NEW METHODS AND EVALUATION EXPERIMENTS ON TRANSLATING TED TALKS IN THE IWSLT BENCHMARK

Amittai Axelrod, Xiaodong He, Li Deng, Alex Acero, Mei-Yuh Hwang

amittai@uw.edu, {xiaohe,deng,alexac,mehwang}@microsoft.com

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

ABSTRACT

Abstract— The IWSLT benchmark task is an annual evaluation campaign on spoken language translation held by the International Workshop on Spoken Language Processing (IWSLT). The task is to translate TED talks (www.ted.com). This task presents two unique challenges: Firstly, the underlying topic switches sharply from talk to talk, and each one contains only tens to hundreds of utterances. The translation system therefore needs to adapt to the current topic quickly and dynamically. Secondly, unlike other machine translation benchmark tasks, only a very small relevant parallel corpus (transcripts of TED talks) is available. Therefore, it is necessary to perform accurate translation model estimation with limited data. In this paper, we present our recent progress and two new methods on the IWSLT TED talk translation task from Chinese into English. In particular, to address the first problem, we use unsupervised topic modeling to select additional topic-dependent parallel data from a globally irrelevant corpus. These additional data slices can then be used to build an unsupervised topic-adapted machine translation system. For the second problem, we develop a discriminative training method to estimate the translation models more accurately. Our experimental evaluation results show that both methods improve the translation quality over a state-of-the-art baseline.

Index Terms— spoken language translation, topic adaptation, discriminative training, IWSLT.

1. INTRODUCTION

The IWSLT benchmark is an annual evaluation of spoken language translation (SLT) held by the International Workshop on Spoken Language Processing (IWSLT) [5]. Since 2010, the main focus of IWSLT has been the translation of TED talks (www.ted.com). TED talks are given by leaders in various fields and cover an open set of topics in Technology, Entertainment, Design, and other domains. Compared to conventional machine translation tasks, this task presents two unique challenges: First, the underlying topic switches sharply from talk to talk, and each talk contains only tens to hundreds of utterances. Therefore, the system needs to adapt to the current topic dynamically and automatically. Second, unlike text based machine translation where a large parallel training corpus is often available, there is only a small amount of talk-style parallel data consisting of human translations of TED talks. Therefore, methods of estimating accurate translation models from limited parallel data are needed.

In this paper, we present our recent progress in preparing for IWSLT 2011 Evaluation. To address the first problem, we propose a topic model-based method for fast unsupervised topic adaptation. Machine translation systems are more effective when used to translate input that closely matches the training and tuning data. Here the wide-ranging subject of the talks contraindicates the use of a single domain-specific system for the task. A topic model [10] is a generative model for explaining broad topical variety in a corpus. The importance of this model is that it is unsupervised, and that after training it can be used to perform statistical inference on new input. This allows previously-unheard utterances to be related to the topics learned during training. Topic models have been used to select additional monolingual data to create a topic-specific language model [11], and these models have been applied to the task of statistical machine translation (SMT) [12][13]. Combining topic models with prior work on selecting relevant out-of-domain sub-corpora [14][15], we propose a method for selecting additional parallel corpora using an unsupervised topic model.

To address the second problem, we develop a discriminative training method to estimate the translation channel models more accurately. The machine translation problem is commonly modeled by a log-linear model with multiple features that capture different dependencies between the source language and the target language [2]. Although the log-linear model is discriminative in nature, many of the feature functions, such as the phrase-level translation probability features and the lexicon-level translation probability features (e.g., lexical weighting), are derived from generative models. Further, these features are usually trained by conventional maximum likelihood (ML) estimation [6], which may not correspond to the translation quality well. This objective mismatch could particularly problematic when data are sparse. In order to address this problem, we propose a discriminative training method for these generative translation models based on a technique called growth transformation (GT) [1].

We conducted experiments on the IWSLT Chinese-to-English machine translation task. Our experimental results show that both of the proposed methods lead to significant translation quality improvement.

2. THE TED-TALK TRANSLATION TASK

The goal of the 2011 IWSLT evaluation campaign is the translation of TED talks. In this work, we conducted our experiments on the IWSLT Chinese-to-English machine translation task. TED talks are originally given in English, In the Chinese to English translation task, we are given human translated Chinese text with partial punctuations inserted. The goal is to match the human transcribed English speech with detailed punctuations. This is an open-domain spoken language translation task, whose training data consist of approximately 110K sentences in the

transcripts of the TED talks, and their translations, in English and Chinese, respectively. In addition, the IWSLT evaluation campaign also provides out-of-domain data for potential usage. These include about 7.7M parallel sentences of UN proceedings, and 115M of monolingual English sentences, from multiple sources including the EuroMatrixPlus project, Europarl corpus, and LDC Gigawords corpus.

We evaluated our work on the 2010 IWSLT MT_CE task test set, consisting of 1664 Chinese sentence with one English reference translation each. This is the most recent IWSLT test set for which the reference translations are available (the 2011 evaluation set had not published at the time of this writing). To evaluate, we used a phrase-based decoder as proposed in [6] to produce all the system outputs, and measured the case-sensitive BLEU scores with the NIST *mt-eval31a* tool used in the IWSLT evaluation [3]. In the following sections, we will present our two new methods in greater details. We will also evaluate them and present the experimental results as appropriate.

3. TOPIC ADAPTATION

3.1 Probabilistic Topic Model

Latent Dirichlet Allocation (LDA) [10] is a probabilistic topic model for decomposing the content of a (heterogeneous) corpus according to some number of topics *K*. In particular, for a fixed number of topics, each part of the corpus is assumed to reflect some combination of all of those topics. Probabilistic inference can then be used to extract an underlying topical structure from the corpus. One advantage to topic models is that they can be trained in an unsupervised manner, using freely-available toolkits such as MALLET [9].

Let $P(z)$ be the distribution, over all Z topics, in a particular utterance W which consists of words w. In LDA, $P(z)$ is taken to have a Dirichlet distribution. Now let $P(w|z)$ is the probability distribution of words given the particular topic z. The generative story of probabilistic topic models supposes that each word w in an utterance is produced by first sampling a topic z from $P(z)$, and then selecting a word w according to $P(w|z)$. The probability of a word within an utterance is thus:

$$P(w) = \sum_{k=1}^{K} P(w|z = k)P(z = k) \quad (1)$$

3.2 Unsupervised Topic Adaptation

Once the topic model has been trained, it can be used to infer the topic mixture of new utterances. These topic scores can be used to cluster the new input relative to the existing *K* topics. Prior work has shown that the data in each topical cluster in a corpus can be used to train targeted language models which outperform the general corpus-wide model on topic-specific input [11]. This approach has applied to Statistical Machine Translation (SMT) as well, whereby language models are adapted to the parallel corpus topics and added to the system to improve translation performance [12] [13].

In this work we instead consider the case where there is both an in-domain and an out-of-domain bilingual parallel corpus. Rather than adapting a topical language model to use in combination with a background model, we wish to identify parts of the external parallel corpus that are similar to the individual topics in the in-domain corpus. The 2011 IWSLT task included the use of 7.7 million sentences of parallel UN data, which can be considered out-of-domain relative to the TED talks in the training corpus. Our experiments show that the UN corpus, when used in its entirety as a second translation model, does not positively impact translation. However, prior work by [14] shows that relevant subsets of an unrelated corpus can be more beneficial for training a second translation model than using the entire additional corpus.This motivates the use of a topic model trained on the input (Chinese) side of the TED talks to select the most relevant subset of the UN corpus for each particular topic, based on threshholding the scores of the single-most-likely topic. In this way, the UN parallel corpus is trimmed to four pieces totaling the 1.4M most topically-relevant sentences. Each of these topic-specific subsets is used to train a topic-specific translation model. The TED training corpus for IWSLT is not large enough to split into topics that are big enough to use to train a reasonable translation model, so all the TED data is used together as a general TED-domain model and adaptation is performed by using a different subset of the UN data to train the topic-adapted model.

The tuning and evaluation data was split into topics via the same model that had been trained on the TED data, and assigning it to the single most likely topic. Even concatenating the 2010 dev and test sets, we were limited to 4 topics to keep each topical tuning set be large enough to prevent overfitting. Each topical subset of the input data was decoded using the corresponding topical model. During MERT learning and run-time testing, two translation models, one general and one topic-specific, were used in combination with two language models trained on the in-domain data and some additional monolingual data (Section 4.3). These four models were tuned for each topic in a log-linear combination.

Table 1. BLEU scores for the baseline and topic-adapted translation system on the combined development and test set.

Method	BLEU
Baseline	11.34%
+ Topic-Adapted UN data	11.75%

Table 1 shows the result of tuning and evaluating the 2598 sentences corresponding to the dev2010 and tst2010 datasets. These systems were also evaluated with the IWSLT 11 benchmark data (result pending). On the combined dataset, using the topic-adapted translation model improved performance by 0.41 BLEU (3.6% relative) compared with the baseline.

4. DICRIMINATIVE TRANINING OF TRANSLATION MODELS

4.1 The Log-Linear Model and Its Features for Translation

In SMT, the optimal translation $\hat{E}$ given the input source sentence F is obtained via the decoding process according to

$$\hat{E} = \underset{E}{\operatorname{argmax}}\, P(E|F) \quad (2)$$

According to [2], we then model the posterior probability of the translation hypothesis E given the source sentence F via a Max-Entropy model that takes the log-linear form:

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

where $Z = \sum_E exp\{\sum_i \lambda_i log\varphi_i(E,F)\}$ being the normalization denominator to ensure that the probabilities sum to one. In the log-linear model, $\{\varphi_i(E,F)\}$ are the feature functions constructed from E and F. In a phrase-based machine translation system, features include hypothesis length, number of phrases, reordering model scores, language model scores, and four translation model-based features as follows:

- 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.
- The backward phrase translation feature is defined similarly.
- 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.
- The backward word translation feature is defined similarly.

In [4], Och proposes tuning the linear weights of these features, e.g., $\boldsymbol{\lambda} = \{\lambda_i\}$, directly via maximizing the BLEU score of the final translation on a development set according to

$$\hat{\boldsymbol{\lambda}} = \underset{\boldsymbol{\lambda}}{\operatorname{argmax}}\, BLEU\big(\hat{E}(\boldsymbol{\lambda}, X), E^*\big) \tag{4}$$

where E^* is the translation reference(s), and $\hat{E}(\boldsymbol{\lambda}, X)$ is the translation output. This leads to efficient training algorithm such as Minimum Error Rate Training (MERT) [4]. In (4), the normalization denominator Z is ignored since it is irrelevant to the decoding process, and the model becomes a linear model. Although the linear model above is discriminative in nature, many of the feature functions, such as the translation models based features, are derived from generative models. Conventionally, these features are usually trained by maximum likelihood (ML) estimation [6]. Recently, effort has been made to further extend the max-BLEU training method. In [16], model parameters are optimized with a perceptron using the best possible translation hypothesis as the approximated reference. On the other hand, in [17], the linear model is extended to include tens of thousands of fine-grained features, where most of them are binary indicators. In order to effectively training the weights of this many features, an MIRA-based optimization method is used.

Extended from our earlier work [18], we propose a discriminative training method for the estimation of translation models based on a technique called growth transformation (GT) [1][7]. Unlike [16], we use the expected BLEU score as the objective function and the true reference is used without approximation. Compared to [17], our focus is on discriminative training of the phrase and lexicon translation probability distributions. With our method, we can train tens of millions of parameters effectively.

4.2 Estimation formula for translation models

Let Λ denote the full parameter set of the translation models. The objective function of our method is expected BLEU:

$$O(\Lambda) = \sum_E p(E|F,\Lambda) C_{DT}(E) \tag{5}$$

where $C_{DT}(E)$ is the evaluation metric, which for translation is BLUE score. In this work, we adopt:

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

Detailed derivation is omitted due to space limitation and will be elaborated in a future paper. In the following, we just present the estimation formula for the phrase and lexicon translation models directly. Using the backward phrase translation model as an example, the GT formula is:

$$p(\tilde{f}|\tilde{e},\Lambda) = \frac{\sum_k \sum_{\substack{E,F:\\ e_k=\tilde{e}\\ f_k=\tilde{f}}} p(F|E,\Lambda')\Delta_E + D_{\tilde{e}} \cdot p(\tilde{f}|\tilde{e},\Lambda')}{\sum_k \sum_{\substack{E:\\ e_k=\tilde{e}}} p(F|E,\Lambda')\Delta_E + D_{\tilde{e}}} \tag{7}$$

where $\Delta_E = [C_{DT}(E) - O(\Lambda')]$ and $D_{\tilde{e}}$ is a constant independent of Λ. Λ' denotes the model obtained from the last iteration. In our implementation, the following formula is used to compute $D_{\tilde{e}}$:

$$D_{\tilde{e}} = \tau + \rho \cdot \sum_k \sum_{\substack{E,F:\\ e_k=\tilde{e}}} p(F|E,\Lambda')\max(-\Delta_E, 0) \tag{8}$$

We set τ to be a small positive value and $\rho \geq 1$, so that the denominator of (7) is guaranteed to be positive. The forward phrase translation model has a similar GT estimation formula and will be omitted here. For the backward lexical weighting feature, the GT formula for the lexicon translation model $p(g|h,\Lambda)$ is:

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

where

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

and D_h is set in a similar way as (8). Again, the forward word translation model has a similar GT estimation formula.

4.3 Experimental results

The baseline is a phrase-based translation system as described in [6], including all the translation features defined in Section 4.1. The translation features such as phrase and word translation models are trained by maximum likelihood. In training, the TED parallel training data are first word-aligned by a lexicalized HMM [8]. Then, a phrase table is extracted from the aligned TED parallel corpus [6]. Similarly we build a second phrase table from 500k UN parallel sentences selected from the UN corpora by the method described in [14]. Two language models are used: one is a 3-gram LM trained on the English side of the TED parallel corpora, and the other one is a 5-gram LM trained on the supplementary monolingual English corpus. The development set for the max-BLEU based feature weight tuning consists of 934 sentences.

In our approach, we discriminatively trained the four translation models for the TED data derived phrase table. The UN data derived phrase table has small impact on this task and is kept with no change. In training, we tune the feature weights of the log-linear model and train the translation models for the phrase table alternatively. At each iteration, the feature weights are optimized on the development set by MERT, and then we fix the features weights and collect sufficient statistics and finally, the translation models are updated according to (7) and (9). These steps iterate until convergence is reached.

Fig.1 shows the convergence of the proposed GT-based discriminative training of translation models. In order to make the expected BLEU comparable across different iterations, $\boldsymbol{\lambda}$ is fixed in Fig 1. The GT-based training produces fast and stable convergence: the value of the objective function, which is the expected sentence-level BLEU score (Expected BLEU), grows smoothly after each iteration.

The results on the dev set and the test set are presented in Table 2. Discriminative training of translation models significantly improves the BLEU points on the test set from 11.48 to 12.21, a gain of absolute 0.73 BLEU points (6.4% relative).

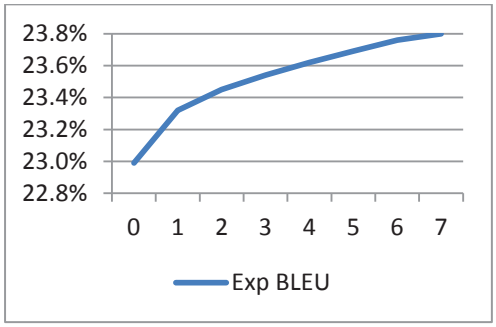

Fig. 1. The expected BLEU score on the training set along with the number of iterations.

Table 2. Translation BLEU scores on the development set and test set, respectively.

Method	Dev set	Test set
Baseline	10.27%	11.48%
DT of trans. mdl.	11.06%	12.21%

5. CONCLUSION

Speech translation technology has many challenges, and compared speech recognition, it is much less mature. In the research described in this paper, we take a task-oriented approach to improving speech translation technology in the context of IWSLT benchmark evaluation. We first identify two rather unique difficulties --- rapid topic switching and shortage of in-domain parallel data --- associated with the IWSLT benchmark task in translating TED talks from Chinese to English.

Targeting these challenges, we have developed two new methods --- dynamic topic adaptation with no supervision and discriminative learning for the translation model. We have rigorously evaluated these two methods. The topic adaptation technique improves a strong baseline by absolute 0.41 BLEU points (3.6% relative). The discriminative learning technique has more significant improvement with 0.73 absolute BLEU point increase (6.4% relative).

Our future research will extend the current success in two fronts. First, the current TED talk translation task has limited the power of topic modeling and adaptation. The shortage of supplied data permits the division of topics with only one level of "resolution". With many real-world tasks where data are less restricted, multi-resolution topic modeling and adaption can be derived from the current work to achieve greater effectiveness. Second, the discriminative learning method described in this paper can also be extended to improve not only the translation model but also simultaneously the ASR model in the full speech translation system in an integrative manner.

REFERENCES

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

[2] F. Och and H. Ney. "Discriminative training and maximum entropy models for statistical machine translation." In Proc. ACL 2002.

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

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

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

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

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

[8] X. He, "Using word-dependent transition models in HMM-based word alignment for statistical machine translation," *Proc. ACL-WMT*, 2007.

[9] A. McCallum, "MALLET: A machine learning for language toolkit", http://mallet.cs.umass.edu, 2002.

[10] D. Blei, A. Ng, M.I. Jordan, "Latent Dirichlet allocation," Journal of Machine Learning Research, 2003.

[11] Y.C. Tam and T. Schultz, "Unsupervised language model adaptation using latent semantic marginals", In *Proc of Interspeech,* 2006.

[12] Y.C. Tam, I. Lane, T. Schultz, "Bilingual-LSA based LM adaptation for spoken language translation". In *Proc of ACL*, 2007.

[13] N. Ruiz, M. Federico. "Topic adaptation for lecture translation through bilingual latent semantic models". In *Proc of WMT* 2011.

[14] A. Axelrod, X. He, J. Gao. "Domain adaptation via pseudo in-domain data selection". Proceedings of Empirical Methods in Natural Language Processing, 2011.

[15] X. He and L. Deng, "Robust speech translation by domain adaptation." *Proc. Interspeech*, 2011

[16] P. Liang, A. Bouchard-Cote, D. Klein and B. Taskar, "An end-to-end discriminative approach to machine translation," in *Proc. COLING-ACL*, 2006

[17] D. Chiang, K. Knight and W. Wang, "11,001 new features for statistical machine translation," in *Proc. NAACL-HLT*, 2009.

[18] 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.