Discriminative Learning of Feature Functions of Generative Type in Speech Translation

Xiaodong He XIAOHE@MICROSOFT.COM
Microsoft Research, One Microsoft Way, Redmond, WA 98052 USA

Li Deng DENG@MICROSOFT.COM
Microsoft Research, One Microsoft Way, Redmond, WA 98052 USA

Abstract

The speech translation (ST) problem can be formulated as a log-linear model with multiple features that capture different levels of dependency between the input voice observation and the output translations. However, while the log-linear model itself is of discriminative nature, many of the feature functions are derived from generative models, which are usually estimated by conventional maximum likelihood estimation. In this paper, we first present the formulation of the ST problem as a log-linear model with a plurality of feature functions. We then describe a general discriminative learning framework for training these generative features based on a technique called growth transformation (GT). The proposed approach is evaluated on a spoken language translation benchmark test of IWSLT. Our experimental results show that the proposed method leads to significant improvement of translation quality. Fast and stable convergence can also be achieved by the proposed method.

1. Electronic Submission

Speech translation (ST) takes the source speech signal as input and produces as output the translated text of that utterance in another language. It can be viewed as automatic speech recognition (ASR) and machine translation (MT) in tandem.

Like many other machine learning problems, the speech translation (ST) problem can be modeled by a log-linear model with multiple features that capture different dependencies between the input voice observation and the output translations. Although the log-linear model itself is a discriminative model, many of the feature functions, such as scores of ASR outputs, are still derived from generative models. Further, these features are usually trained by conventional maximum likelihood estimation. In this paper, we propose a general framework of discriminative training for these generative features based on a technique called growth transformation (GT). The proposed approach is evaluated on a spoken language translation benchmark test called IWSLT. Our experimental results show that the proposed method leads to significant translation performance improvement. It is also shown that fast and stable convergence can be achieved by the proposed GT based optimization method.

2. Previous Work

In [He et al 2006, HeDeng2008], we presented the GT-based discriminative training method of hidden Markov models (HMM) for ASR in a systematic way. More recently, in [HeDeng2011], this optimization method was extended to ST based on the Bayesian framework. In [HeDengAcero2011], we provided experimental evidence that global end-to-end optimization in ST is superior to separate training of ASR and MT components of a ST system. And in [Zhang et.al. 2011], a global end-to-end optimization for ST was implemented using a gradient descent technique with slow convergence. All these earlier work set up the background for the current work, aimed to use more advanced optimization technique of GT for improving the global end-to-end optimization of ST with not only faster convergence but also better ST accuracy.

3. Speech Translation: Modeling and Training

A general framework for ST is illustrated in Fig. 1. The input speech signal X is first fed into the ASR module. Then the ASR module generates the recognition output set $\{F\}$, which is in the source language. The recognition hypothesis set $\{F\}$ is finally passed to the MT module to obtain the translation sentence E in the target language. In our setup, an N-best list is used as the interface between ASR and MT. In the following, we use F to represent an ASR hypothesis in the N-best list. Detailed descriptions of the processes of ASR, MT and ST have been provided in [HeDeng2011].

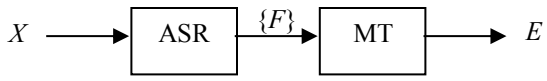

Fig. 1. Two components of a speech translation system

3.1. 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

$$\hat{E} = \underset{E}{\operatorname{argmax}}\, P(E|X) \quad (1)$$

Based on law of total probability, we have,

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

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

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

where $Z = \sum_{E,F} exp\{\sum_i \lambda_i log \varphi_i(E,F,X)\}$ is the normalization denominator to ensure that the probabilities sum to one. In the log-linear model, $\{\varphi_i(E,F,X)\}$ are the feature functions empirically constructed from E, F, and X. The only free parameters of the log-linear model are the feature weights, i.e. $\Lambda = \{\lambda_i\}$. Details of these features used in our experiments are provided next.

3.2. Features in the ST model

The full set of feature functions constructed and used in our ST system are derived from both the ASR and the MT modules as listed below:

- Acoustic model (AM) feature: $\varphi_{AM}(E,F,X) = p(X|F)$, which is the likelihood of speech signal X given a recognition hypothesis F, computed from the AM of the source language. This is usually modeled by a hidden Markov model (HMM).
- Source language model (LM) feature: $\varphi_{SLM}(E,F,X) = P_{LM}(F)$, which is the probability of F computed from a N-gram LM of the source language. This is usually modeled by a N-1 order Markov model.
- Forward phrase translation feature: $\varphi_{F2Eph}(E,F,X) = P_{TMph}(E|F) = \prod_k p(\tilde{e}_k|\tilde{f}_k)$, where $\tilde{e}_k$ and $\tilde{f}_k$ are the k-th phrase in E and F, respectively, and $p(\tilde{e}_k|\tilde{f}_k)$ is the probability of translating $\tilde{f}_k$ to $\tilde{e}_k$. This is usually modeled by a multinomial model.
- Forward word translation feature: $\varphi_{F2Ewd}(E,F,X) = P_{TMwd}(E|F) = \prod_k \prod_m \sum_n p(e_{k,m}|f_{k,n})$, where $e_{k,m}$ is the m-th word of the k-th target phrase $\tilde{e}_k$, $f_{k,n}$ is the n-th word in the k-th source phrase $\tilde{f}_k$, and $p(e_{k,m}|f_{k,n})$ is the probability of translating word $f_{k,n}$ to word $e_{k,m}$. (This is also referred to as the lexical weighting feature.) Note, although this feature is derived from the probability distribution $\{ p(e_{k,m}|f_{k,n}) \}$ which is modeled by a multinomial model.
- Backward phrase translation feature: $\varphi_{E2Fph}(E,F,X) = P_{TMph}(F|E) = \prod_k p(\tilde{f}_k|\tilde{e}_k)$, where $\tilde{e}_k$ and $\tilde{f}_k$ are defined as above.
- Backward word translation feature: $\varphi_{E2Fwd}(E,F,X) = P_{TMwd}(F|E) = \prod_k \prod_n \sum_m p(f_{k,n}|e_{k,m})$, where $e_{k,m}$ and $f_{k,n}$ are defined as above.
- Translation reordering feature: $\varphi_{order}(E,F,X) = P_{hr}(S|E,F)$ is the probability of particular phrase segmentation and reordering S, given the source and target sentence E and F. In a phrase-based translation system, this is usually described by a heuristic function.
- Target language model (LM) feature: $\varphi_{TLM}(E,F,X) = P_{LM}(E)$, which is the probability of E computed from an N-gram LM of the target language, modeled by a N-1 order Markov model.
- Count of NULL translations: $\varphi_{NC}(E,F,X) = e^{|Null(F)|}$ is the exponential of the number of the source words that are not translated (i.e., translated to NULL word in the target side).
- Count of phrases: $\varphi_{PC}(E,F,X) = e^{|\{(\tilde{e}_k,\tilde{f}_k),k=1,\ldots,K\}|}$ is the exponential of the number of phrase pairs.
- Translation length: $\varphi_{TWC}(E,F,X) = e^{|E|}$ is the exponential of the word count in translation E.
- ASR hypothesis length: $\varphi_{SWC}(E,F,X) = e^{|F|}$ is the exponential of the word count in the source sentence F. (This is also referred to as word insertion penalty.)

3.3. Conventional Training Method

The free parameters of the log-linear model, i.e., the weights (denoted by Λ) of these features, are usually trained by minimum error rate training (MERT) [Och 2003]. Specifically, the training is aimed to maximize the BLEU score of the final translation on a validation set according to

$$\hat{\Lambda} = \underset{\Lambda}{\operatorname{argmax}}\, BLEU(E^*, \hat{E}(\Lambda,X)) \quad (4)$$

where E^* is the translation reference(s), and $\hat{E}(\Lambda,X)$ is the translation output. The latter is obtained through the decoding process according to (1) given input speech X and feature weights Λ. The operation of (4) is often carried out using grid search, which is feasible due to a small number of the weights, e.g., 12.

However, the number of free parameters of feature functions is huge, and it is not suitable to train them using

the above grid search method. In most MT and ST systems today, the free parameters of feature functions are usually estimated separately, with maximum likelihood estimation.

In the next sections, we will reformulate the training objective as an expected accuracy of the translation, and derived growth transformation (GT) in optimizing these models.

4. New Discriminative Training Method

We first introduce the discriminative training objective function for ST. Then, we derive the GT of the models.

4.1. The discriminative training objective function

As proposed in [HeDengChou2008] and [HeDeng2011], we denote by $X = X_1 \dots X_R$ the superstring of concatenating all R training utterances, and $E = E_1 \dots E_R$ the superstring of concatenating all R training references, then we can define the objective function:

$$O(\Lambda) = \sum_E p(E|X,\Lambda) \cdot C_{DT}(E) \quad (5)$$

This is the model-based expectation of the classification quality measure $C_{DT}(E)$ for ST, where $C_{DT}(E)$ is the evaluation metric or its approximation. For translation, the quality is usually evaluated by Bi-Lingual Evaluation Understudy (BLEU) scores or Translation Edit Rate (TER). A few example of $C_{DT}(E)$ for ST can be found in [HeDeng2011]. In this work, we adopt:

$$C_{DT}(E) = \sum_r BLEU(E_r, E_r^*) \quad (6)$$

which is proportional (by $1/R$) to the average of sentence level BLEU scores.

4.2. Growth transformation for model training

After some algebra, we have:

$$p(E|X,\Lambda) = \frac{\sum_F \prod_i \varphi_i^{\lambda_i}(E,F,X|\Lambda)}{\sum_E \sum_F \prod_i \varphi_i^{\lambda_i}(E,F,X|\Lambda)} \quad (7)$$

And

$$O(\Lambda) = \frac{\sum_E \sum_F \prod_i \varphi_i^{\lambda_i}(E,F,X|\Lambda) C_{DT}(E)}{\sum_E \sum_F \prod_i \varphi_i^{\lambda_i}(E,F,X|\Lambda)} \quad (8)$$

where $\varphi_i(E,F,X|\Lambda) = \prod_{r=1}^{R} \varphi_i^{\alpha_i}(E_r,F_r,X_r|\Lambda)$ represent all features described in Section 3.2. We call this *feature decomposable* at the sentence level. Similarly, we have,

$$C_{DT}(E) = \sum_r C_{DT}(E_r) \quad (9)$$

where $C_{DT}(E_r) = BLEU(E_r, E_r^*)$ is the BLEU score of the r-th sentence, and we call this *measure decomposable* at 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:

$$F(\Lambda;\Lambda') = \sum_E \sum_F \prod_i \varphi_i^{\lambda_i}(X,E,F|\Lambda)\,[C(E) - O(\Lambda')] \quad (10)$$

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

Then, similar to [Gopalakrishnan et.al. 1991], GT can be derived for estimating Λ based on the extended Baum-Eagon method [BaumEagon1967]. In the following, we will give derivation of two translation feature functions in the ST system to elaborate on the GT-based discriminative training approach for ST.

4.2.1 GT for the phrase translation model

We use the backward phrase translation model, which was described in Section 3.2, as an example to illustrate the GT approach. Given,

$$P(F|E) = \prod_k p(\tilde{f}_k|\tilde{e}_k) \quad (11)$$

we have GT as:

$$p(\tilde{f}|\tilde{e},\Lambda) = \frac{p(\tilde{f}|\tilde{e},\Lambda') \frac{\partial F(\Lambda;\Lambda')}{\partial p(\tilde{f}|\tilde{e},\Lambda)}|_{\Lambda=\Lambda'} + D_{\tilde{e}} \cdot p(\tilde{f}|\tilde{e},\Lambda')}{\sum_{\tilde{f}} p(\tilde{f}|\tilde{e},\Lambda') \frac{\partial F(\Lambda;\Lambda')}{\partial p(\tilde{f}|\tilde{e},\Lambda)}|_{\Lambda=\Lambda'} + D_{\tilde{e}}} \quad (12)$$

Denote by $\Delta_E = [C(E) - O(\Lambda')]$, we have:

$$p(\tilde{f}|\tilde{e},\Lambda) = \frac{\sum_k \sum_{\substack{E,F: \\ e_k=\tilde{e} \\ f_k=\tilde{f}}} p(F,E|X,\Lambda')\Delta_E + D_{\tilde{e}} \cdot p(\tilde{f}|\tilde{e},\Lambda')}{\sum_k \sum_{\substack{E,F: \\ e_k=\tilde{e}}} p(F,E|X,\Lambda')\Delta_E + D_{\tilde{e}}} \quad (13)$$

where $D_{\tilde{e}}$ is a constant independent from Λ. It could be proved that there exists a large enough $D_{\tilde{e}}$ such that the

above transformation can guarantee a *growth* of the value of objective function defined in (5). In practice, this bound is usually too large and leads to very slow convergence, and people have developed some approximation to speed up the convergence. Refer to [HeDengChou 2008] for more discussions.

The forward phrase translation model has a similar GT estimation formula.

4.2.2 GT for the word translation model

We now use the backward lexical weighting feature as another example to illustrate GT. Given

$$P(F|E,\Lambda) = \prod_k \prod_m \sum_n p(f_{k,m}|e_{k,n},\Lambda) \quad (14)$$

we have GT formula for the word translation model $p(g|h,\Lambda)$ as:

$$p(g|h,\Lambda) = \frac{p(g|h,\Lambda')\left(\frac{\partial F(\Lambda;\Lambda')}{\partial p(g|h,\Lambda)}|_{\Lambda=\Lambda'}+D_h\right)}{\sum_g p(g|h,\Lambda')\left(\frac{\partial F(\Lambda;\Lambda')}{\partial p(g|h,\Lambda)}|_{\Lambda=\Lambda'}+D_h\right)} \quad (15)$$

This can be simplified to

$$p(g|h,\Lambda) = \frac{\sum_{k,m:\ f_{k,m}=g} \sum_{E,F} p(E,F|X,\Lambda')\Delta_E\gamma_h(k,m)+D_h \cdot p(g|h,\Lambda)'}{\sum_{k,m}\sum_{E,F} p(E,F|X,\Lambda')\Delta_E\gamma_h(k,m)+D_h} \quad (16)$$

where

$$\gamma_h(k,m) = \frac{\sum_{n:e_{k,n}=h} p(f_{k,m}|e_{k,n},\Lambda')}{\sum_n p(f_{k,m}|e_{k,n},\Lambda')}. \quad (17)$$

The forward word translation model has a similar GT formula.

5. Evaluation

5.1. The discriminative training objective function

In this section, we conduct evaluation on the international workshop on spoken language translation (IWSLT) Chinese-to-English DIALOG task benchmark test. The test includes conversational speech in a travel scenario. The translation training data consisted of approximately 30,000 parallel sentences in both Chinese and English. The test set is the 2008 IWSLT spontaneous speech Challenge test set, consisting of 504 Chinese sentences. In this task, the speech recognition transcriptions are given, so our focus is on the training of translation related feature models, specifically, the forward and backward phrase translation model and word translation model discussed in Section 4.

The baseline is a phrase-based translation system including all the translation features defined in Section 3.2. The parameter set of the log-linear model is optimized by MERT. The translation features such as phrase and word translation models are trained by maximum likelihood. In training, the parallel training data are first word-aligned. Then, phrase tables are extracted from the aligned parallel corpus. The target language model is trained on the English side of the training data.

In our GT approach, the log-linear model is fixed. We first decode the whole training corpus using the current feature models. Then, sufficient statistics are collected. Finally, the model parameters are updated according to (13) and (16). These steps go with several iterations until convergence is reached.

5.2. Experimental results

In evaluation, single-reference based BLEU scores are reported. Fig. 2 shows the convergence of the proposed GT-based discriminative training of all four translation models.

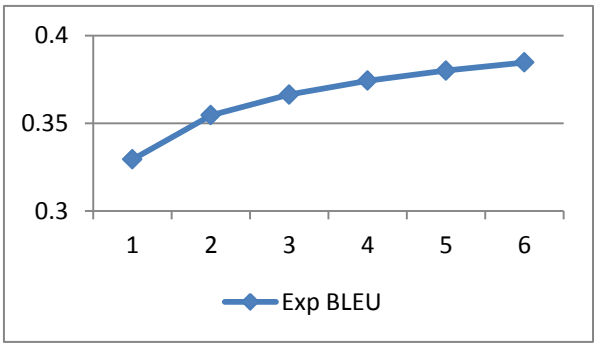

Fig. 2. The expected BLEU score on the training set along with the number of iterations.

It is shown that the GT-based training gives fast and stable convergence, where the value of the objective function, which is the expected sentence-level BLEU score (Expected BLEU), grows monotonically after each iteration, and start to converge after 5 iterations.

Fig.3 shows the relationship between the Expected BLEU and the BLEU score of the top-1 translation hypothesis on the training corpus. It is shown that these two scores correlated very well, indicating that improving the expected BLEU helps improve the BLEU score of the top-1 translation. Fig 4 shows the BLEU score on the test set after different number of iterations. It is shown that after 5 iterations, the BLEU score is improved from the 0.202 (the baseline) to 0.218, a substantial improvement of (absolute) 1.6%.

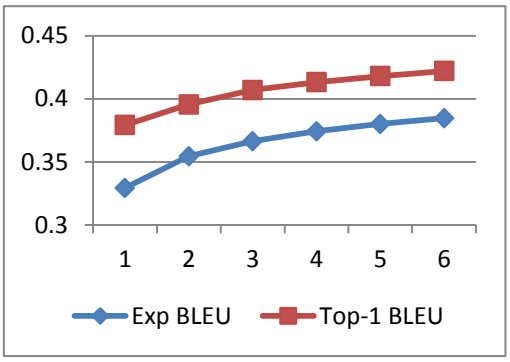

Fig. 3. The expected BLEU vs. the top-1 BLEU scores on the training set, along with the number of iterations.

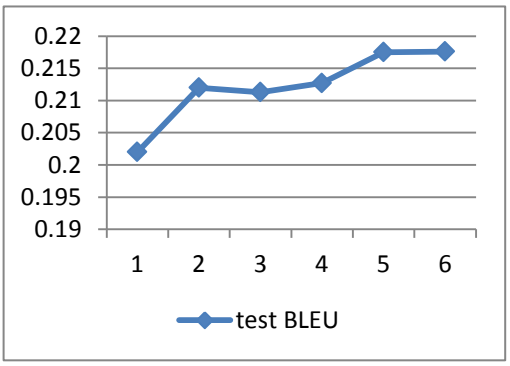

Fig. 4. BLEU scores on the test set over iterations.

6. Conclusion

Speech translation is a serial combination of speech recognition and machine translation. Traditionally, these two components are trained independently. In this paper, we propose an end-to-end learning approach that jointly trains these two components. A new optimization technique based on GT, also called extended Baum-Welch algorithm, is introduced to accomplish this task. This is superior to our earlier approach based on gradient decent.

One major 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 applied to weighting parameters as pioneered in [Och 2003]. The framework presented in this paper provides an approach where discriminative learning is injected into the feature functions themselves.

In the past, GT has been used mainly in speech recognition, and has accounted for the huge success in discriminative training of HMM-based speech recognizers. This is the first time that GT optimization is applied successfully in ST and MT. GT serves as a unifying framework 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.

On the other hand, ASR and MT are the two most important components in speech recognition. Therefore, another important research direction is integration of the end-to-end optimization method with latest advances in these two areas, such as speaker adaptation in ASR [HeZhao2003] [Lei2006] and system combination in MT [HeToutanova2009] [Li et.al. 2009], to achieve even better speech translation performance.

References

Baum L. and Eagon, J. “An inequality with applications to statistical prediction for functions of Markov processes and to a model of ecology,” *Bull. Amer. Math. Soc.*, vol. 73, pp. 360–363, Jan. 1967.

Gopalakrishnan, P., D. Kanevsky, A. Nadas, and D. Nahamoo, “An inequality for rational functions with applications to some statistical estimation problems,” *IEEE Trans. Inform. Theory.*, vol. 37, pp. 107–113, Jan. 1991.

He, X., Deng, L., Chou, W. “A novel learning method for hidden Markov models in speech and audio processing,” in *IEEE MMSP*, October 2006

He, X., Deng, L., Chou, W. "Discriminative learning in sequential pattern recognition," in *IEEE Sig. Proc. Mag.*, vol. 25, 2008, pp. 14-36.

He X. and Deng L. “Speech recognition, machine translation, and speech translation” in *IEEE Sig. Proc. Mag.*, 2011, to appear.

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

He X. and Toutanova K., “Joint optimization for machine translation system combination,” *In Proc. EMNLP*, 2009

He X. and Zhao Y., “Fast model selection based speaker adaptation for nonnative speech,” in *IEEE Transaction on Speech and Audio Processing*, IEEE, 2003

Lei X., Hamaker J., and He X., “Robust feature space adaptation for telephony speech recognition,” in *InterSpeech* , 2006

Li C-H., He X., Liu Y., and Xi N., “Incremental HMM alignment for MT system combination,” in *ACL*, 2009

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

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