RECENT ADVANCES IN SRI'S IRAQCOMM™ IRAQI ARABIC-ENGLISH SPEECH-TO-SPEECH TRANSLATION SYSTEM

Murat Akbacak[1], Horacio Franco[1], Michael Frandsen[1], Saša Hasan[2], Huda Jameel[1], Andreas Kathol[1], Shahram Khadivi[2], Xin Lei[1], Arindam Mandal[1], Saab Mansour[2], Kristin Precoda[1], Colleen Richey[1], Dimitra Vergyri[1], Wen Wang[1], Mei Yang[3], Jing Zheng[1]

[1]SRI International, Menlo Park, CA 94025 USA
[2]RWTH Aachen, D-52056 Aachen, Germany
[3]University of Washington, Seattle, WA 98195 USA

ABSTRACT

We summarize recent progress on SRI's IraqComm™ Iraqi Arabic-English two-way speech-to-speech translation system. In the past year we made substantial developments in our speech recognition and machine translation technology, leading to significant improvements in both accuracy and speed of the IraqComm system. On the 2008 NIST-evaluation dataset our two-way speech-to-text (S2T) system achieved 6% to 8% absolute improvement in BLEU in both directions, compared to our previous year system [1].

Index Terms— speech translation, spoken language translation system

1. INTRODUCTION

The IraqComm™ translation system [1] has been developed primarily under the DARPA Spoken Language Communication and Translation System for Tactical Use (TRANSTAC) program. IraqComm mediates and translates spontaneous, real-time conversations between an English speaker and a speaker of colloquial Iraqi Arabic. It is trained to handle topics of tactical importance, including force protection, checkpoint operations, civil affairs, basic medical interviews, and training. The system runs on standard Windows computers and can be used in a variety of modes, including an eyes-free, nearly hands-free mode. As the system is intended for use in interactive bilingual conversations, it must provide accurate translations, quickly enough so as to minimally impact the human-human interaction. An initial description of the software and hardware configurations can be found in [1]. The present paper details advances in the accuracy and speed of speech recognition and machine translation components.

2. ADVANCES IN SPEECH RECOGNITION

Two automatic speech recognition (ASR) systems are needed for the IraqComm system, one for English and one for Iraqi Arabic, each using SRI's Dynaspeak® speech recognizer. Here we describe improvements subsequent to [1]. These are due to more data, improved data processing, better combination of in-domain (ID) and out-of-domain (OOD) data (for English), hierarchical language modeling, feature space discriminative training, online speaker adaptation, end-pointing and faster 4-gram rescoring of first-pass word lattices.

2.1. End-pointing

Imperfect use of the hold-to-talk mechanism may introduce unnecessary silence at the beginning or end of an utterance. Removing this silence before ASR decoding can lead to faster ASR performance by reducing the number of frames to decode. We performed end-pointing at the beginning and end of each utterance using a two-word (speech and nonspeech) hidden Markov model (HMM) recognizer. The use of end-pointing resulted in ASR decoding (in terms of real-time factors) as much as 16% faster relative for both English (e.g., from 0.663 to 0.561 on the July '07 live test set) and Iraqi Arabic (e.g., from 0.759 to 0.609 on the July '07 live test set) with no loss in ASR accuracy.

2.2. Lattice Rescoring Speedup

IraqComm's ASR uses a two-pass decoding strategy with a first-pass time-synchronous Viterbi search using a small bigram language model (LM) followed by a second lattice rescoring pass using a large 4-gram LM to extract the final hypotheses. This approach allows us to use a large LM to obtain high accuracy without significantly increasing decoding time, if the second pass is very fast.

The previous implementation of lattice rescoring uses SRILM's [2] expansion algorithm to expand the first-pass lattice into a larger one with each node having distinct LM context and LM scores being assigned to arcs. Then a shortest-path algorithm can be applied to the expanded lattice to obtain rescored output. However, with a high-order LM, the expansion algorithm can be very slow for a high-density lattice, as may result from noisy input speech. In IraqComm, only the top or top few hypotheses are needed, so full lattice expansion is often unnecessary and traversal of a small portion of the lattice is adequate. Inspired by this observation we implemented an A* beam-search algorithm, which expands only the lattice nodes with forward-backward scores inside the search beam. Furthermore, using a priority queue we first expand the nodes in the best path with the lowest forward-backward scores, which are then updated according to the new LM scores. This algorithm successfully minimizes unnecessary node expansion and achieves substantial speed improvement without accuracy loss. In some cases, the speedup can be more than an order of magnitude. We also extended this algorithm to extract the top n-best hypotheses, and the implementation has been released in SRILM.

2.3. Feature Space Minimum Phone Error Training

Feature space MPE (fMPE) [3] is a discriminative feature space transform technique that modifies speech feature vectors x_t during both training and decoding:

$$y_t = x_t + \mathrm{M}\, h_t,$$

where x_t is the original feature vector, h_t is a high-dimensional posterior probability vector, and M is a matrix mapping the

posterior vector onto a lower-dimensional feature space. The projection matrix M is trained to optimize the MPE criterion.

The computation of the posterior vector h_t is expensive since it requires the likelihood computation of a large number of Gaussians (i.e., all the Gaussians in the acoustic model). Gaussian selection is typically used to first cluster all the Gaussians into indexing clusters, and then only evaluate the likelihood of the Gaussians in the highest-likelihood clusters. To further speed up the likelihood computation, we adopt Gaussian shortlists. The acoustic space is partitioned into a number of vector quantization (VQ) regions. The VQ codebooks are organized as a tree to quickly locate the VQ region in which a given input feature vector falls. Each VQ region is associated with a list of Gaussians that have high likelihood scores for training samples. During decoding, for each frame, a VQ region is first located according to the feature vector of that frame. Then only the Gaussians in the corresponding list are evaluated and others are ignored. Thus, a significant part of Gaussian computation is avoided. This technique was first proposed in [4]. For fMPE posterior vector computation, we built shortlists for both the indexing Gaussians and the Gaussians in the clusters, resulting in a two-layer shortlist.

By combining Gaussian selection and two-layer shortlists, the final fMPE computation is reduced to around 5% of the decoding time, with minor or no degradation relative to the accuracy without shortlists.

2.4. Improved Language Models

2.4.1 Iraqi Arabic Hierarchical Class LM (HCLM)

To improve language model performance for Iraqi Arabic (IA), we leveraged a language modeling technique previously developed for our Pashto-English translation system. The probability estimation of low-frequency and unseen n-grams is inherently difficult, and data sparsity is more serious for morphologically rich languages such as Iraqi Arabic. Such languages have a high vocabulary growth rate, which results in high language model perplexity and a large number of out-of-vocabulary (OOV) words. Word clustering is useful in that statistics on classes can replace statistics on individual words when the latter are unavailable or unreliable. A traditional class-based language model is built by partitioning the vocabulary into classes and approximating transition probabilities from word to word with transition probabilities from class to class. [5] advanced the approach of hierarchically clustering the vocabulary into a word tree in which the root node represents the entire vocabulary and a leaf node represents a single word. When estimating the conditional probability of a word based on its n-gram prefix, a hierarchical back-off strategy first backs off to its context with the most distant word replaced by its class, from the most specific to the most general (i.e., traversing the tree bottom-up). If none of these back-offs provides a minimum number of occurrences, it backs off to the normal lower-order (n-1)-gram prefix. It is thus likely to achieve more accurate n-gram estimation, in particular for unseen words. We further improved this hierarchical clustering class LM with part-of-speech (POS) information. In this refinement, we automatically generated a class tree for each POS tag. For this IA ASR task, a lexicon including POS information for words is available. As in [5], we generated a class tree for words with unknown POS tags in the decoding vocabulary. The hierarchical clustered class LM without POS information gave a relative perplexity reduction of 8% over the baseline n-gram LM, and the variant growing a cluster tree for each POS tag and employing this information for n-gram probability estimation yields about a 20% relative perplexity reduction compared to the baseline n-gram LM.

2.4.2 English LM

We improved the English ASR LM slightly by conducting data selection on the OOD data before training the OOD LM to interpolate with an ID-trained LM. We investigated two approaches, namely, perplexity based, using the ID-trained LM, and n-gram-hit based, to use ID n-grams directly for selecting more relevant OOD data. With the n-gram-hit-based approach to data selection, using 50% of the full OOD data achieved a 3% relative improvement on perplexity on the OOD-trained LM, compared to using all OOD data.

2.5. Improved Acoustic Models

All models use a 16 kHz front end with 10 ms frame advance rate, and Mel frequency cepstral coefficients with 1st, 2nd, and 3rd order derivatives. Heteroscedastic linear discriminant analysis (HLDA) reduces the dimensionality of the feature vector to 39. Decision tree state-clustered tri-phone models are used for both English and Iraqi Arabic.

All the ID data available for this task was collected under the TRANSTAC program, and currently represents about 110 hours of English and 507 hours of Iraqi Arabic speech and text data. As new training data becomes available, our data processing regime allows us to quickly implement global processing changes and investigate their effects. We also use automatic measures to detect potentially problematic data (e.g., using character-based perplexity measures to flag words with unusual character sequences, running the latest system on the new data to check transcription/translation quality).

2.5.1 English AMs

In the past we trained the English acoustic models mostly on OOD data (using approximately 200 hours of clean data from WSJ, ATIS and Broadcast News), to some of which we also added noise for acoustic noise robustness. The OOD-trained models were then adapted to a small amount (about 10 hours) of ID data. The results from these earlier models are shown in line 1 of Table 1. These were also the models used for last year's July-07 NIST evaluation. It should also be noted that those earlier models were not using HLDA or fMPE.

Since we now have much more ID data, we experimented with merging ID and OOD data with different weights, for a better match to ID speakers. In lines 2-5 of Table 1 we compare the performance of acoustic models (all with HLDA+MPE training and with 4-gram rescoring) for different combinations of ID and OOD training data, on several test sets. We observe that the OOD data performs quite well on the live testsets, but not so well in the offline testsets (which are a very close match to the ID training data). When MAP adapting OOD-trained models to ID data, the offline testset accuracy improves, but live performance degrades somewhat. The best performance is achieved when ID data is weighted (3 times) pooled with the OOD data to train the acoustic models with 128 Gaussians per triphone state.

We also observed that discriminative training was more robust when we trained with merged data, and the improvements we were getting with MPE and fMPE were larger. Line 6 in Table 1 shows the improvement achieved with the fMPE models.

	July'07 -offline	July'07-live	All offline	All live
1: Baseline	27.8	12/6	23.1	15.7
2: ID-only_data	17.7	11.2	19.5	14.1
3: OOD-only data	24.1	11.5	30.0	11.5
4: +ID-MAP-adapt	18.6	10.7	21.4	12.5
5: 3xID + OOD data	17.6	9.8	20.8	11.1
6: +fMPE	16.6	8.4	19.3	10.5
7: + newLM	16.1	8.5	19.0	10.4
8: +fMLLR	15.6	7.5	18.7	9.9

Table 1*: WER (%) results on different testsets*[1] *for different English models. Baseline (line 1) refers to the system we used for last year's July-07 NIST evaluation. Lines 2-5 compare the use of different data for AMs, while lines 6-8 show the effect of adding fMPE, new LM and fMLLR sequentially in the system trained with the best AM data combination (3xID + OOD).*

2.5.2 Iraqi Arabic AMs

Iraqi Arabic acoustic models were updated with the additional 65 hours of transcribed speech data released under the TRANSTAC program in the first half of 2008. Table 2 shows the improvements obtained from using this data, as well as the improvements resulting from employing fMPE-based acoustic modeling, speaker adaptation, and improved language modeling, on all previous live (about 1.8K utterances) and offline (about 3.8K utterances) testsets, as well as July'07 offline and live test sets,.

	July'07-offline	July'07-live	All Offline	All Live
Baseline	41.3	23.8	36.3	23.9
+new data	40.9	24.1	36.1	23.6
+fMPE	39.4	22.7	35.2	22.7
+fMLLR	39.3	21.8	33.9	22.3
+new LM	37.5	16.5	31.0	18.9

Table 2: *Improvements in WER (%) for Iraqi Arabic ASR.*

2.6. Online Speaker Adaptation

To improve DynaSpeak's accuracy under conditions of changing speakers and acoustic environments, we implemented online speaker adaptation based on feature space maximum likelihood linear regression (fMLLR), or constrained maximum likelihood linear regression (CMLLR) [6]. A feature-space affine transform is estimated incrementally and used to adapt the acoustic features for every utterance. The sufficient statistics required for estimation of the adaptation transformation are updated after every utterance within an IraqComm session. To handle the case of speakers changing within a session, the sufficient statistics are updated with a higher weight for more recent utterances. This procedure involves normalizing the sufficient statistics to a fixed window of 1000 frames with a decay weight for the statistics from the previous utterances. Overall, fMLLR produced significant word error rate improvements in ASR for both English and Iraqi Arabic for both offline and live conditions, as shown in Tables 1 and 2.

[1] Offline testsets include the EN and IA parts of pre-recorded human-to-human conversations, with an interpreter translating in both directions. The live testsets were recorded using the S2S translation system, without the help of a human interpreter. The July'07 testsets were the ones used for the NIST evaluations, including 567/561 offline utterances (for EN/IA respectively) and 808/696 online utterances. The all-online, all-offline are bigger testsets including previous NIST-eval data, but also locally selected utterances for a total of 4.5K,/3.8K offline and 2K/1.8K online EN/IA data.

3. ADVANCES IN MACHINE TRANSLATION

IraqComm uses SRI's own statistical translation engine, SRInterp™. SRInterp supports state-of-the-art statistical machine translation (SMT) technologies, including the standard phrase-based translation and hierarchical phrase-based translation. Inside IraqComm, in the Iraqi Arabic to English direction, SRInterp is the sole translation engine. For English to Iraqi Arabic, SRInterp runs in parallel with an interlingua-based translation engine, Gemini™, and provides translations when Gemini fails or is slow to generate output. The following sections describe major improvements to the translation components over the past year.

3.1. Increased Training Data and Improved Lexicon

A major contributor to translation improvement is additional data releases within the TRANSTAC program via the Linguistic Data Consortium (LDC), i.e., new parallel data with about 55K new sentences and an improved Iraqi Arabic lexicon (V5.3), which increased coverage significantly compared to V4.1, used in the 2007 system.

Iraqi Arabic is a highly inflected language, and word segmentation has proven effective in reducing vocabulary size and improving translation quality. We have developed a semi-supervised Iraqi Arabic word segmentation algorithm [7], which uses an initial set of segmentation rules, called *seed rules*, derived from a linguistic lexicon, and iteratively deduces and adds new rules for new words based on a predefined prefix and suffix list, a word stem list, and certain heuristics. As the Iraqi lexicon covers more distinct word forms the word segmentation quality has improved, as shown in Table 3, which also indicates the greater effect of the improved lexicon over additional data.

3.2. Word Alignment Training

We experimented with splitting the training data according to the original language of production and gave the two parts (audio transcript or translation) different weights in the word alignment training. The idea was to give more weight to the part of the data that matched the translation direction, e.g., more weight to Iraqi Arabic audio transcripts and the corresponding English translations into English for Iraqi Arabic to English translation, and vice versa for the other direction. This increased translation quality on one of our test sets only slightly, less than 0.1% BLEU (Bilingual Evaluation Understudy) absolute, which was not deemed significant. We therefore proceeded without this splitting. We also tested several symmetrization heuristics for combining bidirectional GIZA++ alignments, and selected the refined heuristic in [8] based on BLEU.

3.3. Hierarchical Phrase-based Translation

Hierarchical phrase-based translation uses synchronous context-free grammars (SCFGs) in the translation model. The standard phrase-based translation approach can model only local word order change effectively, while hierarchical phrase-based translation provides a more principled way to model long-distance word reordering, and has been very effective for language pairs with very different word orders, such as Chinese and English. However, Iraqi Arabic and English share similar syntactic structure and word orders, and only smaller gains from hierarchical modeling are

expected. In experiments with the same training data and word alignment, we obtained about 1.0% absolute BLEU improvement in the Iraqi Arabic to English direction, and about 0.6% in the other direction. Although the difference is small, because of the large size of the test set (17K sentences), it is statistically significant.

3.4. Speed and Memory Optimization

IraqComm is a real-time application running with limited CPU power and memory on a laptop or smaller platform, and therefore constraints are imposed upon model development. To make the translation engine run with sufficient speed and an acceptable memory footprint, we take several steps. Among these are the pre-computation of all phrase and rule scores with the scaling factors and storage of only the combined score; limiting the phrase length on both source and target sides; limiting the number of target phrases and rules corresponding to a single source phrase; storage of the phrase and rule table on disk in a binary format that allows fast retrieval so as to reduce memory use; and limiting the order and size of the language model. With these optimizations, the translation engine runs on a laptop without noticeable delay in most cases.

	2007 system lexicon v4.1	adding '08 data	using lexicon v5.3
IA→Eng	33.97	34.18	34.86
Eng→IA	20.56	20.70	23.73

Table 3*: BLEU (%) of various systems on a 17k-sentence test set*

4. COMBINED PERFORMANCE

Above we presented the improvements in ASR and machine translation (MT). Table 4 presents how these improvements reflect on speech-to-text (S2T) translation performance, by comparing the July '07 system with the June '08 system on the June '08 evaluation offline testset. This testset contains 651 English utterances and 579 Iraqi Arabic utterances. There are 4 translation references which are provided by NIST.

There is a big improvement (>15% absolute in WER) from all the developments we made on the English models..For Iraqi Arabic ASR, there is an overall improvement of 5.3% absolute in WER, similar to the overall improvement observed in Table 2.

	July '07 system	June '08 system
	WER (%)	
ASR_{EN}	30.0	14.4
ASR_{IA}	36.5	32.6
	BLEU(%)	
$S2T_{EN2IA}$	13.1	21.5
$S2T_{IA2EN}$	34.6	40.3

Table 4*: ASR (WER) and S2T (BLEU) of last year's and this year's systems on June'08 offline testset*.

Compared to last year's speech-to-text (S2T) system [1], we achieved 8.4% absolute BLEU improvement in English to Iraqi Arabic direction. In Iraqi Arabic to English direction, the absolute improvement is 5.7% BLEU score. These are substantial improvements for an S2T system.

5. FUTURE WORK

In addition to the improvements presented in this paper, we are currently working on further enhancements of the abilities of our system. ASR models can further be improved by adding cross-word triphones (we are currently using only within-word triphones which run faster for our system), and vowelized pronunciations for Arabic, which we have not incorporated yet since the current vowelized lexicon does not have a good vocabulary coverage. For MT we are working on better combination between statistical and rule-based components, and improving the name-translation capabilities. Moreover we are working on better integration of ASR/MT components and fine-tuning the whole system to better address the usability targets which include high task completion and concept transfer rates.

Finally we are looking towards generalizing our methods for rapid and robust porting of the technology to new language pairs.

6. CONCLUSIONS

IraqComm is a good example of how far spoken language translation technology has come in recent years. Generating accurate speech-to-text translations quickly is very important in real-life human-human interactions. We have presented advances in the accuracy and speed of speech recognition and machine translation components. We have shown that improvements in both ASR and MT accuracy lead to substantial improvements in S2T translation performance in both English to Iraqi Arabic and also Iraqi Arabic to English directions.

7. ACKNOWLEDGMENT

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) and the Department of Interior-National Business Center (DOI-NBC) under contract number NBCHD040058.

8. REFERENCES

[1] K. Precoda, J. Zheng, D. Vergyri, H. Franco, C. Richey, A. Kathol, and S. Kajarekar, "IraqComm: A next generation translation system," in *Proc. Interspeech*, pp. 2841-2844, 2007.

[2] A. Stolcke, "SRILM - An extensible language modeling toolkit," in *Proc. ICSLP*, vol. 2, pp. 901-904, Denver.

[3] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "FMPE: Discriminatively trained features for speech recognition," in *Proc. ICASSP*, pp. 961–964, 2005.

[4] V. Digalakis, P. Monaco, and H. Murveit, "Genones: Generalized mixture tying in continuous hidden Markov model-based speech recognizers," *IEEE Trans. Speech and Audio Processing* 4(4), 281-289, July 1996.

[5] W. Wang and D. Vergyri, "The use of word n-grams and parts of speech for hierarchical cluster language modeling," in *Proc. ICASSP*, Toulouse, France, May 2006.

[6] M.J.F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech & Language*, vol. 12, pp. 75-98, 1998

[7] M. Yang, J. Zheng and A. Kathol, "A semi-supervised learning approach for morpheme segmentation for an Arabic dialect," in *Proc. Interspeech*, 2007.

[8] F.J. Och and H. Ney, "Improved statistical alignment models," in *Proc. ACL*, 2000.