A Hybrid Phrase-based/Statistical Speech Translation System

David Stallard, Fred Choi, Kriste Krstovski, Prem Natarajan, Rohit Prasad, Shirin Saleem

BBN Technologies
Cambridge, MA, USA
stallard@bbn.com

Abstract

Spoken communication across a language barrier is of increasing importance in both civilian and military applications. In this paper, we present a system for task-directed 2-way communication between speakers of English and Iraqi colloquial Arabic. The application domain of the system is force protection. The system supports translingual dialogue in areas that include municipal services surveys, detainee screening, and descriptions of people, houses, vehicles, etc. N-gram speech recognition is used to recognize both English and Arabic speech. The system uses a combination of a pre-recorded questions and statistical machine translation with speech synthesis to translate the recognition output.

Index Terms: speech-to-speech translation, Iraqi Arabic

1. Introduction

Military and humanitarian personnel often need to communicate with residents of a host country who do not speak English. In a crisis situation, there will be little time to train personnel in the host country language, and human interpreters will often be in short supply. Portable devices for speech-to-speech language translation would therefore be very useful in such environments.

In this paper, we present a system that is being developed to address this need. The system facilitates communication between an English speaker and a speaker of Iraqi colloquial Arabic. Colloquial Iraqi Arabic is the language spoken by ordinary people in Iraq, and differs considerably from the Modern Standard Arabic (MSA) that is used in writing and broadcast news. In particular, the pronunciations of words and sometimes even the words themselves are different between the two, making colloquial Arabic in effect a low-resource language. While the algorithms it uses are general, the system is specifically targeted towards the military domain of force protection, and is intended to support limited translingual dialogues for situations such as checkpoints and municipal services surveys, as well as questions about people, buildings, vehicles, and the like.

The system is being developed as part of DARPA's TRANSTAC program. The systems being developed in this program are broadly classified as being either "1.5 way" or "2 way". The 2-way systems seek, in principle, to translate any utterance, in either direction, typically by using broad-coverage statistical machine translation (SMT) components trained on large parallel corpora. Examples are the systems developed by IBM [1], SRI [2], and CMU [3]. The 1.5-way systems, by contrast, use a task-directed approach to make the communication problem easier, by specifying a fixed set of English questions with pre-recorded foreign-language translations, together with a constrained set of foreign-language answers that can be translated into English. An example question would be "How old is he?" or "Is the roof of the house tiled or concrete". Example systems are a system earlier developed by BBN [4] and a system developed by Sehda [5].

In spite of its more limited coverage, the 1.5-way approach has important advantages. Among these is the advantage of clarity. Specifically, the English-speaking user always knows exactly what the system has understood him to have said, and knows that this meaning was conveyed accurately, fluently, and intelligibly to the Arabic-speaking respondent. By contrast, in a statistically-based 2-way system, the user cannot know for certain whether the system has properly conveyed his meaning, since errors can always arise in the translation process, even if the speech recognition was error-free. Moreover, the 2-way system, by virtue of its generality, depends on synthesized speech, which may not be as intelligible as natural speech, especially for low-resource languages like colloquial Arabics.

Our system is designed to be a hybrid between the two approaches, combining the clarity of the 1.5-way for utterances that correspond to recorded questions, with the coverage of the 2-way for utterances that do not. The system currently runs on a Windows laptop, and a port to an integer-only Windows CE handheld is mostly complete (except for the SMT engine).

In what follows, we first present the overall architecture of our system, followed by a discussion of speech recognition and translation. We conclude with preliminary evaluation results, and a discussion of future work.

2. Design and Architecture

Figure 1 shows a block diagram of the system. English speech is first passed to the English speech recognition component, which is BBN's Byblos system [6]. The resulting English

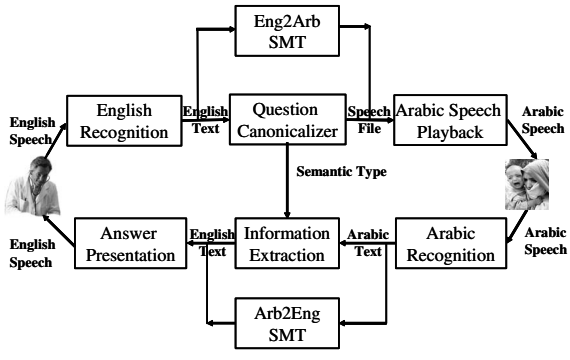

Figure 1: System Architecture

output text is then passed to two different translation components. The first, the Question Canonicalizer, tries to map the text into a canonical utterance for which the system has a recorded audio translation. If a single such utterance is found, it is returned, and the recorded audio for it played out to the Arabic speaker. If no canonical utterance is found, or if more than one canonical utterance is found, the English text is passed to BBN's SMT component, which maps the text into Iraqi Arabic text. In this case, the Arabic text is passed to a speech synthesizer, which synthesizes Arabic speech and plays it out to the Arabic speaker.

The Arabic speaker responds, and his speech is passed to the Arabic speech recognizer (also Byblos), which turns the Arabic speech into Arabic text. The Arabic text is then passed to two alternative translation components. The first is an information extraction-based translation method, which looks for information that is directly relevant to the question asked. The second is an SMT component. The results of the two are combined, and passed to the Cepstral Iraqi speech synthesizer, which turns the resulting text into English speech.

The complete system runs on a Windows laptop with a graphical user interface which is intended to be operated by the English speaker. This interface is shown in Figure 2. It contains separate "listen" buttons to tell the system to listen for Arabic and English speech, respectively. These buttons operate on a push-and-hold basis, in which the operator clicks the button and holds it down until the speech is complete. Using push-and-hold rather than click-to-talk is more robust than automatically detecting the end of speech, and seems to be more intuitive and familiar for military users.

The English speaker can override the result of the Question Canonicalizer and have the SMT translate his question verbatim by clicking in the text window and pressing "enter" on the keyboard. The system can also operate in regular 2-way mode in which the canonicalizer is bypassed by default.

3. Speech Recognition

3.1. Iraqi Speech Recognition

The BBN Byblos speech recognition system [6] models speech as the output of context dependent phonetic Hidden Markov Models (HMMs). The outputs of the HMM states are mixtures of multi-dimensional diagonal Gaussians. Different forms of parameter tying are used in Byblos. In a State-Tied Mixture (STM) model, all triphones with a certain center phoneme and a state position share the same set of Gaussians (512 on average). In a State Clustered Tied Mixture (SCTM) model, states which are automatically clustered according to quinphone context share the same set of Gaussians (64 on average). The mixture weights in both these cases are shared based on a linguistically guided decision tree.

The baseline acoustic models were trained using the Maximum Likelihood framework. We further improved the models by training discriminative models using the lattice-based Maximum Mutual Information (MMI) estimation [8]. The baseline ML models were used as the initial estimate for discriminative training.

Decoding is done in 2 passes as described in [9]. The forward pass is a fast match beam search using an STM model and a bigram language model. The backward pass operates on

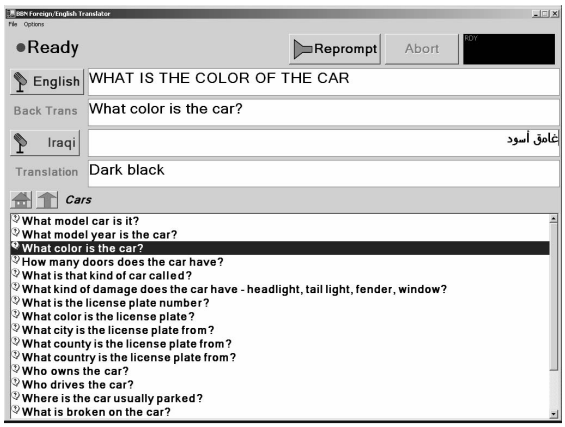

Figure 2: User Interface

the output of the forward pass using the more detailed SCTM quinphone model and a trigram language model to produce the best hypothesis. Shortlists of Gaussians which occur in a certain region of the acoustic feature space are pre-computed to speed up Gaussian computation [10]. In addition, the means and variances of the Gaussians are also quantized.

The acoustic model is trained on about 196 hours of Iraqi Arabic speech collected under the TRANSTAC effort. This consists of 1.5-way and 2-way data of colloquial Iraqi speech from different domains such as Medical/Refugee, Force Protection, Civic Amenities, Detainee Screening, etc. Language modeling data from different domains was interpolated with the interpolation weights tuned on a held-out development set. The language model employs Kneser-Ney smoothing and was trained on about 1.5 million words. The dictionary size is about 52K.

Configuration	%WER
Maximum Likelihood	38.10%
Maximum Mutual Information	33.70%

Table 1. Improvement in WER on held-out Iraqi Arabic test set using different acoustic model estimation methods

Table 1 shows the results on a held-out dev-test set of approximately 9 hours of speech. We get 11.5% relative improvement in word error rate (WER) over the baseline ML models by using the MMI models. Note that all forms of the glottal stop or "hamza" have been normalized for WER computation. The Arabic decoder runs in less than real time on a 2.8GHz processor.

Table 2 shows the results on the 1.5-way and 2-way test sets of the March 2006 Offline Evaluation of the TRANSTAC program.. The acoustic models were adapted for individual

Decoding	%WER	
	1-5way Offline	2-way Offline
Unadapted	28.7%	28.9%
Adapted	23.0%	22.6%

Table 2. Results on TRANSTAC March 2006 Offline Iraqi Arabic test sets

speakers. We got a 20% relative gain in WER with adaptation. Standard orthographic normalizations were applied for WER computation.

3.2. English Speech Recognition

The English decoder has the same configuration as the Arabic system except that it uses the STM model in both the forward and backward pass. The English acoustic model is trained on 36 hours of speech. The language model is an interpolation of 1 million words of in-domain data, and 49 million words of out-of-domain data. The dictionary size is 8K words. The results on the held-out test set of 2 hours of speech are shown in Table 3. Similar to the case of Iraqi Arabic, MMI models give a gain of 11% relative over the ML models. The speed of the English decoder is 0.52xRT on a 2.8GHz processor.

Estimation	%WER
Maximum Likelihood	29.80%
Maximum Mutual Information	26.50%

Table 3. Improvement in WER on held-out English dev-test set using different acoustic model estimation methods.

The results on the March 2006 Offline Evaluation test sets are shown in Table 4. Unsupervised adaptation gave a gain of 6% absolute on the 2-way offline set.

Decoding	%WER	
	1.5-way Offline	2-way Offline
Unadapted	12.00%	12.50%
Adapted	5.80%	6.50%

Table 4. Results on TRANSTAC March 2006 Offline English test sets.

4. Translation

4.1. Overview

To translate recognition output from English to Arabic, our system first tries to map the output to one of the approximately 700 English utterances for which it has a pre-recorded translation into Arabic. If it is unable to do so unambiguously, it translates the utterance via its SMT component.

In the Arabic-to-English direction, the system applies both SMT and a concept translation component that is described in [4]. The concept translation component looks for information that is directly relevant to the question being asked (e.g. an age amount in response to "How old is he?"), while the SMT attempts to translate the entire utterance. The two outputs are concatenated unless one is a substring of the other, in which case only the larger of the two is presented.

4.2. Utterance Canonicalization

To translate speech recognition output from English to Iraqi, two different modules are used. The first is the English Canonicalizer, which tries to map the recognition output onto one of the approximately 700 canonical utterances for which the system has a pre-recorded audio translation. The canonicalizer was developed using a corpus of approximately 70,000 variants of the canonical utterances. (Both the canonical utterances and their variants were generated by experts in the application domain). Variants not only differ in wording, but may also differ in underlying concepts. For example, "How high is the house?" and "How many stories does the house have?" are both considered variants of "Is it a one, two, or three story house?". (Note that word-based metrics such as BLEU score are thus not valid for evaluating this procedure.)

We have experimented with both a probabilistic semantic RTN parser and automatically generated rules to solve this problem. Here we report on the probabilistic parser approach. The parser is similar to the HUM parser [7], except that all probability computations have been integerized to allow the system to run on handheld computers with integer-only hardware. The parser was trained on an annotated version of the corpus of variants. In particular, each variant was annotated with a simple tree, whose root label was the high-level sub-domain (e.g. house questions, car questions, etc) and whose pre-terminal labels were base concepts like 'NAME', 'HEIGHT', etc.

The output of the parser is treated as an N-dimensional Boolean concept vector, where N is the number of concepts in the system. This vector is matched against a database of concept vectors derived from the annotated variants. In particular, the system searches for the vector with the least distance to the input vector, where this distance is weighted by the inverse document frequency of each concept. Thus a concept is weighted more strongly if it is associated with fewer questions. The match is rejected if no vector can be found within a maximum allowed distance.

In an offline test of 172 1.5-way utterances conducted under the TRANSTAC program, the canonicalization component achieved error rates of 1.3% on transcription, and 6.9% on ASR output.

4.3. Statistical Machine Translation

If the canonicalizer cannot find a question to map to, or if more than one match is found with the same score, the system uses the SMT component to generate a verbatim translation into Arabic. BBN's SMT engine was originally developed for translation of text from Modern Standard Arabic (MSA) and Mandarin into English. As part of this work, we applied it to translation of recognition output from colloquial Iraqi Arabic into English, and from English to Iraqi Arabic.

The SMT component uses a phrase-based approach to translation [11]. In particular, given an input foreign language sentence 'f', we estimate the most likely translation into the target sentence 'e' as

$$\hat{e} = \arg\max_{e} P(e \mid f)$$

Word alignments between source-target sentence pairs are first generated. Phrase pairs are extracted from the word alignments by merging nearby alignment groups using a set of rules. Phrase pair statistics can be automatically extracted from word aligned corpora.

The system tries to find the most likely target sentence among all possible segmentations of the source sentence into

phrases, all possible phrase reorderings, and all possible translations of the source phrases into target phrases. This is determined using a log-linear interpolation of statistical models. The parameters of the models are estimated using the statistics of the phrase pairs extracted from the word alignments, and the interpolation weights are optimized by trying to minimize the translation errors on a development set.

The data used to develop the translation models was collected under the TRANSTAC program. All forms of the hamza (a glottal stop) were normalized during training, and decoding. Table 5 shows the 4-reference BLEU score for Arabic-to-English (A2E) translation on a blind test set, the 240-utterance TRANSTAC December 2005 2-way Offline Evaluation. Statistics are given for both the December 05 and March 06 configurations, including the number of parallel utterances used to train the translation model and number of words used to develop the English target LM.. For Dec-05 the target LM was developed from the English half of the parallel corpus, whereas for Mar-06 the target LM was developed from the ASR training set A 26% relative gain in speech output BLEU is observed with the larger training set (the larger LM gave only a minimal gain).

System	Utt Pairs	LM Words	Text	Speech Output	
			BLEU	WER	BLEU
Dec-05	172K	846K	0.54	27.5	0.43
Mar-06	236K	48.5M	0.61	21.3	0.54

Table 5. Arabic-to-English SMT performance on the TRANSTAC December 2005 2-way offline test set

Table 6 shows the English-to-Arabic (E2A) results on the December Offline 2-way test set, where we again show a comparison between the Dec-05 and Mar-06 configurations. Only about 13K E2A utterances were available for training, so these were augmented with the reverse of the A2E data above.

The performance on E2A is notably poorer than for A2E. Two reasons can be posited for this. One is simply that less data was available for E2A. Although 236K A2E pairs were available for reversal, most were simple Arabic answers to questions (e.g. "near the mosque"), and thus not well matched to the English inputs. Another possible reason is the morphological complexity of the Arabic language itself, in which elements that are words in English (e.g. conjunctions, prepositions) appear in Arabic as affixes attached to words. As a result, the vocabulary size of Arabic is 3 times larger than that of English. Thus, a large amount of data may be needed to learn the alignments from English to Arabic.

System	Utt Pairs	LM Words	Text	Speech Output	
			BLEU	WER	BLEU
Dec-05	184K	854K	0.32	12.87	0.22
Mar-06	250K	2360K	0.43	9.90	0.36

Table 6. English-to-Arabic SMT performance on the TRANSTAC December 2005 2-way offline test set

5. Current Status and Future Work

The current system runs on a Windows laptop computer, and is undergoing further improvement and testing. With the exception of the SMT component, all modules of the system have also been ported to an integer-only Windows CE handheld. In the near future, we plan to port the SMT component to this environment as well.

6. Acknowledgements

This work reported here was done under the DARPA TRANSTAC Program. We would like to thank Jinxi Xu and Mike Kayser of BBN for help in applying their machine translation component to this effort. The statistical machine translation engine used here was developed under a separate BBN effort. We would also like to thank Luise Malloy and John McCary at DARPA for developing the canonical questions and variants for the TRANSTAC program.

7. References

[1] Zhou, B., Chen, S., and Gao, Y., "Constrained Phrase-Based Translation Using Weighted Finite-State Transducers". Proc ICASSP 2005, Philadelphia, PA. March 2005 pp. 1017-1020

[2] Kathol, A., Precoda, K., Vergyri, D., Wang, W., and Riehemann, S., "Speech translation for low-resource languages: the case of Pashto", Proc. INTERSPEECH-2005, 2273-2276.

[3] Schulz, T., and Black, A., "Challenges with Rapid Adaptation of Speech Translation Systems to New Language Pairs" Proc. ICASSP 2006, Toulouse, France, May 2006 (to appear)

[4] Stallard, D., Makhoul, J., Choi, F., Macrostie, E., Natarajan, P., Schwartz, R., and Zawaydeh, B., "Design and Evaluation of Limited Two-Way Speech Translation". Proc. Eurospeech 2003, Sept. 2003. 2221-2225

[5] Ehsani, F., Kimzey, J., Master, D., Hunil, P., Sudre, and K., "Rapid Development of a Speech Translation System for Korean", Proc. ICASSP 2006, Toulouse, France, May 2006 (to appear).

[6] Prasad R., et al., "The 2004 BBN/LIMSI English Conversational Telephone Speech Recognition System," Proc. EUROSPEECH, ISCA, Lisbon, Portugal, Sept. 2005.

[7] Miller, S., Stallard, D., Bobrow, R., and Schwartz, R., "A fully statistical approach to natural language interfaces". In Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, pp. 55--61 Santa Cruz, CA. June 1996

[8] Woodland, P. and Povey, D., "Large Scale Discriminative Training of Hidden Markov Models for Speech Recognition," Computer Speech and Language, Vol. 16, pp. 25-47, 2002.

[9] L. Nguyen and R. Schwartz, "Efficient 2-pass N-best Decoder," Proc. EUROSPEECH, ISCA, Rhodes, Greece, Sept. 1997.

[10] J. Davenport et al., "Towards a Robust Real-Time Decoder," Proc. ICASSP 1999, IEEE, Phoenix, AZ, March 1999.

[11] Koehn, P., Och, F., and Marcu D., "Statistical Phrase-Based Translation", NAACL/HLT 2003, Proc. NAACL/HLT, Edmonton, Canada, May 27--June 1 2003