Rapid Development of an English/Farsi Speech-to-Speech Translation System

Chia-lin Kao, Shirin Saleem, Rohit Prasad, Fred Choi, Prem Natarajan,
David Stallard, Kriste Krstovski, Matin Kamali

Presented at: IWSLT 2008, Hawaii

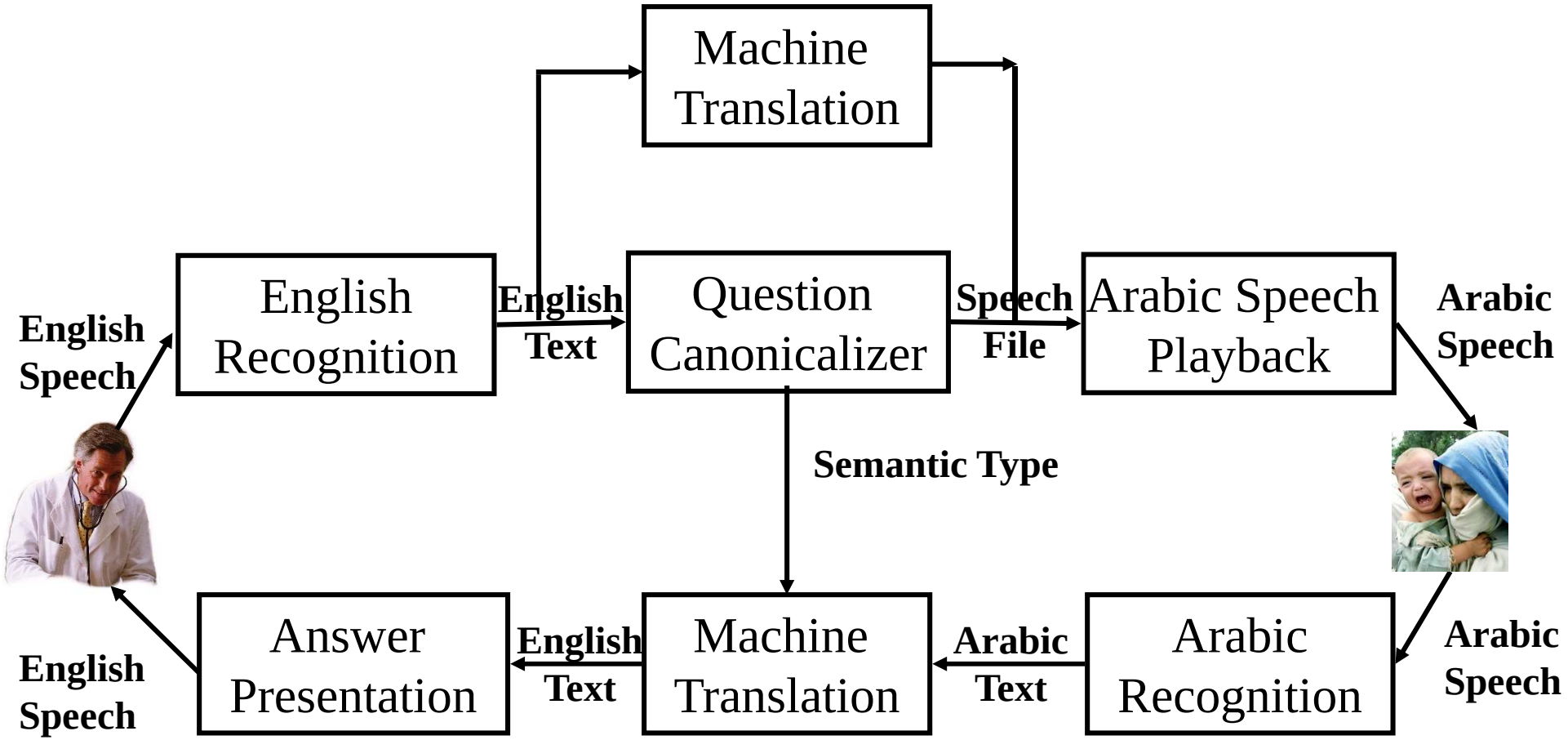
Outline

- Overview of Speech-to-Speech (S2S) Translation System
- DARPA TRANSTAC Surprise Language Evaluation Task
- Challenges in Configuring for English/Farsi (E/F)
- Expansion of Training Data
- Automatic Speech Recognition for E/F
- Machine Translation for E/F
- Error Analysis and Conclusions

BBN TransTalk: Speech-to-Speech Translator

- **Two-way speech-to-speech translation system for facilitating information exchange over a language barrier**
  - **Advanced under DARPA TRANSTAC program**
  - **TRANSTAC focuses on English/Iraqi language pair with emphasis on evaluating in a new language every year**
- **BBN Byblos speaker-independent speech recognizer**
  - **Context-dependent hidden Markov models (HMMs)**
  - **Multi-pass search**
- **Phrase-based statistical machine translation**
- **Question Canonicalizer for mapping frequently asked questions to recorded speech**
- **Eyes-free, one-handed control interface**

Block Diagram for English/Arabic Operation



DARPA TRANSTAC Surprise Language Evaluation Task

- 90-day constraint for completing an end-to-end system in a new language pair
- One shot chance to train the system on a new language that it has not previously been trained on
  - Little time for experimenting and conducting additional research
- Training data provided is a factor of 5 less than what the Iraqi system is trained on
  - Need to scale the system down to train with less data
- Additional data collection infeasible due to limited time
  - Can we efficiently reuse existing data and how?
- Lack of sufficient linguistic resources and phonetic information
  - Need to configure the system components to address linguistic differences without prior experience with the new language

Challenges in Farsi

- **Complex word structure and irregularity in orthography representation:**
  - **Formal or informal/colloquial endings**
  - **Multiple forms of compound words**
  - **Variances of plural and verb affixes in terms of different persons and tenses**
- **Little technical work has been done in Farsi, no baseline to compare to during development**
- **Significant number of homographs due to lack of short vowels Arabic script writing system**
- **Large number of ambiguous homophones due to multiple alphabets sharing the same pronunciation**
- **Lack of standardized normalization guidelines and romanization schemes**

Multiple Variants of Compound Words

Compound Word Form I: نمیتونم →I can't
سلام سلام آره من صداتون میشنوم نمیتونم ببینمت ولی صداتون میشنوم
Hi, hi ya. I hear you but I can't see you, but I hear you.
Compound Word Form II (contains zero width joiner): نمی تونم → I can't
به من کمک کن کمک کن نمی تونم بالا بیام که
Help me, help. I can't come up

Training Data Provided by NIST

- **Audio and associated transcriptions and translations provided from two different types of collections**
  - **1.5-way: answers to a fixed set of questions**
  - **2-way: role-played dialog between speakers**

Acoustic Training Data

Set	#Hours Farsi	#Hours English
Train	82.2	3.6
Dev	3.5	0.6
Test	3.4	0.6

Training Data Provided by NIST – cont'd

Parallel Sentence Pairs for MT

Farsi-to-English (F2E)				
#Sentence Pairs	#Farsi Words		#English Words	
	Total	Unique	Total	Unique
75K	537K	24.6K	602K	11.3K
3.9K	28K	4.7K	32K	2.8K
3.9K	27K	4.7K	30K	2.8K
English-to-Farsi (E2F)				
#Sentence Pairs	#Farsi Words		#English Words	
	Total	Unique	Total	Unique
31.3K	178K	14.1K	200K	8.2K
1.6K	9.5K	2.6K	11.2K	1.9K
1.6K	9.6K	2.7K	11.2K	1.9K

E2F training data above includes English data from DARPA Babylon CAST program that was translated into Farsi

Training Data Expansion

- **Harvested text data from the web based on n-gram queries for language modeling (LM)**
  - **44 million words of text for Farsi**
  - **12 million words of text for English**

- **Directed interactive data collection through the system**
  - **Native English speaker and a Farsi speaker were instructed to use the S2S translation system to role-play scenarios**
  - **Bilingual speaker reviewed and corrected the incorrect translations produced by the system**
  - **Collected a total of 1800 parallel sentence pairs, where the source sentence was either recognized or translated incorrectly**

Training Data Expansion – Reuse of E/I data

- **Motivation: Use data from existing collection in a different language but preferably from similar domain**
- **Generated English/Farsi (E/F) sentence pairs by translating English source sentences and translations from English/Iraqi (E/I) collection**
  - **Domain and conversational style is similar in both collections**
- **Selected 3000 English sentences from E/I collection that improve n-gram coverage on the E/F collection**
  - **Sentences with low out-of-vocabulary but high perplexity with respect to English LM trained from E/F collection**
  - **Named entities (NEs) were all mapped to a single token for selection because NEs are usually language-specific**
- **Bilingual speaker of Farsi/English translated the selected 3000 English sentences into Farsi**

Farsi ASR Highlights – Vowelized vs. Grapheme

- **Baseline acoustic and language models trained on acoustic training data provided by NIST**
  - **Acoustic models estimated using maximum likelihood on 82 hours of Farsi training**
  - **Trigram LM trained using Knesser-Ney smoothing on 941K words of in-domain text using a lexicon of 36K words**
- **Compared vowelized lexicon to letter-to-phone (grapheme) approach for lexicon creation**
  - **Used vowelized lexicon provided from Appen (uses USC's pronunciation conventions)**
  - **8% relative improvement in WER for vowelized lexicon**

Lexicon	# Phones	%WER
Grapheme	28 speech + 4 non-speech	42.2
Vowelized USC Pronunciation	29 speech + 4 non-speech	38.7

Farsi ASR Highlights – Improved Language Model

- **Added harvested web data into the language model**
- **Discarded documents for LM training that had 20% OOV rate w.r.t. 36K Farsi lexicon**
- **Modest improvement in perplexity and WER on the internal E/F test set**

Farsi LM Data	Perplexity	%WER
E/F Farsi	200	32.6
E/F Farsi +Web	191	32.2

English ASR Highlights

- **Acoustic model trained on all English training data available in TRANSTAC E/I and E/F collections**
  - **135 hours available for acoustic training**
  - **Baseline acoustic models estimated using ML**
- **Experimented with multiple language models**

English LM Data	Lexicon Size	%WER
E/F	22K	31.1
E/F	29K	30.9
E/F + E/I	29K	26.0
E/F + E/I + Babylon CAST (BC)	29K	25.8
E/F + E/I + BC + Web	29K	25.6

Acoustic Modeling Improvements

- **Experimented with discriminative training and discriminative feature transformations**
- **Minimum phone error (MPE) training results in a 10% and 15% relative gain in Farsi and English respectively**
- **Hetroscedastic LDA (HLDA) features gives additional improvement in WER**

Training Criterion	Farsi %WER	English %WER
ML	36.4	34.7
MPE	33.1	29.3
HLDA-MPE	32.2	25.6

English/Farsi MT Development

- **Trained our SMT engine on the training data provided by NIST as well as on the expanded corpus**
- **Significant gains in all metrics for both data expansion techniques even with a small collection**

Farsi-to-English (F2E)			
F2E Data	**BLEU**	**METEOR**	**100-TER**
Baseline	31.2	61.0	38.9
Baseline + Directed	32.1	61.6	38.7
Baseline + Directed + Reused E/I	32.5	62.3	39.5
English-to-Farsi (E2F)			
E2F Data	**BLEU**	**METEOR**	**100-TER**
Baseline	18.0	47.2	40.4
Baseline + Directed	18.7	47.8	41.5
Baseline + Directed + Reused E/I	18.9	47.9	41.5

TRANSTAC July 2007 Evaluation Results

- **Offline evaluations on pre-recorded audio to measure WER and translation accuracy from text (T2T) and ASR output (S2T)**
- **Of the 4 systems evaluated by NIST, the BBN system ranked first in**
  - **"Completely Adequate" Likert percentages with smallest "Inadequate" Likert percentages for both directions**
  - **For automated metrics, all the numbers in "red" in the table below are the best results amongst all systems evaluated**

Condition		BLEU	METEOR	100-TER	100-WER
S2T	E2F	19.3	45.5	**41.0**	**84.7**
	F2E	**29.7**	**56.7**	**37.7**	72.3
T2T	E2F	**23.3**	50.3	**44.3**	N/A
	F2E	**35.7**	**63.2**	**45.7**	N/A

TRANSTAC July 2007 Evaluation Results – cont'd

- **Live evaluation of the S2S system with users for measuring**
  - **Complete Exchange: the number of high-level concepts that the English speaker was able to successfully retrieve in a 10-minute period**
  - **Proper Question: the number of English utterances correctly translated**
  - **Proper Answer: the number of Farsi utterances correctly translated**
- **BBN's system retrieved most number of concepts in all three metrics above**

Error Analysis of MT output – Farsi to English

- **Applied our novel error analysis methodology to find most damaging errors†**
- **Damage attributed as fraction of total Likert error (TLE) caused by an error type**
  - ***Likert Error*** **(*LE*) for a sentence is 5 – Likert Score**
  - ***Total Likert Error*** **(*TLE*) for a set of sentences is the sum of the Likert Error over that set**

Farsi-to-English			
Error Category	%Count	Weight	%TLE
Missing Concept Word	39.1	1.13	41.5
Wrong Word Sense	21.2	1.08	21.5
Wrong Word Order	20.5	0.89	17.2
Wrong Concept	10.4	1.29	12.6
Pronoun Error	2.9	0.86	2.4
MT OOV Word	2.0	1.0	1.8
Other	2.1	1.33	2.9

† David Stallard, et. al, "Recent Improvements and Performance Analysis of ASR and MT in a Speech-to-Speech Translation System," *Proc. ICASSP 2008.* Mar. 30 – Apr. 4 2008, Las Vegas, Nevada USA.

Error Analysis of MT output – English to Farsi

- **Most damaging errors: wrong word sense, wrong word order, wrong concept, missing concept words**
- **Top-4 error categories are the same in both directions but their relative ranking changes**

English-to-Farsi			
Error Category	**%Count**	**Weight**	**%TLE**
Wrong Word Sense	29.8	1.22	30.6
Wrong Word Order	23.4	1.13	22.3
Wrong Concept	11.1	1.58	14.7
Missing Concept Word	15.7	1.09	14.4
Extra Concept Verbiage	6.0	0.9	4.5
MT OOV Word	2.6	2.06	4.4
Pronoun Error	4.7	1.02	4.0
Other	6.8	0.87	5.0

Conclusions

- **Successfully developed an English/Farsi two-way S2S system in 90 days**
- **Small amount of targeted data collection at filling gaps in the system as well as reuse of existing collection improves translation performance**
- **Vowelized manual lexicon is better than the grapheme approach for Farsi ASR**
- **Most damaging errors: wrong word sense, wrong word order, wrong concept, and missing concept words**

Backup Slides

How Can We Quantify An Error Category's Importance?

- **Simply counting the number of the category's occurrences isn't adequate**
  - **Doesn't take the severity of the category into account**
- **We define importance as the total "damage" the error category does to a representative set of translations**
  - **Quantified as the relative reduction in Likert Score it causes**
- **We define the following measures:**
  - ***Likert Error*** **(*LE*) for a sentence is 5 – Likert Score**
  - ***Total Likert Error*** **(*TLE*) for a set of sentences is the sum of the Likert Error over that set**
  - **The *Weight* (W) of *C* is the average damage done by its instances.**
  - **The TLE for a category *C* is the damage done to the corpus by instances of *C*. It can be computed as:**

$$TLE(C) = Count(C)*Weight(C)$$

How Can We Compute An Error Category's Weight?

- **Many sentences have multiple errors. How should we apportion the blame among them?**
  - **Just "splitting the check" is unfair – it penalizes minor errors simply for appearing in the company of major ones**
- **Instead, treat each annotated utterance as an equation:**
  - **E.g. "SING2PLU + MISSING_CONCEPT = Likert Error of 3"**
- **The corpus then becomes a system of simultaneous equations, whose variables are the category weights**

  $Aw = k$ (*w* = weight vector, *k* = Likert Errors)
- **This system is unlikely to be consistent, but we can solve an approximation of it as least squares**

  $Aw = k + e$ (*e* is the error vector, minimize |e|)
- **Complete error weighting tool implemented in Java**
  - **Uses Java linear algebra package by Boisvert et al. at NIST**