

INTEGRATING TEXT AND PHONETIC INFORMATION FOR ROBUST STATISTICAL SPEECH TRANSLATION

Liang Gu, Yonggang Deng, Wei Zhang and Yuqing Gao

IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA

ABSTRACT

This paper focuses on the use of both text and phonetic information in a speech translation system in order to make translation results more robust to speech recognition errors. Conventional statistical speech translation formulas are extended to exploit both text-form and phonetic speech recognition results. A novel data-driven word/text tying algorithm is then proposed to group words based on both pronunciation similarity and meaning equivalency. In our speech-to-text translation experiments, significant improvement was achieved by using phonetic information and the proposed word tying algorithm.

1. INTRODUCTION

Automatic speech translation extracts information from spoken utterances in one language and translates it into another language. Most state-of-the-art statistical speech translation systems consist of two cascaded components: automatic speech recognition (ASR) followed by statistical machine translation (SMT). Typically, these components are designed and optimized separately. The resulting translation systems are usually very sensitive to speech recognition errors, which hence greatly deteriorates speech translation robustness and accuracy, especially when the ASR performance is poor.

A straightforward approach towards robust speech translation is to jointly optimize the ASR and SMT procedures during speech translation. One typical way is to tightly couple the ASR process and the SMT process by inheriting as much information as possible from the ASR process and utilize it in the following SMT process [1, 2, 3, 4]. The corresponding translation hypotheses can be re-scored based on pre-trained integration models such as weighted finite state transducer (WFST) [2, 4] or log-linear models [3]. While promising experimental results have been reported, the information shared between ASR and SMT in these approaches is all in text form. In particular, the phonetic information is not explicitly utilized, if not ignored at all, in the integration of the ASR and the SMT processes.

In this paper, we believe that the phonetic recognition results represent important supplementary information and hence

should be exploited and integrated with text-form recognition results during the joint optimization process of ASR and SMT in order to achieve more robust speech translation performance. Therefore, we propose a novel statistical speech translation approach that utilizes both the text and phonetic information in the ASR outputs. The new SMT models are trained upon not only text-to-text parallel corpus but also phone-to-text parallel corpus. During speech translation, both the text-based features and phone-based features are extracted from the ASR results and then sent to the SMT decoders, by which the best translation hypothesis is achieved.

Another common issue in spoken language translation is the data sparseness problem. To improve speech translation robustness and accuracy when the training data is limited, we further propose a new word/text tying algorithm that automatically ties the words with both the same pronunciation in the source language and the same corresponding translation in the target language. The total number of unique words to be translated is hence reduced while the effective vocabulary coverage is unchanged, which could significantly alleviate the training data sparseness problem and hence enhance speech translation robustness.

2. PHONETIC INFORMATION VS. TEXT INFORMATION IN SPEECH TRANSLATION

2.1. Problem Statement

Let \mathbf{x} denote acoustic observations of a speech utterance in the source language. In order to obtain the best target language sentence, we typically look for a target language translation with maximum posterior translation probability given the observed speech:

$$\hat{e} = \operatorname{argmax}_e P(e|\mathbf{x}) = \operatorname{argmax}_e \sum_f P(e, f|\mathbf{x}) \quad (1)$$

$$\approx \operatorname{argmax}_e \sum_f P(e|f)P(f|\mathbf{x}) \quad (2)$$

where f is a text-form speech recognition hypothesis in the source language. Conventional cascaded speech translation approach further approximates equation 1 as:

$$\hat{e} \approx \operatorname{argmax}_e P\{e|\hat{f} = \operatorname{argmax}_f P(f|\mathbf{x})\} \quad (3)$$

This work was supported in part by the DARPA TransTac program.

ASR	Text output: 华尔街一百好 Phone output: (HU_A_ER_JI_IE) (Y_I_B_AI) (H_AO) Text-based Translation: "Wall Street one hundred good" Phone-based Translation: "100 Wall Street"
True	True Text: 华尔街一百号 True Phone: (HU_A_ER_JI_IE) (Y_I_B_AI) (H_AO) Text-based Translation: "100 Wall Street" Phone-based Translation: "100 Wall Street"

Table 1. Example of Mandarin-to-English Translation with erroneous text recognition result but correct phonetic recognition result

ASR	Text output: 似的 Phone output: (SH_IH_D_E) or (S_IH_D_E) Text-based Translation: "like" Phone-based Translation: "yes" or "like"
True	Text: 是的 Phone: (SH_IH_D_E) Text-based Translation: "yes" Phone-based Translation: "yes"

Table 2. Example of Mandarin-to-English Translation errors that are caused by multiple pronunciations and accents

where the inside argmax represents the ASR process and the outside argmax represents the SMT process, respectively.

In this paper, we propose to incorporate phonetic information into the above speech translation scheme. Assume phone sequence ϕ is hypothesized, equation 1 can be re-written as:

$$\hat{\mathbf{e}} = \underset{\mathbf{e}}{\text{argmax}} P(\mathbf{e}|\mathbf{x}) = \underset{\mathbf{e}}{\text{argmax}} \sum_{\mathbf{f}, \phi} P(\mathbf{e}, \mathbf{f}, \phi|\mathbf{x}) \quad (4)$$

$$\approx \underset{\mathbf{e}}{\text{argmax}} P\{\mathbf{e} | (\hat{\mathbf{f}}, \hat{\phi}) = \underset{\mathbf{f}, \phi}{\text{argmax}} P(\mathbf{f}, \phi|\mathbf{x})\} \quad (5)$$

where $(\hat{\mathbf{f}}, \hat{\phi})$ is the best pair of text-form and phonetic ASR hypotheses.

2.2. Phonetic information vs. text information

Note that Phonetic information ϕ and text information \mathbf{f} are usually coupled together via a pre-defined pronunciation lexicon $\phi = g(\mathbf{f})$. This text-to-phone and phone-to-text mapping is often not deterministic and the ASR process typically relies on language models to get both of them right. In practice, however, current monolingual language models are not powerful enough to pick out the correct words all the time. Table 1 shows an example in Mandarin-to-English translation (Without loss of generality, tones are not modeled in the phonetic information). In this case, the text-form ASR error causes serious mistakes in the text-based translation. Nevertheless, if we can optimize and apply equation 4 properly, we may exploit the context constraints in both the source language and target language and therefore achieve accurate translation even if the text ASR result is incorrect, as the perfect phone-based translation example shown in Table 1.

Moreover, phonetic and text-form ASR results represent information in the input speech utterance on different lev-

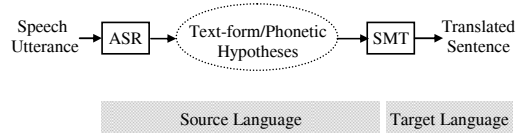


Fig. 1. Speech Translation via both Text-form and Phonetic information.

els. While the text-form ASR result conveys most of the written form message, the phonetic ASR result maybe more colloquial and conveys most of the spoken form message. In colloquial speech, many words have several pronunciations and, to make things worse, these pronunciations may be further expressed in various ways because of accents. Table 2 illustrates one example in Mandarin-to-English translation. The most popular response “是的(yes)” could be pronounced as (S_IH_D_E) rather than (SH_IH_D_E) for speakers with southern accents as they often pronounce “SH” as “S”. In the meantime, there is another word “似的(like)” that may be spoken as both (S_IH_D_E) and (SH_IH_D_E). Consequently, “是的(yes)” may be frequently misrecognized as “似的(like)” and, as shown in the table, the corresponding translations differ a lot because of this simple ASR mistake. Therefore, we propose to integrate ASR results at both the phonetic and text-form levels as a natural way to exploit both the phonetic message embedded in the observed speech utterance and the text-form information exists in current extensive text-form bi-lingual corpora.

3. SMT USING PHONETIC/TEXT INFORMATION

Our SMT system employs the state-of-the-art log-linear translation model [5],

$$\hat{\mathbf{e}} = \underset{\mathbf{e}}{\text{argmax}} P(\mathbf{e}|\mathbf{f}, \phi) = \underset{\mathbf{e}}{\text{argmax}} \sum_{m=1}^M \lambda_m h_m(\mathbf{e}, \mathbf{f}, \phi) \quad (6)$$

where λ_m is model parameters of feature function h_m , $m = 1, 2, \dots, M$. They are tuned automatically via discriminative minimum error rate training on a development set. Active features include language model $P(\mathbf{e})$, sentence length, distortion model, and most importantly, translation distribution in both directions, $P(\mathbf{e}|\mathbf{f}, \phi)$ and $P(\mathbf{f}, \phi|\mathbf{e})$.

The conditional probability distribution $P(\mathbf{f}, \phi|\mathbf{e})$, and similarly the other translation direction, can be optimized upon a bilingual corpus containing both phonetic and text information. The optimal distribution may then be achieved as [6]:

$$\hat{P}(\mathbf{f}, \phi|\mathbf{e}) = \underset{P}{\text{argmax}} \prod_{i=1}^I P(\mathbf{f}^i, \phi^i|\mathbf{e}^i) \quad (7)$$

where $(\mathbf{f}^i, \phi^i, \mathbf{e}^i)$ is the i -th sentence pair in the training corpus. In practice, it is not trivial to obtain bilingual parallel corpora containing both text-form and phonetic information.

Word/Phrases	Phone Sequence	Meanings
ليجوه ليجوه	(L.J.G.W.HH)	down
إنّا إنّا إنّا	(HM.THL.AL)	if

Table 3. Example of different Arabic words with both almost the same pronunciations and equivalent meanings

An alternative way is to setup such bilingual corpora only in limited domains and interpolate it with larger generic corpora $\{\mathbf{f}^i, \phi^i = g(\mathbf{f}^i), \mathbf{e}^i\}$ that derive phonetic information from the text information.

The phonetic and text based speech translation procedure is illustrated in Figure 1. In this new speech translation scheme, for an input speech utterance \mathbf{x} , the most appropriate target language sentence $\hat{\mathbf{e}}$ is searched as follows:

1. Given \mathbf{x} , a set of text-form and phonetic hypotheses is obtained via an ASR decoder. Either the best hypothesis or the top N-best results may be generated;
2. The recognition hypotheses are translated by the SMT models optimized according to equation 7;
3. The best target sentence $\hat{\mathbf{e}}$ is selected from all translation hypotheses in step (2) by maximizing the linear combination of feature functions as in equation 6.

4. WORD TYING BASED ON PHONETIC AND MEANING EQUIVALENCY

One of the most critical issues in speech translation is the data sparseness problem. Compared to traditional bilingual written text corpora, the bi-lingual spoken training data is dramatically limited due to the difficulty involved in recording, transcription and translation. As a result, bi-lingual spoken or colloquial data is practically almost always insufficient for spoken language translation applications. A common way to alleviate this problem is to perform word clustering. Conventionally, the automation of this work is achieved by iteratively clustering words based on SMT models [7]. However, the extremely challenging nature of word clustering (in many cases even human being can not decide whether some words should be grouped together or not), coupled with inevitable word alignment errors when SMT models are applied, makes highly accurate word clustering and hence, the improvement in speech translation quality, doubtful. Motivated by the above concerns, we propose a new scheme to tie only a small portion of words in the vocabulary list but with a guaranteed grouping accuracy, as described next.

In speech translation, it is not uncommon that different words or phrases have both the same pronunciations and the same meanings. Examples are shown in Table 3 for Arabic-to-English speech translation. These words all have the same sounds and meanings while their written form is different. Although some of these variations may be normalized by hand-crafted rules, these rules are typically neither designed for nor

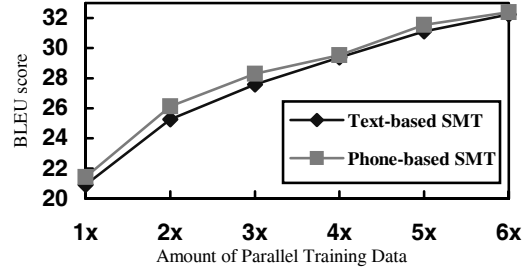


Fig. 2. Arabic-to-English speech translation performance using text or phonetic information.

matched with our speech translation purpose. Instead, we propose to group these words automatically.

Without losing generosity, let single word \mathbf{f}_1 and \mathbf{f}_2 have the same phone sequence ϕ as well as the same target translation \mathbf{e} . We further assume that \mathbf{f}_1 and \mathbf{f}_2 are the only two words that can be the translation of \mathbf{e} . Denote the corresponding conditional translation probabilities as $p_1 = P(\mathbf{f}_1, \phi|\mathbf{e})$ and $p_2 = P(\mathbf{f}_2, \phi|\mathbf{e})$, respectively.

By tying \mathbf{f}_1 and \mathbf{f}_2 as $\tilde{\mathbf{f}}$, we get $\tilde{p} = P(\tilde{\mathbf{f}}, \phi|\mathbf{e}) = p_1 + p_2$ and achieve:

1. higher conditional translation probability \tilde{p} for both \mathbf{f}_1 and \mathbf{f}_2 ;
2. more accurate word alignment of \mathbf{f}_1 and \mathbf{f}_2 in equation 7 and, in turn, more accurate and robust estimation of conditional translation probability \tilde{p} .

The tying of these words is even more beneficial for phrase-based SMT (such as that in [8]) as all the phrases with the only difference in \mathbf{f}_1 and \mathbf{f}_2 can be tied together when generating the phrase translation tables.

In practice, while it is straightforward to find out words with the same pronunciations if a pronunciation lexicon $\phi = g(\mathbf{f})$ exists, it is a substantial challenge for any data-driven approaches to determining the word groups with equivalent translation. Here we propose a statistical word/text tying algorithm using translation distribution distance. With the above detonation, \mathbf{f}_1 will be tied with \mathbf{f}_2 into $\tilde{\mathbf{f}}$ if and only if the Kullback-Leibler divergence of two translation distributions is less than a pre-defined threshold:

$$D_{KL}(P(\mathbf{e}|\mathbf{f}_1)||P(\mathbf{e}|\mathbf{f}_2)) < \beta \quad (8)$$

where β is a threshold for word tying that can be optimized on a development set.

5. EXPERIMENTAL RESULTS

Experiments were carried out on speech translation for colloquial Arabic to English. The word vocabulary size is 80K and 50K for Arabic and English, respectively. The test set consists of 1440 Arabic utterances. Automatic speech recognition results show a 15.5% word error rate (WER), 14.7% phone sequence error rate (PSER), and 6.7% phone error rate

	Speech Translation of ASR reference	Speech Translation of ASR output
Text-based SMT	34.18	31.85
Phone-based SMT	35.16	32.07
SMT with word tying	36.18	33.52

Table 4. Comparison of Arabic-to-English speech translation BLEU score with and without proposed word tying algorithm

(PER). The consistent decreasing of recognition error rate not only demonstrated the importance of phonetic information in speech translation, but also the importance of integrating the text-form and phonetic information within the entire translation procedure.

The statistical translation models are trained on 271K parallel Arabic-English utterances, which consist of 1.97M English words and 1.37M Arabic words. We use the GIZA++ Toolkit [9] to train IBM models [6] and extract statistical phrase translation models using heuristics based on word alignments.

To evaluate and compare the usefulness of text and phonetic information in speech translation, two sets of SMT models were trained. Set A is with text information only and Set B with phonetic information only. For the latter experiments, text-form bilingual parallel corpus was converted to phone-to-text parallel corpus according to a pre-defined pronunciation lexicon. In order to check the impact of training data size, each model set contains six models trained on various amount of training data. Figure 2 shows the experimental results in BLEU [10] score. The phone-based models are better than text-based models when the training data is small, mainly due to the implicit smoothing effect of phonetic information that discussed in section 2.2. When more training data is available, the performances of these two SMT models become closer, which indicates the importance of combining the two distinct types of information in the speech translation processes.

Additional experiments were carried out to couple text and phonetic information by tying the words with same pronunciations and equivalent meanings. The pronunciation of each word is represented by its phoneme sequence. The meaning of each word is represented by its statistical phrase-based translation table derived from SMT models. The words in the vocabulary list are grouped and tied according to equation 8, where we found $\beta = 3$ achieved best performance on a development set. Experimental translation results on input stream with or without ASR errors are illustrated in Table 4. In both cases, the number of unique words was reduced, which leads to the superior translation accuracy of translation via phonetic/meaning based word tying achieved superior compared with either text-based or phone-based translation.

6. CONCLUSIONS

We have investigated the potential benefit of integrating phonetic information with conventional text-form information in

speech translation in order to achieve translation results that are more robust to speech recognition errors. An extended statistical machine translation scheme is proposed to utilize both phonetic and text information. During speech translation, text-form hypotheses as well as phone sequence hypotheses are derived from speech recognizers and passed to translation models. To alleviate data sparseness problem, we further propose a novel word tying algorithm that groups words with same pronunciations and equivalent meanings. Superior speech translation performance was achieved in our Arabic-to-English speech translation experiments. In the future, we plan to more extensively explore the use of phonetic information in speech translation and test our approach on more language pairs .

7. REFERENCES

- [1] H. Ney, “Speech translation: Coupling of recognition and translation,” in *Proc. of ICASSP*, 1999, pp. 517–520.
- [2] L. Mathias and W. Byrne, “Statistical phrase-based speech translation,” in *Proc. of ICASSP*, 2006.
- [3] R. Zhang and et al, “A unified approach in speech-to-speech translation: integrating features of speech recognition and machine translation,” in *Proceedings of the 20th International Conference on Computational Linguistics*, 2004, pp. 1168–1174.
- [4] E. Matusov, S. Kanthak, and H. Ney, “On the integration of speech recognition and statistical machine translation,” in *Proc. of Interspeech*, September 2005.
- [5] F. J. Och and H. Ney, “Discriminative training and maximum entropy models for statistical machine translation,” in *Proc. of ACL*, 2002, pp. 295–302.
- [6] P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer, “The mathematics of machine translation: Parameter estimation,” *Computational Linguistics*, vol. 19, pp. 263–312, 1993.
- [7] Y. Wang, J. Laerty, and A. Waibel, “Word clustering with parallel spoken language corpora,” in *Proc. of IC-SLP*, 1996.
- [8] P. Koehn, F. Och, and D. Marcu, “Statistical phrase-based translation,” in *Proc. of HLT-NAACL*, 2003.
- [9] F. J. Och and H. Ney, “A systematic comparison of various statistical alignment models,” *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.
- [10] K. Papineni, S. Roukos, T. Ward, and W. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in *Proc. of ACL*, 2002, pp. 311–318.