we also demonstrate the effectiveness of the proposed approach on a speech translation task using the spoken language translation benchmark test of IWSLT. Speech translation is a serial combination of ASR and MT. Traditionally, these two components are trained independently. In this paper, we present an end-to-end learning approach that jointly trains these two components following the general principle underlying the unifying framework.

A specific contribution of this work is the pervasive use of discrimination in the full MT and ST system. In previous work of MT and ST, discriminative learning was only applied to weighting parameters as pioneered in [12]. The framework presented in this paper provides an approach where discriminative learning is injected into the feature functions themselves.

In the past, EBW method has been used mainly in ASR, and has accounted for the huge success in discriminative training of HMM-based speech recognizers. This paper represents the first work where EBW-based optimization is applied successfully in ST and MT. EBW method serves as a unifying optimization technique in learning complex systems where sub-components of the full system are serially connected and where the objective function of the system parameter learning can be expressed as a rational function. We are hopeful that in addition to speech and language processing problems, system parameter learning of other information processing problems can also benefit from the EBW-based approach presented in this paper.

REFERENCES

[1] L. Baum and J. Eagon, "An inequality with applications to statistical prediction for functions of Markov processes and to a model of ecology," *Bull. Amer. Math. Soc.*, Jan. 1967.

[2] P. Gopalakrishnan, D. Kanevsky, A. Nadas, and D. Nahamoo, "An inequality for rational functions with applications to some statistical estimation problems," *IEEE Trans. Inform. Theory*, Jan. 1991.

[3] X. He and L. Deng, "Speech recognition, machine translation, and speech translation -- A unified discriminative learning paradigm," *IEEE Sig. Proc. Mag., Sept*. 2011.

[4] X. He, L. Deng, W. Chou, "Discriminative learning in sequential pattern recognition." *IEEE Sig. Proc. Mag.,* 2008.

[5] G. Heigold, P. Dreuw, S. Hahn, R. Schlu¨ter and H. Ney, "Margin-based discriminative training for string recognition," *IEEE Journal of STSP*, Dec. 2010.

[6] R. Hsiao and T. Schultz, "Generalized Baum-Welch algorithm and its implication to a new extended Baum-Welch algorithm", In *Proc. Interspeech* 2011.

[7] K. Jarvelin, and J. Kekalainen, "IR evaluation methods for retrieving highly relevant documents." In *Proc. SIGIR*, 2000.

[8] D. Kanevsky, D. Nahamoo, T. Sainath, B. Ramabhadran, P. Olsen,"A-Functions: a generalization of extended Baum-Welch transformations to convex optimization", in *Proc. ICASSP*, 2011.

[9] D. Kanevsky, D. Nahamoo, T. Sainath, and B. Ramabhadran, "Convergence of line search A-Function methods", In *Proc. Interspeech*, 2011.

[10] P. Koehn, F. Och, and D. Marcu. "Statistical phrase-based translation," In *Proc. HLT-NAACL*, 2003.

[11] E. Matusov, S. Kanthak, and H. Ney, "Integrating speech recognition and machine translation: Where do we stand?" In *Proc. ICASSP*, 2006.

[12] F. Och, "Minimum error rate training in statistical machine translation." In *Proc. ACL*, 2003.

[13] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: a method for automatic evaluation of machine translation." In *Proc. ACL*, 2002.

[14] M. Paul, M. Federico, and S. Stücker, "Overview of the IWSLT 2010 evaluation campaign." In *Proc. IWSLT*, 2010.

[15] Y. Wang, D. Yu, Y. Ju, A. Acero, "An introduction to voice search", *IEEE Sig. Proc. Mag*., May 2008.

[16] S. Yaman, L. Deng, D. Yu, Y. Wang, and A. Acero, "An integrative and discriminative technique for spoken utterance classification," *IEEE Trans. ASLP*, 2008.

[17] J Nocedal, "Updating Quasi-Newton Matrices With Limited Storage," *Mathematics of Computation*, July 1980.

[18] X. He, L. Deng, and A. Acero, "Why word error rate is not a good metric for speech recognizer training for the speech translation task?" in *Proc. ICASSP*, 2011.

[19] M. Lehr and I. Shafran, "Discriminatively estimated joint acoustic, duration and language model for speech recognition," In *Proc. ICASSP*, 2010.

[20] Y. Zhang, L. Deng, X. He, and A. Acero, "A novel decision function and the associated decision-feedback learning for speech translation," in *Proc. ICASSP*, 2011.