UNSUPERVISED TRAINING FOR FARSI-ENGLISH SPEECH-TO-SPEECH TRANSLATION

Bing Xiang, Yonggang Deng, Yuqing Gao

IBM T. J. Watson Research Center, Yorktown Heights, NY 10598
{bxiang, ydeng, yuqing}@us.ibm.com

ABSTRACT

Speech-to-speech translation has evolved into an attractive area in recent years with significant progress made by various research groups. However, the translation engines usually suffer from the lack of bilingual training data, especially for low-resource languages. In this paper we present an unsupervised training technique to alleviate this problem by taking advantage of available source language data. Different approaches are proposed and compared through extensive experiments conducted on a speech-to-speech translation task between Farsi and English. The translation performance is significantly improved in both directions with the enhanced translation model. A state-of-the-art Farsi automatic speech recognition system is also established in this work.

Index Terms— Speech Recognition, Machine Translation, Unsupervised Training

1. INTRODUCTION

In recent years, automatic speech-to-speech (S2S) translation has drawn much attention in research community due to its wide range of practical applications. It aims at breaking down the communication barriers between people who do not speak a common language. Significant progress has been made by various research groups. However, people are still facing many challenges. One of them is the lack of training data in both automatic speech recognition (ASR) and machine translation (MT), especially for low-resource languages. For MT the creation of human translated bilingual corpus is expensive and time-consuming. On the other hand, monolingual data is usually cheap and easier to collect. In this work we present a few unsupervised training techniques to enhance the translation model by better utilizing the source languages, in contrast to the usual approaches that concentrate on target languages to improve language models.

There have been some similar work proposed previously in both speech recognition and machine translation, which fall into unsupervised or semi-supervised training areas. In speech recognition, it has been shown in [1] [2][3] that unsupervised training is an effective approach to improve the acoustic models. A baseline system is trained with limited transcribed speech data first. Then the system is used to decode some untranscribed data that is available. The decoded hypotheses are selected based on various criteria and added into the training corpus, on which a better acoustic model is trained iteratively. Significant improvements have been achieved in these work.

In MT, several semi-supervised training techniques have been proposed to improve the word alignments [4][5]. For example, in [5] a semi-supervised training approach is proposed to alternate the unsupervised training step with a discriminative step trained on a small, manually word-aligned corpus, which leads to improved alignments. More recently, in [6][7] a technique was proposed to utilize source languages to improve the translation performance, especially for different domains. An initial system is used to translate the source text in a test set. Then translation output is filtered based on confidences and used to create an additional phrase table, which serves as a new component in a log-linear model. It is shown that the translation performance can be improved significantly on test sets that are in a domain different from that of the original training corpus. However, the experiments conducted in these work are limited by the size of test sets and hence the number of new phrases obtained is small. Also, it is not feasible for an online real-time MT system.

In this work, we concentrate on a different scenario, where we have relatively more in-domain source language data available, e.g. from audio transcripts or another translation task that shares the same source language but a different target language. Different approaches are proposed and compared through extensive experiments on Farsi-to-English and English-to-Farsi translations.

The rest of the paper is organized as follows: Section 2 presents different approaches on data selection and retraining with available source language data. Section 3 briefly describes the development of a state-of-the-art Farsi speech recognition system given that it is rarely addressed in speech recognition community. Section 4 reports extensive experimental results obtained in both Farsi-to-English and English-to-Farsi translations. The paper ends with some conclusions and future work discussion in Section 5.

2. UNSUPERVISED TRAINING FOR MT

In this section we first describe our baseline MT system, then present three different algorithms within the framework of unsupervised training for MT.

2.1. Baseline System

Our baseline is a phrase-based statistical machine translation system. We start from a sentence-aligned parallel training corpus and generate word alignments with GIZA++ [8] based on IBM Model 1-4 and hidden Markov model. Then we extract phrase pairs based on the word alignments and some symmetrization heuristics [8]. A phrase table is built upon them with the probabilities estimated based on relative frequency. Our decoder is a phrase-based multi-stack implementation of log-linear model similar to Pharaoh [9]. As most of the statistical MT systems, the features used in the decoder include phrase translation probabilities and lexical probabilities in both directions, language model, distortion penalty, phrase penalty and word penalty. The decoding weights are optimized to maximize BLEU scores [10], where BLEU is a metric to evaluate translation qualities based on n-gram precisions and a brevity penalty. A 3-gram language model is trained with Kneser-Ney smoothing [11] and used in the decoder.

2.2. Algorithm I

Suppose we have some source language data available besides a human-annotated parallel training corpus. The question is whether we can utilize such untranslated source language data to improve the translation. A straightforward way is to translate these source language data with current baseline system. Then select the translation hypotheses and add them to the original parallel corpus along with their corresponding source text. The procedure can be divided into the following steps.

1. Translate the untranslated source language data with baseline system. For each source sentence, generate N-best hypotheses, each composed of multiple target phrases.
2. Compute confidence scores and rerank the N-best list.
3. Select top-n hypotheses from the reranked N-best list and add into the original parallel corpus.
4. Rerun GIZA++ to generate word alignments.
5. Extract phrase pairs based on new alignments and build a new phrase table for the system.
6. The procedure can be iterative until no more effect observed on the development set.

The confidence score for each N-best is calculated with a log-linear model as in Eq. (1).

$$c_n = w_1 logP_{s,n} + w_2 logP_{ph,n} + w_3 logP_{st,n} + w_4 logP_{ts,n} + w_5 logP_{lm,n} + w_6 T_n \quad (1)$$

It is a combination of six features with the weights optimized on a development set under the maximum BLEU criterion. $P_{s,n}$ is the sentence posterior probability of the n-th best hypothesis as computed in Eq. (2).

$$P_{s,n} = \frac{exp(s_n)}{\sum_{\hat{n}=1}^{N} exp(s_{\hat{n}})}, \quad (2)$$

where s_n is the nbest score provided by the decoder for the n-th hypothesis. As mentioned above in the section of baseline system, s_n is the score from a log-linear combination of several different features.

$P_{ph,n}$ is the phrase posterior probability for the n-th hypothesis. It is a product of the phrase posterior probabilities of all target phrases in that hypothesis.

$$P_{ph,n} = \prod_{i=1}^{M_n} p_{n,i} \quad (3)$$

The phrase posterior probability for each target phrase is defined in Eq. (4).

$$p_{n,i} = \frac{\sum_{\hat{n}} \sum_{\hat{i}, e_{\hat{n},\hat{i}} = e_{n,i}} P_{s,\hat{n}}}{\sum_{\hat{n}} \sum_{\hat{i}} P_{s,\hat{n}}}, \quad (4)$$

where $e_{n,i}$ is the i-th target phrase in the n-th nbest hypothesis.

As defined in Eq. (5) and (6), $P_{st,n}$ and $P_{ts,n}$ are the lexical probabilities in two directions, where $p(e_t|f_s)$ is the probabilty of the t-th target word in the target hypothesis given the s-th source word in the source sentence.

$$P_{st,n} = \prod_{t=1}^{T_n} \prod_{s=1}^{S_n} p(e_t|f_s) \quad (5)$$

$$P_{ts,n} = \prod_{t=1}^{T_n} \prod_{s=1}^{S_n} p(f_s|e_t) \quad (6)$$

The last two features in Eq. (1) are language probabilty $P_{lm,n}$ and the number of target words T_n. In this work, we use 4-gram language model in N-best reranking. The weights for confidence score are optimized on the development set.

2.3. Algorithm II

We also explore a different approach of utilizing the untranslated source data. Instead of retraining the system from scratch after merging the selected data with the original corpus, we build a phrase table with the selected data only, then interpolate it with the original phrase table based on a mixture model. The data selection part is the same as in Algorithm I, i.e. picking reranked n-best hypotheses from the N-best list. Again, the selection and training procedure can be iterative.

The mixture models for the phrase translation probabilities in both directions are shown in Eq. (7) and (8), where e and f are target and source phrases. P_{org} and P_{add} are the probabilities estimated from the original parallel corpus and the additional selected data, respectively.

$$P(e|f) = \alpha logP_{org}(e|f) + (1-\alpha)P_{add}(e|f) \quad (7)$$

$$P(f|e) = \alpha logP_{org}(f|e) + (1-\alpha)P_{add}(f|e) \quad (8)$$

The interpolation weight α can be tuned on the development set as other decoding weights.

2.4. Algorithm III

When the amount of available source language data is relatively large, we may also incrementally select them to improve the translation model. The iterative procedure is illustrated below.

1. In iteration i, translate subset i of the untranslated source language data with current MT system and generate N-best hypotheses.
2. Calculate confidence scores for each hypothesis and rerank the N-best list.
3. Select top-n hypotheses from the reranked N-best list and combine with the corresponding source text.
4. Generate word alignments on the selected data with GIZA++.
5. Extract phrase table based on new alignments and then interpolate with the table from previous iteration, T_{i-1}, to get new phrase table T_i.
6. Go back to the first step unless no more effect observed on the development set.

In this algorithm, we partition the untranslated source language data into multiple subsets randomly. In each iteration, we translate one of the subsets with currently the best system, then improve the translation model with the interpolated phrase table. The benefit is that in iteration i ($i > 1$), we have a system better than the baseline to translate the new data, and hence higher translation quality is expected in the final translation system.

3. FARSI-ENGLISH SPEECH-TO-SPEECH TRANSLATION

A Farsi-English speech-to-speech translation system includes the following components.

1. Farsi and English speech recognition
2. Two-way machine translation between Farsi and English
3. Speech synthesis for Farsi and English

In this section we briefly describe the development of Farsi speech recognition. For more details about other components, please refer to [12].

3.1. Farsi Speech Recognition

The Farsi acoustic training data consists of around 80 hours of speech collected under the DARPA Transtac program, which covers mainly the military domain. All the audio data are re-sampled at 16kHz before the feature extraction. Every 10 ms a 24-dimensional MFCC feature vector is computed and then mean normalized. Sequences of 9 vectors are then stacked together leading to a 216-dimensional new feature vector. This new feature space is finally reduced to 40 dimensions with a combination of linear discriminant analysis (LDA) and maximum likelihood linear transformation (MLLT).

Two types of context-dependent quinphone models are built in this work. One is a grapheme model, with each Farsi letter modeled by one grapheme. The other is a phonetic model, where the phonemes for each Farsi word are extracted from a pronunciation lexicon released in Transtac. The short vowels in Farsi words are explicitly modeled in the phonetic system. There are 43 graphemes in the grapheme model and 31 phonemes in the phonetic model. The number of Gaussians is around 60K. Both models are trained under either maximum likelihood (ML) or minimum phone error (MPE) [13] criterion. An fMPE [14] feature transform is also applied on the final model. A statistical 3-gram language model is trained with around 100K Farsi sentences. The number of unique words is around 30K.

Table 1 shows the roadmap of the development of our Farsi ASR system. The word error rate (WER) is measured on a 2-hour Farsi test set randomly selected from the Transtac data and excluded from the training.

System	Phone	Train	Iter	WER
S1	grapheme	ML	1	40.9
S2	phonetic	ML	1	36.6
S3	phonetic	ML	2	33.5
S4	phonetic	ML	3	31.4
S5	phonetic	MPE	4	29.1
S6	phonetic	fMPE	4	26.0

Table 1. ASR results on the Farsi development set

As shown in the table, the vowelization in the phonetic model provided 4.3% absolute WER reduction. In the first iteration of ML training, the phone alignments were generated by a context-independent model. While in the second and third iterations, we used the previous iteration model to align the audio data with the transcripts. About 5% absolute WER reduction was achieved in the end during the ML training. Another 2-3% absolute improvement was achieved by using MPE and fMPE separately. The IBM Farsi speech recognition system achieved the lowest word error rate among all the systems participating in the Transtac July 2007 Evaluation.

4. EXPERIMENTAL RESULTS

In this section, we report the results we obtained on a set of experiments conducted in Farsi to English and English to Farsi translations. In both directions, the translation model in the baseline system is trained with around 110K parallel sentences. The target 3-gram language models are trained with the target side data from the parallel corpus. All the experiments use a 1430-sentence set as the development set to tune the weights. The systems are evaluated on the Transtac July 2007 Evaluation Farsi-English offline test data with close to 1K sentences in each direction. There are two types of source input. One is source language speech references in the text-to-text (T2T) task. The other is the IBM Farsi or English speech recognition system output from the speech-to-text (S2T) task. All test sets have four sets of human-annotated MT references.

4.1. Farsi to English

All the unsupervised training experiments on Farsi to English translation utilized 50K-sentence Farsi audio transcripts that have no English translations.

The results from Algorithm I with or without N-best reranking are compared with that from the baseline in Table 2. The second row corresponds to the experiment in which we applied Algorithm I without reranking. That means we simply picked the raw 1-best output generated by the baseline and added them to the parallel corpus. There is little effect on the two test sets. However, when a reranking is applied based on confidence scores, we see that the system achieved slightly better results, with 0.3 to 0.5 gain in BLEU on the test sets, as shown in the third row of Table 2. This shows that the confidence score based reranking helped to find more reliable translations. Another observation we can have from the 4th column of Table 2 is that the number of phrases increased by 40% since we re-generated the word alignments and hence obtained many new phrase pairs appearing in the extra source language text and their corresponding translations.

Algorithm	Rerank	Sent	Phrase	T2T	S2T
Baseline	N/A	110K	1.7M	31.5	25.6
I	No	160K	2.5M	31.6	25.6
I	Yes	160K	2.4M	32.0	25.9
II	Yes	160K	2.5M	32.6	26.6
II	Yes	610K	3.6M	32.5	26.0

Table 2. Comparison of Algorithm I and II with baseline

In Table 2, we also showed the results from Algorithm II, where we selected reranked 1-best (4th row) or 10-best (last row) and built a new phrase table before the interpolation with the baseline phrase table. This method achieved better performance than simply merging data with the original corpus and retraining from scratch. When selecting 1-best only, the BLEU scores are increased by 1% absolute compared to the baseline results. But adding 10-best instead results in worse score on the S2T set, although still slightly better than the baseline. This is mainly due to more noise from the lower-rank N-best, where we have one source sentence translated into multiple hypotheses with different qualities.

The iterative training results from Algorithm II are shown in Table 3. In each iteration, we translate the 50K Farsi sentences with current system, then build a phrase table and interpolate it with that from the previous iteration. We see an increase in the phrase table size after each iteration. The translation performance improved in

the first two iterations, but no more in the third iteration, which is a sign of saturation for the good quality phrase pairs. Overall, the extra Farsi sentences contributed more than 1.5% gain in BLEU on the T2T task and more than 1% gain on the S2T task.

Algorithm	Iter	Sent	Phrase	T2T	S2T
Baseline	0	110K	1.7M	31.5	25.6
II	1	160K	2.5M	32.6	26.6
II	2	210K	2.8M	33.1	26.8
II	3	260K	3.0M	33.3	26.3

Table 3. Iterative results with Algorithm II

4.2. English to Farsi

We also conducted a set of experiments in English to Farsi translation. In this direction we have relatively more source language data. There are 800K English sentences extracted from bilingual corpora in other language pairs, which are in the similar domain as our Farsi-English test sets.

In Table 4, we first show the results of adding 200K randomly selected and then translated English sentences. Even without reranking, the BLEU scores are increased by 1.7% on T2T and 0.9% on S2T. We believe that this is largely due to the significant increase of the phrase table size. When selecting the 1-best based on confidence scores, further improvements are achieved as shown in the third row.

Algorithm	Rerank	Iter	Sent	Phrase	T2T	S2T
Baseline	N/A	0	110K	1.6M	24.2	20.8
I	No	1	310K	5.0M	25.9	21.7
I	Yes	1	310K	4.9M	26.5	22.1
II	Yes	1	310K	4.8M	26.6	22.3
II	Yes	1	910K	12.0M	26.7	22.3
III	Yes	2	910K	12.4M	27.0	22.5

Table 4. Comparison of Algorithm I, II and III with baseline

In Table 4, we also experimented with the interpolation of phrase tables. Algorithm II provided similar gain in the 4th row as that from Algorithm I. In the 5th row, we translated all 800K sentences with the baseline but no big effect on the test sets. In the last row, we implemented Algorithm III. Its first iteration is the same as the 4th row, i.e. translating 200K sentences with the baseline, while in the second iteration we use the first-iteration model to translate the rest 600K English sentences. The final results are around 2% better than the IBM baseline that already achieved the highest BLEU scores in the Transtac July 2007 Evaluation.

Here we also show the increase of n-gram precisions in Table 5 for the S2T test set. We believe that they are mainly benefited from two facts. The first is that we have significantly more phrases in the interpolated phrase table, which provides wider coverage. The second reason is that we selected reliable translation hypotheses, which helped to reinforce some of the correct translations.

Algorithm	1-gram	2-gram	3-gram	4-gram	BLEU
Baseline	62.1	28.8	14.7	7.1	20.8
III	64.7	31.6	16.3	8.2	22.5

Table 5. N-gram precision and BLEU scores

5. CONCLUSIONS AND FUTURE WORK

This paper tried to alleviate the problem of lacking bilingual training data in statistical machine translation, especially for low-resource languages. Different algorithms and approaches have been proposed and compared through a set of experiments conducted in the task of Farsi-English speech-to-speech translation. It is shown that the translation performance can be largely improved through these techniques in both translation directions. Future work will include other intelligent data selection and filtering methods. Whether similar techniques can be applied to update other components in the MT engine, such as target language model, is also worth to investigate.

6. ACKNOWLEDGEMENT

This work was partially supported by the DARPA Transtac program.

7. REFERENCES

[1] T. Kemp and A. Waibel, "Unsupervised Training of a Speech Recognizer: Recent Experiments," in *Proc. Eurospeech*, Sept. 1999, pp. 2725–2728.

[2] L. Lamel, J. L. Gauvain, and G. Adda, "Unsupervised Acoustic Model Training," in *Proc. ICASSP*, 2002, pp. 877–880.

[3] J. Ma, S. Matsoukas, O. Kimball, and R. Schwartz, "Unsupervised Training on Large Amounts of Broadcast News Data," in *Proc. ICASSP*, 2006, pp. 1056–1059.

[4] C. Callison-Burch, D. Talbot, and M. Osborne, "Statistical Machine Translation with Word- and Sentence-Aligned Parallel Corpora," in *Proc. ACL*, 2004.

[5] A. Fraser and D. Marcu, "Semi-Supervised Training for Statistical Word Alignment," in *Proc. ACL*, 2006, pp. 769–776.

[6] N. Ueffing, "Using Monolingual Source-Language Data to Improve MT Performance," in *Proc. IWSLT*, 2006, pp. 174–181.

[7] N. Ueffing, G. Haffari, and A. Sarkar, "Transductive Learning for Statistical Machine Translation," in *Proc. ACL*, Jun 2007, pp. 25–32.

[8] F. J. Och and H. Ney, "A Systematic Comparison of Various Statistical Alignment Models," in *Computational Linguistics*, vol. 29, 2003, pp. 9–51.

[9] P. Koehn, F. J. Och, and D. Marcu, "Pharaoh: A Beam Search Decoder for Phrase Based Statistical Machine Translation Models," in *Proc. AMTA*, 2004.

[10] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation," in *Proc. ACL*, July 2002, pp. 311–318.

[11] R. Kneser and H. Ney, "Improved Backing-off for M-gram Language Modeling," in *Proc. ICASSP*, 1995, pp. 181–184.

[12] Y. Gao, B. Zhou, L. Gu, R. Sarikaya, H.-K. Kuo, A. I. Rosti, M. Afify, and W. Zhu, "IBM MASTOR: Multilingual Automatic Speech-to-Speech Translator," in *Proc. ICASSP*, 2006, pp. 1205–1208.

[13] D. Povey and P. C. Woodland, "Minimum Phone Error and I-Smoothing for Improved Discriminative Training," in *Proc. ICASSP*, 2002, pp. 105–108.

[14] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "fMPE: Discriminatively Trained Features for Speech Recognition," in *Proc. ICASSP*, 2005, pp. 961–964.