



The MIT-LL/AFRL IWSLT-2008 MT System

**Wade Shen, Brian Delaney, Tim Anderson and
Ray Slyh**

23 October 2008

This work is sponsored by the United States Air Force Research Laboratory under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the United States Government.



Outline



System Overview

What's New in 2008?

- Segmentation Models
- System Combination
- Improved Arabic Morphological Processing
- Added Data
- Additional Improvements

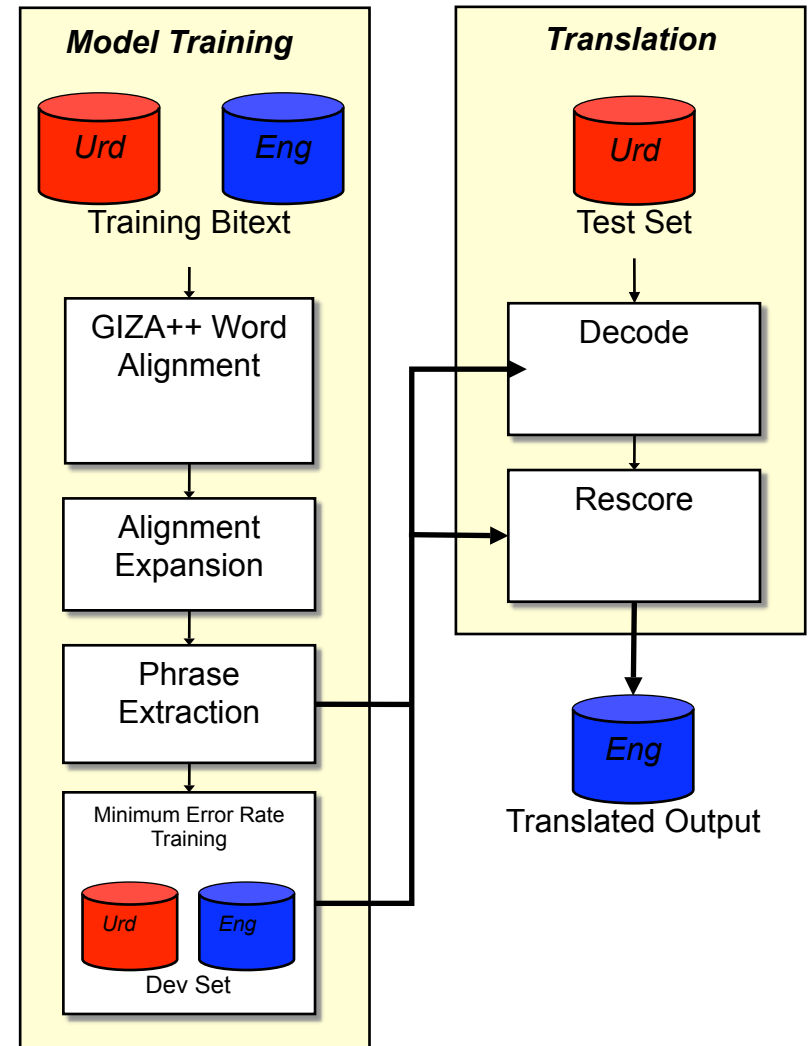
Summary



Statistical Translation System



- **Standard Statistical Architecture**
- **Developed in-house to support SMT experiments**
 - Framework for experiments with low-resource languages
 - Test-bed for S2S MT system
- **Custom Components**
 - FST-based Decoder
 - Segment EM Aligner/Decoder
 - MBR Rescoring
 - System Combination
 - Phrase Training/Minimum Error Rate Training
- **Use Moses decoder for baseline systems**





Phrase Based FST Decoder



- Based on MIT FST toolkit: <http://people.csail.mit.edu/ilh/fst/>
- The target language hypothesis is the best path through the following transducer:

$$E = I \circ P \circ D \circ T \circ L$$

- where,
 - I = source language input acceptor
 - P = phrase segmentation transducer
 - D = weighted phrase swapping transducer
 - T = weighted phrase translation transducer (source phrases to target words)
 - L = weighted target language model acceptor
- Apply phrase swapping twice for long distance reordering
- OOV words are inserted during decoding as parallel links to P, D, T, and L models.
- Allows for direct decoding on pruned ASR lattices



Outline



System Overview

What's New in 2008?

- **Segmentation Models**
- **System Combination**
- **Improved Arabic Morphological Processing**
- **Added Data**
- **Additional Improvements**

Summary



Segmentation Models

Introduction





Segmentation Models

Introduction



- **Scoring the Phrase-based Models**

$$P(\mathbf{E}|\mathbf{F}) \propto P(\mathbf{E})P(\mathbf{F}|\mathbf{E})$$

$$\approx P(\mathbf{E}) \max_{(\mathbf{f}, \mathbf{e})_1^k \in \text{seg}(\mathbf{F}, \mathbf{E})} p((\mathbf{f}, \mathbf{e})_1^k) * \prod_{i=1}^k p(\mathbf{f}_i | \mathbf{e}_i)$$



Segmentation Models

Introduction



- Scoring the Phrase-based Models

$$P(\mathbf{E}|\mathbf{F}) \propto P(\mathbf{E})P(\mathbf{F}|\mathbf{E})$$

$$\approx P(\mathbf{E}) \max_{(\mathbf{f}, \mathbf{e})_1^k \in \text{seg}(\mathbf{F}, \mathbf{E})} p((\mathbf{f}, \mathbf{e})_1^k) * \prod_{i=1}^k p(\mathbf{f}_i | \mathbf{e}_i)$$

Normally uniform or phrase penalty: $p((\mathbf{f}, \mathbf{e})_1^k) \approx e^{-\rho k}$





Segmentation Models

Introduction



- **Scoring the Phrase-based Models**

$$P(\mathbf{E}|\mathbf{F}) \propto P(\mathbf{E})P(\mathbf{F}|\mathbf{E})$$

$$\approx P(\mathbf{E}) \max_{(\mathbf{f}, \mathbf{e})_1^k \in \text{seg}(\mathbf{F}, \mathbf{E})} p((\mathbf{f}, \mathbf{e})_1^k) * \prod_{i=1}^k p(\mathbf{f}_i | \mathbf{e}_i)$$

Normally uniform or phrase penalty: $p((\mathbf{f}, \mathbf{e})_1^k) \approx e^{-\rho k}$



- **Uniform Segmentation probability**

$$P(\mathbf{E}|\mathbf{F}) \approx P(\mathbf{E}) * \max_{(\mathbf{f}, \mathbf{e})_1^k \in \text{seg}(\mathbf{F}, \mathbf{E})} \prod_{i=1}^k p(\mathbf{f}_i | \mathbf{e}_i)$$



Segmentation Models

Introduction



- **Scoring the Phrase-based Models**

$$P(\mathbf{E}|\mathbf{F}) \propto P(\mathbf{E})P(\mathbf{F}|\mathbf{E})$$

$$\approx P(\mathbf{E}) \max_{(\mathbf{f}, \mathbf{e})_1^k \in \text{seg}(\mathbf{F}, \mathbf{E})} p((\mathbf{f}, \mathbf{e})_1^k) * \prod_{i=1}^k p(\mathbf{f}_i | \mathbf{e}_i)$$

Normally uniform or phrase penalty: $p((\mathbf{f}, \mathbf{e})_1^k) \approx e^{-\rho k}$



- **Uniform Segmentation probability**

$$P(\mathbf{E}|\mathbf{F}) \approx P(\mathbf{E}) * \max_{(\mathbf{f}, \mathbf{e})_1^k \in \text{seg}(\mathbf{F}, \mathbf{E})} \prod_{i=1}^k p(\mathbf{f}_i | \mathbf{e}_i)$$

- **constant phrase penalty (bonus)**

$$P(\mathbf{E}|\mathbf{F}) \approx P(\mathbf{E}) \max_{(\mathbf{f}, \mathbf{e})_1^k \in \text{seg}(\mathbf{F}, \mathbf{E})} e^{-\rho k} * \prod_{i=1}^k p(\mathbf{f}_i | \mathbf{e}_i)$$



Segment EM Models

Framework





Segment EM Models

Framework



- **Explicit Segmentation Model**

$$P(\mathbf{E}|\mathbf{F}) \approx P(\mathbf{E}) \max_{(\mathbf{f}, \mathbf{e})_1^k \in \text{seg}(\mathbf{F}, \mathbf{E})} p((\mathbf{f}, \mathbf{e})_1^k) * \prod_{i=1}^k p(\mathbf{f}_i | \mathbf{e}_i)$$



Segment EM Models

Framework



- **Explicit Segmentation Model**

$$P(\mathbf{E}|\mathbf{F}) \approx P(\mathbf{E}) \max_{(\mathbf{f}, \mathbf{e})_1^k \in \text{seg}(\mathbf{F}, \mathbf{E})} p((\mathbf{f}, \mathbf{e})_1^k) * \prod_{i=1}^k p(\mathbf{f}_i | \mathbf{e}_i)$$

- **Segment EM models add non-uniform segmentation prob:**

$$P((\mathbf{f}, \mathbf{e})_1^k) \approx \prod_{i=1}^k p(\mathbf{f}_i | \mathbf{F}) * p(\mathbf{e}_i | \mathbf{E})$$



Segment EM Models

Framework



- **Explicit Segmentation Model**

$$P(\mathbf{E}|\mathbf{F}) \approx P(\mathbf{E}) \max_{(\mathbf{f}, \mathbf{e})_1^k \in \text{seg}(\mathbf{F}, \mathbf{E})} p((\mathbf{f}, \mathbf{e})_1^k) * \prod_{i=1}^k p(\mathbf{f}_i | \mathbf{e}_i)$$

- **Segment EM models add non-uniform segmentation prob:**

$$P((\mathbf{f}, \mathbf{e})_1^k) \approx \prod_{i=1}^k p(\mathbf{f}_i | \mathbf{F}) * p(\mathbf{e}_i | \mathbf{E})$$

- **Approx. monolingual segmentation probs:**

$$p(\mathbf{f}_i | F) \approx p(\mathbf{f}_i | \lambda) \approx \frac{E_{\mathcal{F}}(\mathbf{f}_i | \lambda)}{N_{\mathcal{F}}(\mathbf{f}_i)}$$

$$p(\mathbf{e}_i | E) \approx p(\mathbf{e}_i | \lambda) \approx \frac{E_{\mathcal{E}}(\mathbf{e}_i | \lambda)}{N_{\mathcal{E}}(\mathbf{e}_i)}$$



Segment EM Models

Framework



- **Explicit Segmentation Model**

$$P(\mathbf{E}|\mathbf{F}) \approx P(\mathbf{E}) \max_{(\mathbf{f}, \mathbf{e})_1^k \in \text{seg}(\mathbf{F}, \mathbf{E})} p((\mathbf{f}, \mathbf{e})_1^k) * \prod_{i=1}^k p(\mathbf{f}_i | \mathbf{e}_i)$$

- **Segment EM models add non-uniform segmentation prob:**

$$P((\mathbf{f}, \mathbf{e})_1^k) \approx \prod_{i=1}^k p(\mathbf{f}_i | \mathbf{F}) * p(\mathbf{e}_i | \mathbf{E})$$

- **Approx. monolingual segmentation probs:**

$$p(\mathbf{f}_i | F) \approx p(\mathbf{f}_i | \lambda) \approx \frac{E_{\mathcal{F}}(\mathbf{f}_i | \lambda)}{N_{\mathcal{F}}(\mathbf{f}_i)}$$

$$p(\mathbf{e}_i | E) \approx p(\mathbf{e}_i | \lambda) \approx \frac{E_{\mathcal{E}}(\mathbf{e}_i | \lambda)}{N_{\mathcal{E}}(\mathbf{e}_i)}$$

Expected number of f_i
from forced alignment

Number of possible
occurrences of f_i



Segment EM Models

Training Procedure





Segment EM Models

Training Procedure



1. Train standard phrase-based model



Segment EM Models

Training Procedure



- 1. Train standard phrase-based model**
- 2. Augment phrase model probabilities with initial segmentation probabilities**



Segment EM Models

Training Procedure



- 1. Train standard phrase-based model**
- 2. Augment phrase model probabilities with initial segmentation probabilities**
- 3. Force align training bitexts and dump lattices**



Segment EM Models

Training Procedure



1. Train standard phrase-based model
2. Augment phrase model probabilities with initial segmentation probabilities
3. Force align training bitexts and dump lattices
4. Compute phrase-pair expected values using fixed λ s from lattices (E-step)



Segment EM Models

Training Procedure



- 1. Train standard phrase-based model**
- 2. Augment phrase model probabilities with initial segmentation probabilities**
- 3. Force align training bitexts and dump lattices**
- 4. Compute phrase-pair expected values using fixed λ s from lattices (E-step)**
- 5. Reestimate segmentation probabilities using equations (M-step)**



Segment EM Models

Training Procedure



1. Train standard phrase-based model
2. Augment phrase model probabilities with initial segmentation probabilities
3. Force align training bitexts and dump lattices
4. Compute phrase-pair expected values using fixed λ s from lattices (E-step)
5. Reestimate segmentation probabilities using equations (M-step)
6. MER training to optimize model exponents (λ s)



Segment EM Models

Training Procedure



1. Train standard phrase-based model
2. Augment phrase model probabilities with initial segmentation probabilities
3. Force align training bitexts and dump lattices
4. Compute phrase-pair expected values using fixed λ s from lattices (E-step)
5. Reestimate segmentation probabilities using equations (M-step)
6. MER training to optimize model exponents (λ s)
7. Repeat 2-6



Chinese Experiments



System	dev7	dev3
Baseline (no rescoring)	39.6	52.9
+ phrase segmentation models	40.3	53.6
Baseline (with rescoring)	42.1	53.8
+ phrase segmentation models (iter=3)	42.8	54.1

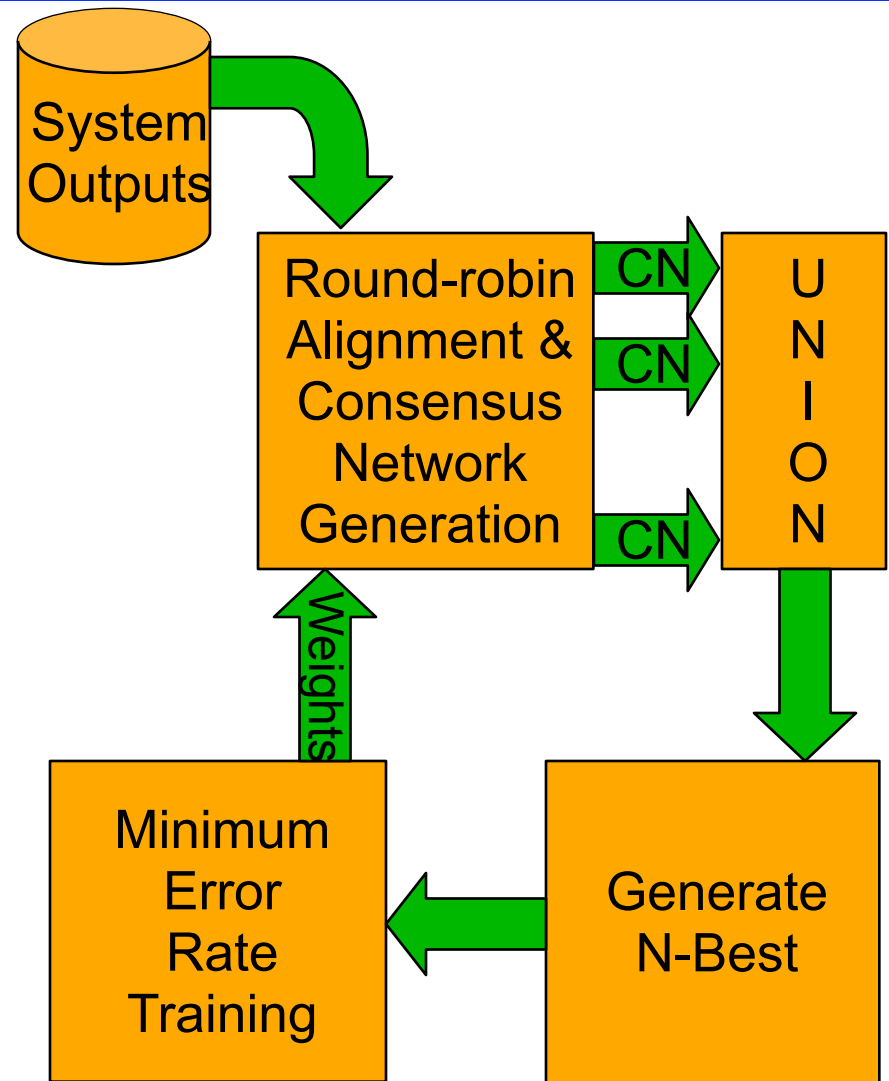
- **Segmentation Models improve overall on dev experiments**
 - Even with one iteration of training only
- **Multiple iterations + rescoring (post-eval) showed improvements**
 - Submitted results had no rescoring, one iter only



System Combination



- **Generate consensus networks using round-robin TER alignment, where each system gets to be the skeleton alignment**
- **Take union of all consensus networks and apply a language model**
- **Weight optimization via Nelder-Meade simplex on a development set using n-best lists**
 - Individual system weights, language model, word penalty, system priors
- **Final combination on unseen data using optimized weights**





System Combination Results

Chinese-to-English



CRR Input

System	dev3	eval
CE-contrast4	53.75	36.91
CE-contrast1	52.92	37.78
CE-contrast3	52.76	35.35
CE-contrast2	52.45	36.51
Combined		37.92

ASR Input

System	Input	dev3	eval
CE-contrast4	Conf. Net	45.80	31.93
CE-contrast3	1-Best	41.70	31.13
CE-contrast2	1-Best	41.65	31.41
CE-contrast7	Lattice	39.70	30.66
CE-contrast6	Lattice	38.84	31.02
Combined			35.38

- Used dev3 to train system combination weights
- CRR input condition
 - +0.14 BLEU → combination weights prefer CE-contrast4 which outperforms CE-contrast1 on dev3 but not on eval
- ASR input condition
 - +3.45 BLEU → **significant gain by combining systems with varying input sources** (1-Best vs. lattice vs. confusion network)



System Combination Results

Arabic-to-English



CRR Input

System	dev5	eval
AE-contrast4	27.95	55.07
AE-contrast3	27.91	54.91
AE-contrast1	26.03	50.81
AE-contrast2	28.25	51.79
Combined		56.51

ASR Input

System	Input	dev5	eval
AE-contrast4	Conf. Net	25.69	45.31
AE-contrast3	1-Best	25.34	45.63
AE-contrast1	Lattice	24.53	44.49
AE-contrast2	1-Best	23.44	44.35
Combined			48.92

- Used dev5 to train system combination weights (different ASR systems?)
- CRR input condition
 - +1.44 BLEU → decent gain despite change in system ranking between dev5 & eval sets
- ASR input condition
 - +3.29 BLEU → **significant gain by combining systems with varying input sources** (1-Best vs. lattice vs. confusion network)



Arabic Preprocessing



Preprocessing Method	Dev6
<i>Baseline (No normalization or AP5)</i>	42.06
<i>Remove all diacritics except tanween, no AP5</i>	49.40
<i>Remove all diacritics, no AP5</i>	50.39
<i>Remove all diacritics, apply AP5</i>	53.55

- Diacritics removed:
 - **Short vowels**
 - **Sukuun:** Marks absence of sort vowel
 - **Shadda:** Marks consonant gemination (i.e., doubling)
 - **Tanween:** Case markers for indefinite forms & other uses
 - **Tatweel:** Stretches letters in Arabic typography (not a true diacritic)
- AP5 segments the following from stems:
 - **Prefixes:** al-, bi-, fa-, ka-, li-, wa-
 - **Suffixes:** Attached pronouns



Additional Parallel and Monolingual Data

Chinese ↔ English Tasks



- **Additional parallel data (out of domain) improves system**
 - *Added ISI Chinese-English parallel corpus to training*
- **Additional English data also helps: *Added Gigaword LMs***
- **ISI data not used for evaluation systems**



Additional Parallel and Monolingual Data

Chinese ↔ English Tasks



- **Additional parallel data (out of domain) improves system**
 - *Added ISI Chinese-English parallel corpus to training*
- **Additional English data also helps: Added Gigaword LMs**
- **ISI data not used for evaluation systems**

Chinese to English	Configurations	Language Models			Eval Set	
		Decode	Rescore	Class	MAP	MBR
	Baseline System	4g,5g		7g	53.32	53.50
	+ ISI -corpus	4g	4g ISI	7g	55.30	54.28
	+ English Gigaword	4g	6g GIGA	7g	54.60	54.10

English to Chinese	Configurations	Language Models			Eval Set	
		Decode	Rescore	Class	MAP	MBR
	Baseline System	4g,5g		7g	29.75	29.88
	+ ISI -corpus	4g	5g ISI	7g	30.76	31.60



Other Improvements



- **Minimum Bayes Risk Rescoring**
 - Results mixed: Improve E->C task (eval), C->E (dev only)
- **Added Phrases from Berkeley Aligner**
 - Consistent +0.5 point gain
- **Additional Lexicon Data (C->E) from CEDICT**
 - 0.5-1.0 improvement depending on DEV set
- **Improved Confusion Network decoding**
 - Language model/acoustic model optimization separated
 - Better word splitting from LIG [Besacier 07]
 - Improvement: ~0.5-1.0 points
- **Lexical Approximation [Mermer 07]: +0.5 points**



Outline



System Overview

What's New in 2008?

- Segmentation Models
- System Combination
- Improved Arabic Morphological Processing
- Added Data
- Additional Improvements

Summary



Summary



- **Segment-EM Model**
 - Preliminary results show modest improvement
- **Improved System Combination**
 - Results from ASR are especially promising for both Arabic and Chinese
- **Better morphology for Arabic**
 - Significant improvements from AP5 diacritic norm, and Lexical Approximation
 - More normalizations could further improve phrase estimation
- **Make better use of out-of-domain data**
 - Some improvements this year from ISI arabic data and GIGAWORD