Robust Speech Translation by Domain Adaptation

Xiaodong He and Li Deng

Microsoft Research, Redmond, WA, USA

{xiaohe|deng}@microsoft.com

Abstract

Speech translation tasks usually are different from text-based machine translation tasks, and the training data for speech translation tasks are usually very limited. Therefore, domain adaptation is crucial to achieve robust performance across different conditions in speech translation. In this paper, we study the problem of adapting a general-domain, writing-text-style machine translation system to a travel-domain, speech translation task. We study a variety of domain adaptation techniques, including data selection and incorporation of multiple translation models, in a unified decoding process. The experimental results demonstrate significant BLEU score improvement on the targeting scenario after domain adaptation. The results also demonstrate robust translation performance achieved across multiple conditions via joint data selection and model combination. We finally analyze and compare the robust techniques developed for speech recognition and speech translation, and point out further directions for robust translation via variability-adaptive and discriminatively-adaptive learning.

Index Terms: speech translation, robustness, domain adaptation, data selection, model combination

1. Introduction

Speech translation takes the source speech signal as input and produces as output the translated text of that utterance in another language [1]. It can be viewed as automatic speech recognition (ASR) and machine translation (MT) in tandem. Since ASR inevitably introduces errors and spoken utterances often contain disfluency, the ASR output in speech translation can be treated as "noisy" text to the MT, unlike the typically "clean" text dealt with in the standard MT. An analogous problem is robust ASR where the input to a speech recognizer is "noisy" or distorted speech rather than "clean" speech.

There have been a multitude of techniques developed in robust ASR over the past nearly 30 years. Six classes of compensation techniques have been analyzed in a taxonomy-oriented overview in [2]. They include feature-domain, model-domain, and hybrid compensation approaches, each being based on the unstructured or structured scheme. On the other hand, the robustness issue in MT has received rather recent attention, mainly in the context of speech translation (e.g. [3]). In this paper, we report our recent work on a particularly effective approach to handling robustness in speech translation, akin to the unstructured, hybrid class of compensation techniques in robust ASR.

The robustness problem of speech translation dealt with in this paper pertains to translation domain adaptation. Most state-of-the-art statistical MT systems are trained from large amount of bi-lingual parallel data. However, just as for ASR, the MT performance depends on the quantity and quality of the available training data. There is rarely adequate training data that are directly relevant to the translation task at hand. This problem is especially acute for speech translation where only very limited parallel spoken language data are available.

In this paper, we investigate a variety of domain adaptation methods to achieve robust translation performance on speech translation tasks. The task of domain adaptation is to adapt an MT system trained on one domain, e.g., general-domain to a more specific, target domain, for which only a small amount of training data are available. Existing domain adaptation methods, such as in [4] and [5], often fall into two broad categories. Adaptation can be done at the corpus level, by manipulating the datasets upon which the systems are trained. This is analogous to the unstructured, feature-domain approach to ASR robustness. Adaptation can also be achieved at the model level by combining multiple translation or language models together. This is in connection with the unstructured, model-domain approach to ASR robustness.

The domain adaptation methods based on data selection (i.e., feature-based) first define a metric to measure the similarity between the targeting domain, where only a small amount of data are available, and the data in the general domain. Then, based on that measure, a subset of data that are similar to the targeting domain are selected from the general domain database, and used to train the system for the targeting domain. These methods have been widely studied for language model adaptation. In [6], a perplexity (or cross-entropy) based measure is proposed to select monolingual data from a large general database to build model for a target domain. More recently, the new metric of cross-entropy difference is proposed in [7], which considers both of how close a sentence is to a language model in the target domain and how far away it is from a language model in a background domain.

Separately, there has been considerable interest in methods for effectively exploiting two translation models, one trained on a larger general-domain corpus and the other on a smaller in-domain corpus, to translate in-domain text. In [8], a method is proposed to interpolate the in-domain and general-domain phrase tables, assigning either linear or log-linear weights to the entries in the tables before combining overlapping entries. In [9], instead of directly combining phrase tables, the use of multiple phrase tables is reported.

In this work, we explore and integrate both categories of domain adaptation methods above to attack the problem of adapting a general-domain, writing-text-style MT system to a travel-domain, spoken-language translation task. We have shown that after domain adaptation, significant translation improvement on the targeting scenario is achieved. Concurrently, we also quantitatively study the impact of different adaptation methods, as well as the robustness of their performance as measured on both the targeting scenario and the non-targeting scenario. Our results in this study show that highly robust translation performance can be achieved across multiple conditions via judiciously integrated techniques of data selection, model combination, and proper training of the log-linear model.

2. Bilingual Parallel Data Selection for the Target Domain

We first focus on the task of adapting the general-domain translation system itself. Here we train and tune new MT systems on a selected subset of the original general-domain corpus. We consider a method for extracting domain-targeted parallel data from a general corpus. In our method, we use the in-domain data to rank the individual sentences of the general-domain corpus, select the top N, and then train an MT model on these N parallel sentences plus the original in-domain data.

Domain adaptation based on data selection has been shown to be effective for language model adaptation. In [6], the sentences in the general domain corpus are ranked by their perplexity score according to a language model LM_I trained on the small in-domain corpus.

The perplexity of some string s with empirical n-gram distribution p given a language model q is

$$H(p,q) = -\sum_x p(x)\log q(x) \tag{1}$$

where x represents the n-grams in s, and $H(p,q)$ is the cross-entropy between $p(x)$ and $q(x)$. We simplify this notation to just $H_I(s)$, meaning the cross-entropy of string s according to a language model LM_I which has distribution q.

Selecting sentences with the lowest perplexity is therefore equivalent to choosing the sentences with the lowest cross-entropy according to the in-domain language model.

In [7], beside the LM_I, a language model LM_O of similar size over the general-domain corpus is also constructed. Then the general-domain corpus sentences are ranked using:

$$H_I(s) - H_O(s) \tag{2}$$

and again taking the lowest-scoring sentences. This has the effect of selecting sentences that are not only similar to the in-domain corpus but simultaneously dissimilar to the average of the general-domain corpus.

For MT, bi-lingual parallel data selection is needed for translation model training, and a bi-lingual cross-entropy difference as described in [19] is used in our experiments:

$$\left[H_{I,src}(s) - H_{O,src}(s)\right] + \left[H_{I,tgt}(s) - H_{O,tgt}(s)\right] \tag{3}$$

again, the sentence selection criterion is based on low scores above.

3. Incorporating Multiple Translation Tables in a Log-Linear Model

Using the standard statistical MT terminology, the optimal translation $\hat{E}$ given the input sentence F is obtained via the decoding process according to

$$\hat{E} = \underset{E}{\operatorname{argmax}}\, P(E|F) \tag{4}$$

where the posterior probability in (4) of the output sentence E given F is computed through a log-linear model:

$$P(E|F) = \frac{1}{Z} exp\left\{\sum_i \lambda_i log\varphi_i(E,F)\right\} \tag{5}$$

In (5), $Z = \sum_E exp\{\sum_i \lambda_i log\varphi_i(E,F)\}$ is the normalization denominator to ensure that the probabilities sum to one and $\{\varphi_i(E,F)\}$ are the feature functions empirically constructed from E and F. The only free parameters of the log-linear model are the feature weights, i.e., $\Lambda = \{\lambda_i\}$.

The free parameters of the log-linear model, $\{\lambda_i\}$, are trained by maximizing the BLEU score [13] of the final translation on a hold-out *dev set*, i.e.,

$$\hat{\Lambda} = \underset{\Lambda}{\operatorname{argmax}}\, BLEU(E^*, \hat{E}(\Lambda, X)) \tag{6}$$

where E^* is the translation reference(s), and $\hat{E}(\Lambda, X)$ is the translation output. In our system, the minimum error rate training (MERT) method proposed in [10] is used.

The features used in the log-linear model usually include language models, translation models and other scoring functions. In our experiments, we use two translation models, one trained on in-domain and one on general domain data, so as to expand the coverage of the translation model. Fig. 1 presents the flow chart for information flow for data selection and multi-translation-model incorporation, which we explain below. First, we select a subset of pseudo in-domain data from a large general-domain corpus, based on their similarity to the intended target domain. Then, these pseudo-in-domain data in combination with the real in-domain data are used to train an in-domain translation model (TM). At the same time, the entire general-domain corpus is used to train a separate general-domain translation model. Finally, the scores associated with the two models trained above are treated as separate features incorporated in a unified log-linear model. In the next section, details of the features used in our experiments are provided.

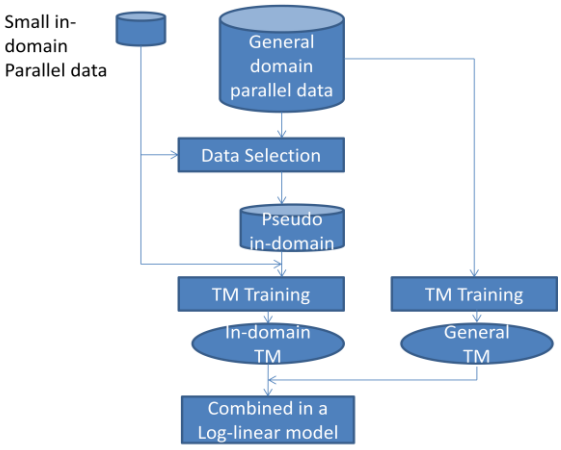

Fig. 1. Information flow diagram illustrating how domain adaptation is accomplished via data selection and integration of two translation models (TM) for speech translation.

2.2. Features in the speech translation model

The full set of feature functions constructed and used in our speech translation system is listed below:

- Forward phrase translation feature: $\varphi_{F2Eph}(E,F) = P_{TMph}(E|F) = \prod_k p(\tilde{e}_k|\tilde{f}_k)$, 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 $\tilde{e}_k$.
- Forward word translation feature: $\varphi_{F2Ewd}(E,F) = P_{TMwd}(E|F) = \prod_k \prod_m \sum_n p(e_{k,m}|f_{k,n})$, 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.)

- Backward phrase translation feature: $\varphi_{E2Fph}(E,F) = P_{TMph}(F|E) = \prod_k p(\tilde{f}_k|\tilde{e}_k)$, where $\tilde{e}_k$ and $\tilde{f}_k$ are defined as above.
- Backward word translation feature: $\varphi_{E2Fwd}(E,F) = P_{TMwd}(F|E) = \prod_k \prod_n \sum_m p(f_{k,n}|e_{k,m})$, where $e_{k,m}$ and $f_{k,n}$ are defined as above.
- Count of phrases: $\varphi_{PC}(E,F) = e^{|\{(\tilde{e}_k,\tilde{f}_k),k=1,\ldots,K\}|}$ is the exponential of the number of phrase pairs.
- Translation length: $\varphi_{TWC}(E,F) = e^{|E|}$ is the exponential of the word count in translation *E*.
- Syntactic phrase reordering feature: $\varphi_{syn}(E,F) = P_{syn}(S|E,F)$ is the probability of particular phrase segmentation and reordering *S*, given the source and target sentence *E* and *F* [11].
- Target language model (LM) feature: $\varphi_{TLM}(E,F) = P_{LM}(E)$, which is the probability of *E* computed from an N-gram LM of the target language.

Note that, in our experiments, we have two feature sets of the form $\{\varphi_{F2Eph}(E,F), \varphi_{F2Ewd}(E,F), \varphi_{E2Fph}(E,F), \varphi_{E2Fwd}(E,F)\}$ — one for the in-domain TM and one for the general-domain TM.

4. Experimental Evaluations

4.1. Experimental conditions

In the corpora used for our experiments, the language pair is English-to-Chinese. Our target domain is the conversational-style travel domain. Two corpora are needed for the adaptation task: one in-domain to serve as a target, and one external corpus which we use and seek to adapt to the target. We have collected about 800K in-domain sentence pairs. Our general-domain corpus was a private set of 12 million parallel sentences amassed from a variety of sources. We evaluated our work on an internal test set, which contains 2000 sentences of travel relevant data with conversational style. We denote this test set as *tvl_test* hereafter. Meanwhile, we also tested our techniques on a general-domain test set that contains 5000 sentences of general-domain data, which we call *gen_test* hereafter. Both test sets have one reference per testing sentence.

In our speech translation experiments reported in this section, we used a syntax-based translation system as the base system [11], and used MERT [10] to train and tune the MT systems. Word alignment is based on the word-dependent HMM-based alignment method proposed in [12].

In our experiments, a total of four language models, for both travel and general domains as well as for both source and target sides, are trained to perform data selection. Based on the method described on section 2, we have selected about 800K sentences of pseudo travel domain data from the 12M general domain sentences. We constructed 4-gram language models. The two target-side language models are subsequently used in the log-linear model for decoding.

4.2. Experimental results

In our experiments, we first evaluate the effectiveness of domain adaptation by data selection. The translation accuracy is measured in BLEU scores [13]. We compare two systems: 1) the baseline system which uses a general-domain translation model and a general-domain *dev set* for the log-linear model training; and 2) the adaptive system which is trained on the combination of selected and original travel domain data, and which uses a travel-domain *dev set* for the log-linear model training. From the comparative results in Table 1 (middle row labeled "Data selection"), the adaptive system clearly improves the performance, by more than 6 BLEU points on the travel domain test set. This indicates that the data selection based method can effectively adapt the MT to the target domain. However, the results on the general domain test set also show serious translation performance degradation, by actually more than 8 BLEU points. This demonstrates lack of robustness in the adapted system for non-target-domain tests.

Table 1: BLEU scores for domain adaption via data selection and the use of multiple models

	tvl_test	*gen_test*
General (baseline)	16.22	18.85
Data selection	22.32 (+6.1)	10.81(-8.0)
Multi-TMs	22.12 (+5.9)	13.93(-4.9)

We then evaluated the data-selection method with the use of both travel-domain and general-domain translation models in the log-linear model. Here we used the same travel domain *dev set* for log-linear model training. The results are shown in the final row of Table 1 (labeled "Multi-TMs"). Compared with the use of a single TM with data-selection (middle row), the BLEU score with the use of multi-TMs is virtually the same (middle column) on the travel test. However, it has a smaller drop on the general test (right column); i.e., outside-travel-domain data. This provides evidence of robustness against non-target-domain tests for the joint use of data selection and multiple models.

The log-linear model is critical for achieving a robust translation performance. In our experiments, we further investigated the roles and effectiveness of using multi-TMs with respect to different ways of training the log-linear model. We tested the systems using three separate ways of designing the held-out *dev set* for log-linear model training, with the results shown in Table 2 where the three Multi-TMs systems have the identical features. The only difference is that their associated log-linear models are optimized on separate *dev sets* containing the travel- and general-domain data with different proportions.

As shown in Table 2, robustness of the system with the use of multiple models depends strongly on how the log-linear models are trained. After carefully choosing the *dev set,* we are able to achieve 4~6 BLEU point improvement on the target travel-domain tests, with very minor performance degradation on the general-domain tests.

Table 2: BLEU scores with multiple TMs for three different ways of training the log-linear translation model

	tvl_test	*gen_test*
General (dev: general)	16.22	18.85
Multi-TMs (dev: travel only)	22.12	13.93
Multi-TMs (dev: *tvl* : *gen* = *1:1*)	22.01	16.89
Multi-TMs (dev: *tvl* : *gen* = *1:2*)	20.24	18.02

4.3. Illustrative translation examples

In this section, we further present a set of typical examples to analyze where the performance improvement comes from with the application of the domain adaptation techniques discussed so far in this paper.

In Table 3, several translation examples are shown, all drawn from the travel test set. For each English source sentence, we present the translated Chinese sentences from the general-domain baseline system (MT1) and that from the adapted system (MT2).

Examining a few dozens of examples, five of which are shown in Table 3, we observed that the improvement of the

adaptive system comes typically from better handling of syntax and word re-ordering, and better picking of the right sense of the ambiguous words. E.g., in Example 1 of Table 3, the adapted system re-orders the translation of "please" to the beginning of the Chinese sentence, leading to a better Chinese sentence than the baseline system. Example 3 provides evidence of the adapted system picking more travel-relevant sense from ambiguous word in the translation. English word "tip" could be "hint" or "change-for-service", and here the travel-domain adapted system picked the right translation "**小费**" (change-for-service), while the general domain system picked "**提示**" which, though more popular in the general text, is not the right translation in the travel context. The same correction has been shown to other words in Table 3, such as "open" (which could mean un-fold, or in-business), and "change" (which could mean make-a-difference or tip-for-service).

Table 3: Examples of E2C translation extracted from the outputs of our baseline vs. domain-adaptive systems

1	source	A glass of cold water, please .
	MT1	一杯冷水请。
	MT2	请来一杯冷的水。
2	source	Please keep the **change** .
	MT1	请保留更改。
	MT2	不用找了。
3	source	Thank , this is for your **tip** .
	MT1	谢谢，这是为您**提示**。
	MT2	谢谢，这是给你的**小费**。
4	source	Do you know of any restaurants **open** now ?
	MT1	你现在知道的任何**打开**的餐馆吗？
	MT2	你知道现在还有餐馆在**营业**的吗？
5	source	I'd like a restaurant with cheerful atmosphere
	MT1	我想就食肆的愉悦的气氛
	MT2	我想要一家气氛活泼的餐厅。

5. Discussion and Future Work

In this paper, we reported two approaches and their integration for robust domain adaptation in a speech translation system. The first approach replies on data selection and processing, pertaining to "feature"-domain robustness technique in the analogous ASR terminology. The second approach makes use of multiple translation models built from different domains, relating to multi-style training or the (unstructured) "model"-domain robustness technique in the analogous ASR terminology (rf. [2]). The joint use of the two approaches, which has been demonstrated to be highly successful for speech translation as reported in this paper, bears some resemblance to the "unstructured" hybrid category, a terminology coined in [2] in the taxonomy of robust ASR techniques.

The analysis and comparisons of robust techniques developed for speech recognition and translation can shed insightful light onto how to further improve speech translation robustness. First, the essence of hybrid feature-model techniques in robust ASR is to normalize out all irrelevant variability in the recognition task. Noise adaptive learning [2] is the most prominent approach of this kind, which has been extended in various directions (e.g., [15]) since its inception ten years ago. This essence, however, has not been captured in the hybrid domain-adaptive approach reported in this paper. Extension of the current method with joint use of data selection and multiple models along the direction of normalizing the text variability expressed in a hidden-semantic domain can likely lead to better performance gain.

Second, discriminative adaptation has been shown to be a powerful technique for robust ASR (for an overview, see [2]). Our recent work on speech translation (e.g., [1, 16, 17, 18]) pioneered the direction of end-to-end discriminative learning for the entire speech translation system design. Integration of discriminative learning with the problem of translation system's adaptation as an extension of the straightforward domain-adaptation method presented in this paper is another key direction of our future work.

6. Acknowledgements

We thank Alex Acero and Rico Malvar for the encouragement and support of this work.

7. References

[1] X. He and L. Deng, "Speech recognition, machine translation and speech translation — a unified discriminative learning paradigm," *IEEE Signal Proc. Mag.* 2011. Accepted.

[2] L. Deng, "Feature-domain, model-domain, and hybrid approaches to noise-robust speech recognition," in D. Kolossa and R. Hab-Umbach (eds.) *Robust Speech Recognition of Uncertain Data*, Springer Verlag, 2010.

[3] W. Wang, G. Tur, J. Zheng, N. Fazil Ayan, "Automatic disfluency removal for improving spoken language translation," in *Proc. ICASSP*, 2010.

[4] M. Eck, S. Vogel, and A. Waibel, "language model adaptation for statistical machine translation based on information retrieval," *Proc. of Language Resources and Evaluation*. 2004.

[5] P. Nakov, "Improving English-Spanish statistical machine translation: experiments in domain adaptation, sentence paraphrasing, tokenization, and recasing," *Proc. of WMT ACL*. 2008.

[6] J. Gao, J. Goodman, M. Li, and K-F. Lee, "Toward a unified approach to statistical language modeling for Chinese. *ACM Trans. on Asian Lang. Infor. Proc.* 2002.

[7] R. Moore and W. Lewis, "Intelligent selection of language model training data," *Proc. of ACL*. 2010.

[8] G. Foster and R. Kuhn, "Mixture-model adaptation for SMT," *Proc. of WMT ACL*. 2007.

[9] P. Koehn and J. Schroeder, "Experiments in domain adaptation for statistical machine translation," *Proc. of WMT ACL*. 2007.

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

[11] C. Quirk, A. Menezes, and C. Cherry, "Dependency treelet translation: syntactically informed phrasal SMT," *Proc. of ACL*. 2005.

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

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

[14] L. Deng, A. Acero, M. Plumpe, and X.D. Huang, "Large vocabulary speech recognition under adverse acoustic environments," *Proc. of ICSLP*, 2000.

[15] Y. Hu and Q. Huo, "Irrelevant variability normalization based HMM training using VTS approximation of an explicit model of environmental distortions," *Proc.of Interspeech*, 2007.

[16] 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?" *Proc. of ICASSP,* 2011.

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

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

[19] A. Axelrod, X. He, and J. Gao, "Domain adaptation via pseudo in-domain data selection," *Proc. of EMNLP*, 2011