Development of SRI's Translation Systems for Broadcast News and Broadcast Conversations

Jing Zheng, Wen Wang, Necip Fazil Ayan

Speech Technology and Research Laboratory, SRI International
{zj,wwang,nfa}@speech.sri.com

Abstract

We present our recent work on developing large-vocabulary Arabic-to-English and Chinese-to-English speech-to-text translation systems for the January 2008 Global Autonomous Language Exploitation (GALE) retest evaluation. Two audio genres were involved in the evaluation: broadcast news and broadcast conversation.

Our system, following the hierarchical phrase-based translation approach, has a two-pass decoding strategy, with the first-pass integrated search generating 3000 unique n-best lists, which are then reranked by several different language models in the second pass.

We emphasize our work on adapting the system, which was mostly trained on text data, to the speech genres, including number tokenization, punctuation compensation, and various optimization techniques. We present our results on several different tuning and testing data sets used for system development.

Index Terms: speech-to-text translation, hierarchical-phrase-based translation, n-best reranking

1. Introduction

SRI's latest large-vocabulary speech-to-text translation systems were developed for the recent Global Autonomous Language Exploitation (GALE) retest evaluation in January 2008. A major goal of the GALE program sponsored by the U.S. Defense Advanced Research Projects Agency (DARPA) is to develop high-quality translation technology that can translate both text and speech documents. Arabic-to-English and Chinese-to-English are the two language pairs of interest, and broadcast news (BN) and broadcast conversation (BC) are the two pursued speech genres. Compared to text-to-text translation, speech documents bring additional challenges in two aspects.

The first aspect lies in data mismatch. Currently, the vast majority of parallel data resources for building statistical machine translation (SMT) systems come from text translation corpora accumulated during the past two decades. For the GALE program, only a small amount of BN and BC data was translated, far from enough to build a high-quality translation system. Therefore, it is essential to make use of the training data from other genres, especially from text genres. To achieve this we need to process data in the way that minimizes difference between text and transcribed speech.

The second aspect relates to information loss. The speech-to-text translation system takes input from automatic speech recognition (ASR), which does not contain any punctuation information. However, for the purpose of readability, appropriate punctuation marks are desired in the translation output, to help users understand the content. In addition, ASR is error prone, especially for the BC genre, where people speak mostly in conversational style and ASR tends to make more errors. Information loss makes speech translation more difficult than text translation, and requires different modeling.

The rest of the paper is organized as follows. Section 2 gives a brief overview of our SMT system and the related models. Section 3 reports our work adapting translation system to speech genres. Section 4 discusses system optimization issues. Section 5 gives results of the full system on three different test sets. Section 6 concludes the paper.

2. SRI SMT System Overview

The SRI SMT system uses a two-pass decoding approach built on SRInterp™ [1] technology. In the first pass, a hierarchical phrase-based decoder [2] is used to generate 3000 unique n-best hypotheses, in an integrated search with a 4-gram language model. In the second pass, five high-order language models are applied to rescore the n-best lists from the first pass. The final translation output is selected from the reranked n-best list by combining the scores from all knowledge sources via log-linear combination.

2.1. Hierarchical Phrase-based Translation

In the statistical machine translation framework, the system is given a sentence in the source language $f_1^J = f_1 \ldots f_J$ that is to be translated into the target language $e_1^I = e_1 \ldots e_I$ that maximizes $P(e_1^I | f_1^J)$:

$$\hat{e}_1^{\hat{I}} = \arg\max_{I,e_1^I} P(e_1^I \mid f_1^J) \qquad (1)$$

where $P(e_1^I | f_1^J)$ is modeled with log-linear models using several feature functions [3]:

$$P(e_1^I \mid f_1^J) = \frac{1}{Z(\Lambda, f_1^J)} \exp\left(\sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J)\right) \qquad (2)$$

The denominator part, which ensures proper probabilistic distribution, only depends on the source sentence f_1^J and scaling factors $\Lambda = \{\lambda_1, \ldots \lambda_m\}$ and therefore can be omitted during the decoding process:

$$\hat{e}_1^{\hat{I}} = \arg\max_{I,e_1^I} \exp\left(\sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J)\right) \qquad (3)$$

For the hierarchical phrase-based translation model [1], Equation 3 can be factorized into

$$\hat{e}_1^{\hat{I}} = \arg\max_{I,e_1^I} \max_{D: f(D)=f_1^J; e(D)=e_1^I} \omega(D) P_{LM}(e_1^I)^{\lambda_{LM}} I^{\lambda_L} \qquad (4)$$

where D is a derivation that generates source sentence f_1^J and target sentence e_1^I based on parsing with synchronous context-free grammars (SCFGs). λ_{LM} and λ_L are scaling factors for language model and sentence length. $\omega(D)$ is the total score of all the SCFG rules involved in D:

$$\omega(D) = \prod_{r \in D} \exp\left(\sum_{k=1}^{K} \lambda_k h_k(r)\right) \qquad (5)$$

where $h_k(.)$ are rule-level feature functions. In our system, we use the following seven features, in addition to the language model and sentence length features:

- Relative frequency [6] in both directions
- Lexical weights [6] in both directions
- Binary: rule containing non-terminals
- Binary: rule containing terminals only
- Binary: gluing rule

The goal of decoding is to find the target sentence generated by the optimal derivation defined by Equations 4 and 5. A CKY-style search algorithm with language model intersection is implemented in our in-house SRInterp™ decoder for this purpose, following Chiang's approach. The feature weights are optimized by maximizing dev set BLEU [4] score with minimum error rate training (MERT) [5].

We observed significant improvement by switching from the phrase-based translation approach [6] to the hierarchical approach in text translation tasks. Therefore we used hierarchical phrase-based translation for speech translation as well. Table 1 shows some comparison results in the NIST-MT06 GALE portion of text data, denoted Eval06 test set.

Table 1. Comparing hierarchical phrase-based translation and phrase-based translation in Eval06 set. Score reported in case-insensitive BLEU (%) with single reference.

	Arabic-to-English		Chinese-to-English	
	Phrase	Hier	Phrase	Hier
Newswire	24.1	27.2	14.9	17.6
Web text	13.4	15.4	12.8	13.8

Table 2. Comparing three punctuation compensation approaches in eval06 test set. Results reported in case-insensitive BLEU (%) with single reference.

	Chinese eval06 test set	
	BN	BC
Manual punctuation	17.4	16.4
Approach 1	16.3	15.4
Approach 2	15.0	14.3
Approach 3	17.0	16.2

2.2. Rescoring Language Models

In the second pass, we used five additional language models (LMs) to rescore the n-best lists generated from the integrated search:

- Google n-gram LM, a 5-gram count-based language model estimated on the Google Tera Word n-gram corpus with deleted-interpolation smoothing [7].
- Yahoo Web LM, a 4-gram language model estimated from the Yahoo n-gram corpus with the modified Kneser-Ney (KN) smoothing algorithm [8]. Since the singletons are not included in the corpus, to estimate KN discounting parameters, the count of singleton n-grams was extrapolated from higher-occurrence n-gram count statistics using an empirical rule [9].
- SARV LM, an adaptive super-ARV class-based 5-gram language model with unsupervised adaptation. The first-pass decoding hypotheses were used to adapt the language model [9].
- Parser-based LM, a 5-gram language model derived from dependency parsing, which used a simplified base-NP model. Some approximations were made to speed computation of language model scores [9].
- BBN web LM, a count-based 5-gram language model built from BBN automatically collected web data.

3. Adaptation to Speech Genres

Most SMT training data came from text genres, and the vast majority is newswire text. Comparing automatic speech transcription and newswire text, two salient differences must be addressed: numbers and punctuation.

3.1. Number Tokenization

In English written text, most numbers (especially large numbers) are represented in digit forms, such as 123. However, in spoken language, numbers needed to be spelled out, such as "one hundred and twenty three." In the final translation, the digit form is preferred because it is much easier to read. To effectively use text training data, we used language-specific tools to convert spelled-out numbers to digits before translation, and tagged those numbers for MT systems. The tools were developed and made available by our collaborators within the NIGHTINGALE consortium led by SRI. This processing makes speech data consistent with text data in terms of number representation.

3.2. Punctuation Compensation

Punctuation is typically not pronounced in natural spoken languages. Therefore, in ASR output there are no punctuation marks. However, punctuation marks are very important for SMT. First, strong punctuation, including the period ".", question mark "?", and exclamation mark "!", provides sentence boundary information. Almost all state-of-the-art SMT systems operate on sentence level, so it is essential to recover sentence boundaries in ASR output. Second, punctuation provides relatively stable anchor points for the widely used statistical word alignment algorithms based on bilingual co-occurrence, as their co-occurrence rate is typically much higher than that of ordinary words. Removing punctuation from training data to match speech conditions will certainly hurt SMT performance. Finally, punctuation marks are also important to make translation output easy to understand. Therefore, even though no punctuation information exists in the input sentence, the SMT system needs to inject it into the output.

We applied a sentence boundary detector and a sentence type classifier using lexical and prosodic information developed by our collaborators to segment continuous ASR output into sentence-like units, and append the strong punctuation marks ".!?" to the end of the sentence units according to sentence type classification results. Comma prediction was also investigated [10], though the accuracy was not high enough to be directly used as SMT input. Some research investigated the use of commas to constrain search in SMT, which is not within the scope of this paper.

So the question becomes what to do with the rest of punctuation, especially commas. One simple but effective approach is to remove commas from the source side of the SCFG rules, but keep them in the target side [11]. In this way, the rules can match the input sentence, which does not have punctuation, but can produce punctuation in the target side, constrained by the language model. As stated earlier, punctuation is important for word alignment, so we kept the word alignment process intact. When extracting rules from the aligned bitexts, we can remove the source side punctuation. To implement the idea, we experimented with the following three approaches:

1. Extract and score rules as usual, directly filter out source-side punctuation from the final extracted rule tables, and remove empty source-side rules.
2. Remove punctuation from the source side of training data, map the alignment to reflect the source-side index change, and extract rules as usual.
3. Extract rule instances as usual, and filter out source-side punctuation before computing feature scores for the extracted rule instances.

Approaches 1 and 3 seem similar, but the former will generate duplicated rules after punctuation filtering. The marginal counts of source-side rules will not reflect punctuation filtering, and therefore will produce different relative frequency scores compared to the third approach.

We compared the three different approaches in a Chinese-to-English translation task, as commas and other punctuation occur more frequently in Chinese text than in Arabic text. Table 2 shows some experimental results in the eval06 test set, in which manual speech transcriptions are used for MT input. We removed all the commas and other weak punctuation from the transcripts to simulate ASR output. We also computed the result of keeping all punctuation from the manual transcripts with number tokenization and treated transcripts like text. This result can serve as the upper bound for any punctuation compensation approach.

From Table 2 we can see that among the three approaches, the third worked best. The loss caused by missing punctuation is 0.4% BLEU for BN and 0.2% for BC. We therefore used this approach for the rest of the experiments described in this paper.

4. Optimization

As we use log-liner models, we need to optimize the scaling factors, also called *feature weights,* for each of the features involved. For the first-pass integrated search, we have nine features, and therefore nine scaling factors. For the reranking pass, we have an additional five features, and therefore need to optimize fourteen scaling factors. The scaling factors are optimized by minimum error training, to maximize the BLEU score on a tuning set, using an n-best-based simplex search. Selecting the right tuning set is very important for system performance. In this work, we use Dev07, a subset of LDC defined GALE Phase 3 dev data, and a portion of LDC GALE P3R1 Release for parameter tuning.

4.1. Genre-dependent Optimization

Broadcast news and broadcast conversations have many different characteristics. A major portion of BN is collected from anchors reading prepared material, which is typically grammatical, and has relatively rich vocabulary and long sentences. BC, mostly collected from recorded talk shows, mainly contains spontaneous conversational speech, with many ungrammatical sentences filled with disfluencies. However, the vocabulary is typically small, and sentences are usually shorter than with BN. The state-of-the-art ASR technique exhibits very different performance in documents of these two genres. The error rate on BC is usually more than twice as high as on BN.

Because of the difference between BN and BC, it is natural to consider genre-specific scaling factors, which probably will fit each genre better than the genre-independent factors. On the other hand, genre-specific optimization is more likely to lead to overfitting, and to hurt performance in unseen test sets.

Table 3. Comparing genre-dependent (GI) and genre-independent (GD) optimization on manual transcription. Results reported in case-sensitive BLEU(%)

		Eval06 (blind test set)			
Languag e	Genre	BN	BC	NW	WT
Arabic	GI	22.6	22.4	28.4	16.0
	GD	22.6	22.3	28.9	16.7
Chinese	GI	17.2	16.3	18.4	15.2
	GD	17.4	16.3	18.7	15.6

Table 4. Comparing optimization on truth vs. ASR output. Results reported in case-sensitive BLEU(%). M means optimizing on manual transcription; A means optimizing on ASR output.

Lang	Genre	Eval06		LDC-test		SRI-test	
		M	A	M	A	M	A
Arab	BN	22.6	22.4	30.4	31.6	30.5	31.2
Chin	BN	15.5	15.5	18.5	19.0	19.5	19.4
	BC	13.3	13.2	15.7	15.9	13.1	13.2

Table 3 lists results of using manual transcriptions as tuning data for optimization. The results are reported in the blind test set Eval06. For comparison, results on text genres, newswire (NW) and web text (WT) of the same test set are also shown. As we can see, there is very little difference between the two approaches for speech genres. This result is somewhat different from optimizing text genres, for which we see bigger and more consistent improvement.

4.2. Manual Transcription versus ASR Output

The above experiment used manual transcriptions for optimization, and was also tested on manual transcriptions, which do not have noise from ASR. For speech-to-text translation, we need to face ASR errors, especially for the BC genre. It is imaginable that the optimal scaling factors may vary along with the level of error rate. It may be beneficial to directly optimize on ASR output (ASR optimization) instead of the manual transcriptions (truth optimization). However, it is not completely clear if the optimized scaling factors generalize well on new test sets, as ASR errors can be totally random.

Table 4 shows the results of an experiment in which we compare optimizing reranking scaling factors on manual transcriptions versus ASR output for transcribed speech translation. The optimization was performed on genre-specific tuning sets, and we report results on three different test sets, namely, Eval06, LDC-test, and SRI-test, which are described in Section 5. For optimizing on ASR, the tuning sets used LDC provided manual sentence segmentation to obtain one-to-one correspondence between sentences and references.

We see mixed results in Table 4. In general, the difference between the two approaches is small; except for some test sets in BN, there are over 0.5 absolute improvements. We finally chose to use ASR-optimized weights for BN and truth-optimized weights for BC. One of the reasons of the decision is ASR error rate of BC is high and tuning on ASR output may be not very reliable.

5. Experimental Results

As the recent GALE retest evaluation did not include Arabic BC genre, we report only Arabic BN results for the entire system.

Table 5. Reference sentence / word counts of Eval06, LDC-test and SRI-test

	Eval06	LDC-test	SRI-test
A2E BN	956/18199	320/9467	226/7029
C2E BN	518/15759	127/3381	548/11738
C2E BC	979/15273	191/3866	476/6631

Table 6. First-pass / second-pass case-insensitive BLEU score of Eval06, LDC-test and SRI-test.

	Eval06	LDC-test	SRI-test
A2E-BN-truth	22.5/23.6	30.0/31.3	30.5/30.6
A2E-BN-asr	22.0/22.6	28.7/30.4	29.7/30.5
C2E-BN-truth	16.7/17.4	18.8/19.5	19.1/20.3
C2E-BN-asr	14.9/15.5	18.0/19.0	18.2/19.4
C2E-BC-truth	16.6/16.7	20.0/20.2	16.1/16.6
C2E-BC-asr	12.7/13.3	16.0/15.7	12.5/13.1

Table 5 shows statistics of the test sets used in the experiments. Among the three LDC-test is the unused subset of LDC P3 dev data, and SRI-test is a reserved portion of LDC GALE P3R1 release.

Table 6 summarizes results from the first-pass decoding and final reranking. Part of results are already shown in Table 4, but copied here for completeness. As we can see, language model reranking did bring consistent improvement in most test sets. We also noticed that there is relatively small difference between translating manual transcriptions of Arabic BN and ASR outputs, in the range of 0.1% ~ 1.0% absolute BLEU score. The difference is larger for Chinese BN, around 0.5% ~ 1.9%, and quite significant for Chinese BC, 3.4% ~ 4.5%. This indicates that we need to improve ASR quality in BC significantly in order to further improve final translation quality.

6. Conclusion

We have presented our work of building speech-to-text translation systems with training data mostly from text genres, including number tokenization, punctuation compensation, and parameter optimization. The final system results show that in comparing the translation of manual transcription and ASR output, Arabic BN has the smallest degradation, while Chinese BC has the largest degradation and is the genre needing the most future work.

7. Acknowledgment

This work is supported by the DARPA GALE program under contract HR0011-06-C-0023. We thank our SRI colleagues, our collaborators from RWTH Aachen University, University of Washington, Columbia University and International Computer Science Institute for sharing tools and data, and providing help in the system development and evaluation.

8. References

[1] Zheng, J., “SRInterp: SRI’s Scalable Multipurpose SMT Engine,” *Technical Report,* June 2008, http://www.speech.sri.com/projects/translation/srinterp.pdf

[2] Chiang, D., “Hierarchical phrase-based translation,” *Computational Linguistics* 33(2):201–228, 2007

[3] Och, F., and Ney, H., “Discriminative training and maximum entropy models for statistical machine translation,” in *Proc. ACL2002*

[4] Papineni, K., Roukos, S., Ward, T., and Zhu, W.J., “BLEU: A method for automatic evaluation of machine translation,” in *Proc. ACL 2002.*

[5] Och, F., “Minimum error rate training in statistical machine translation,” in *Proc. ACL 2003.*

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

[7] Jelinek, F., *Statistical Methods for Speech Recognition, Cambridge*, MA: MIT Press, 1997.

[8] Chen, S.F., and Goodman, J., “An empirical study of smoothing techniques for language modeling,” *Computer Speech and Language*, 13:359-394, 1999.

[9] Wang, W., Stolcke, A., and Zheng, J., “Reranking machine translation hypotheses with structured and Web-based language models,” in Proc. ASRU 2007, Kyoto.

[10] Matusov, E., Hillard, D., Magimai-Doss, M., Hakkani-Tur, D., Ostendorf, M., and Ney, H., “Improving speech translation with automatic boundary prediction,” in *Proc INTERSPEECH2007.*

[11] Bender, O., Matusov, E., Hahn, S., Hasan, S., Khadivi, S., and Ney, H., “The RWTH Arabic-to-English spoken language translation system,” in *Proc. INTERSPEECH 2007.*