The MITLL/AFRL IWSLT-2007 MT System

Wade Shen, Brian Delaney, Tim Anderson and Ray Slyh

27 November 2006

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.

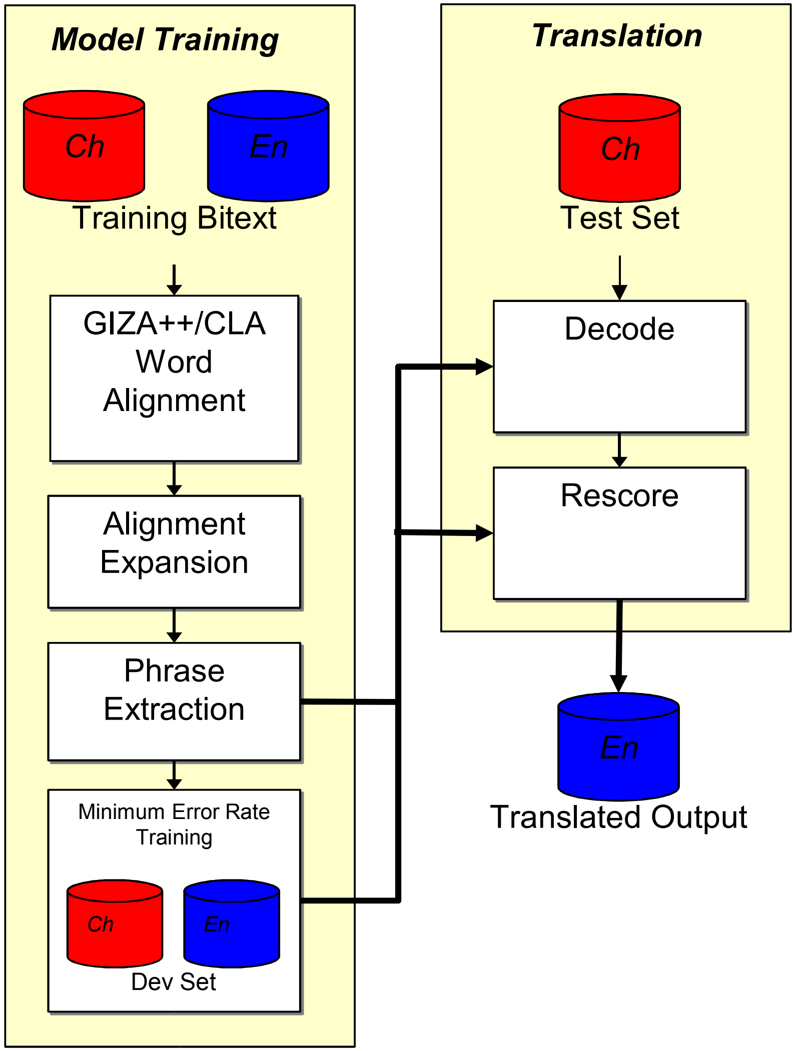
Statistical Translation System

Experimental Architecture

- **Standard Statistical Architecture**

- **New this year**
  - **Light Morphology for Arabic**
  - **Better Speech Decoders**
    - **Lattice-based decoder**
    - **Better conf-net decoding w/**`moses`
  - **Rescoring Features**

- **Participated in**
  - **Chinese ⇨ English**
  - **Arabic ⇨ English**
  - **Italian ⇨ English**



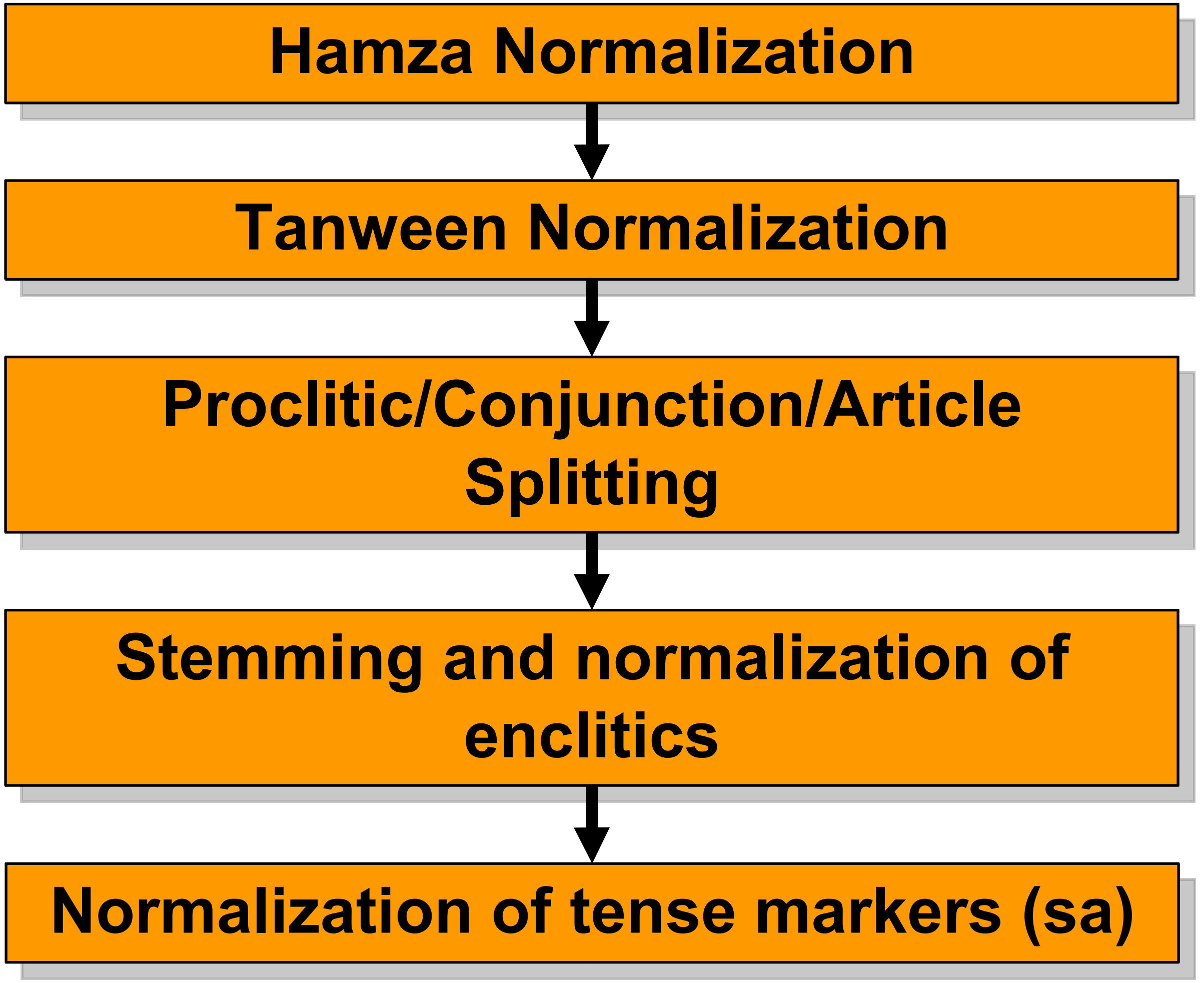
The MITLL/AFRL MT System

Overview

➡ - Light Arabic Morphology
- Improved Confusion Network Decoding
- Direct Lattice Decoding
  - Reordering Constraints
  - Higher-order LMs
- Experiments
  - Lattice vs. Confusion Network Decoding
  - Arabic Preprocessing
- Summary

Light Arabic Morphology

AP5 Process



Hamza Normalization

Tanween Normalization

Proclitic/Conjunction/Article Splitting

Stemming and normalization of enclitics

Normalization of tense markers (sa)

Light Arabic Morphology

Effect

ستحط ال طائرة خلال ساعة ستحط
سنقدم وجبة الغذاء بعدثلاثين دقيقة من
الإقلاع
الحمام في مؤخرة الطائرة اتبعني رجاء

No Processing

ستحط ال طائرة خلال ساعة تقريبا
سنقدم و جبة ال غذاء ب عد ثلاثين دقيقة من ال
إقلاع
ال حمام ف ي مؤخرة ال طائرة اتبع ني رجاء

AP5 Processing

- **Marker `(post)` used to disambiguate suffixes and prefixes**
- **Reduce OOV Rate: 12.3% → 7.26%**
- **Regularized affix and suffix forms**

The MITLL/AFRL MT System

Overview

- Light Arabic Morphology
- Improved Confusion Network Decoding
- Direct Lattice Decoding
  - Reordering Constraints
  - Higher-order LMs
- Experiments
  - Lattice vs. Confusion Network Decoding
  - Arabic Preprocessing
- Summary

Confusion Network Repunctuation

- No longer rely on 1-best repunctuation alone
- Process
  - Convert lattice to confusion network
  - Insert punctuation between columns using all possible n-gram contexts surrounding current column
  - Sum Posteriors of different contexts per punctuation mark
- Significantly more processing requirements
  - Average: n^k where n is n-gram order of punctuation model and k is average column depth

Improved Confusion Network Decoding

- Use of component ASR scores
  - No longer rely on ASR posterior and fixed scaling
  - Expose Source LM and acoustic model scores

- MER Training with ASR scores
  - Interaction of source/target word penalties
  - Use ASR path posterior to optimize source word penalty

    i.e. optimize E(source length)

- Results in improved performance in all languages

The MITLL/AFRL MT System

Overview

- Light Arabic Morphology
- Improved Confusion Network Decoding
- ➡ Direct Lattice Decoding
  - Reordering Constraints
  - Higher-order LMs
- Experiments
  - Lattice vs. Confusion Network Decoding
  - Arabic Preprocessing
- Summary

Direct ASR Lattice Decoding using Finite State Transducers

- As an alternative to decoding on confusion networks, we perform direct decoding of ASR lattices using finite state transducers
- The target language hypothesis is the best path through the following transducer:

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

- where,
  - I = weighted 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

FST Decoder Implementation

- Based on MIT FST toolkit: http://people.csail.mit.edu/ilh/fst/

- Phrase swapping transducer can be applied twice for long distance reordering → inefficient but simple

- Pruning strategy
  - Apply wide beam on full path scores after composition with T
  - Viterbi search with narrow beam during language model search

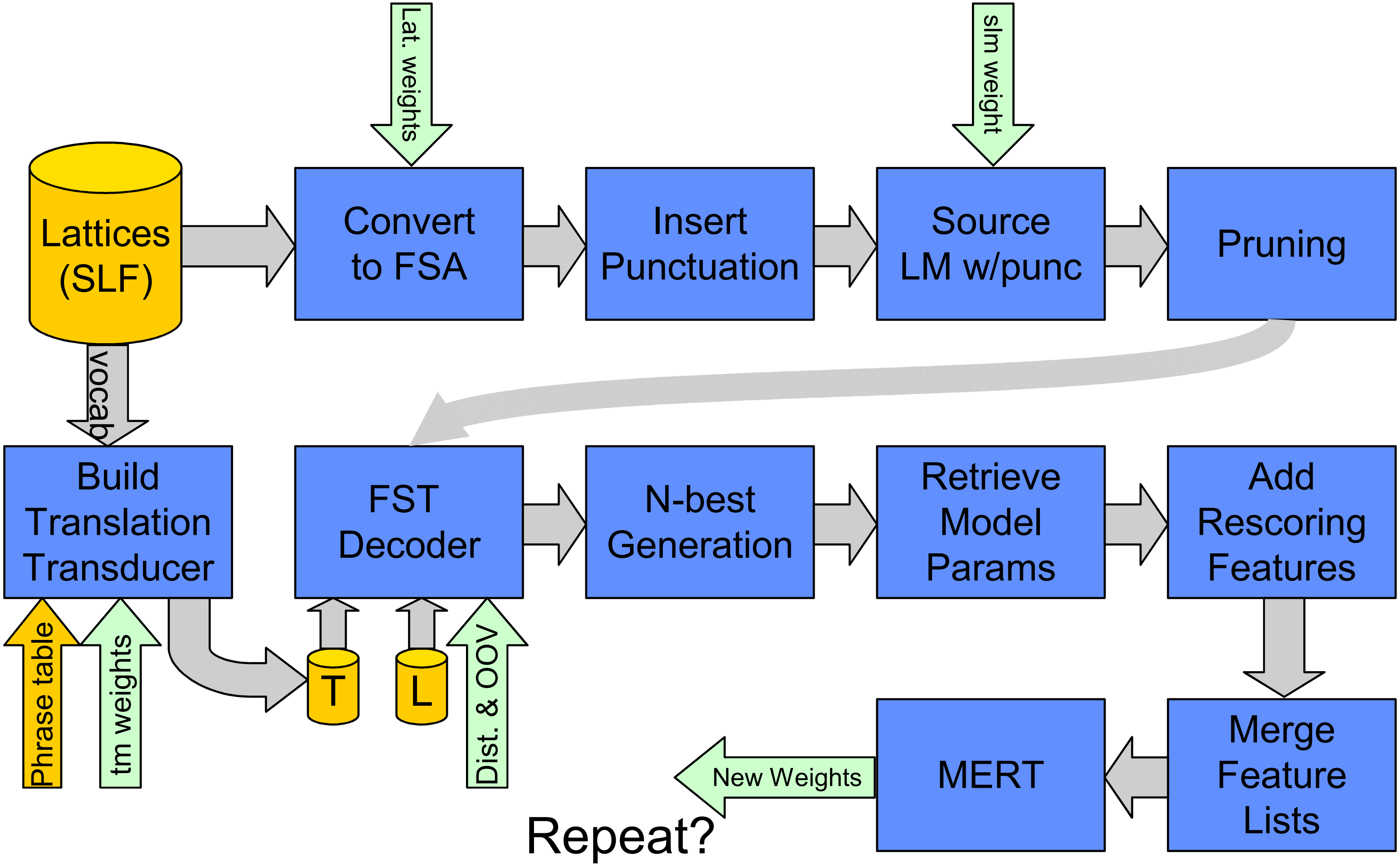
- OOV words are detected and added as parallel paths to P, T, and L transducers → OOV penalty discourages OOV words when multiple paths exist

- Minimum error rate training requires some extra work to recover individual model parameters for weight optimization

Model Parameters For ASR Lattice Decoding

Input lattice parameters	
$P(X\|f)$	Source acoustic model
$P(f)$	Source language model
$W_{pen}(f)$	Source word insertion penalty
$P_{punc}(f)$	Source language model w/punctuation
OOV_{pen}	Out-of-vocabulary penalty
Translation model parameters	
$P(e\|f), P(f\|e)$	Bi-directional phrase probabilities
$P_{lw}(e\|f), P_{lw}(f\|e)$	Bi-directional lexical probabilities
P_{pen}	Phrase penalty
$Lexbo$	Lexical back-off penalty
$W_{pen}(e)$	Target word insertion penalty
Other models	
$P(e)$	Target language model
P_d	Distortion penalty
Rescoring Features	
$P_{IBM1}(f\|e)$	Sentence level IBM Model 1
$P_{clm}(e)$	Class n-gram target LM
$P_{rlm}(e)$	High order target LM

Minimum Error Rate Training with FST Decoder



The MITLL/AFRL MT System

Overview

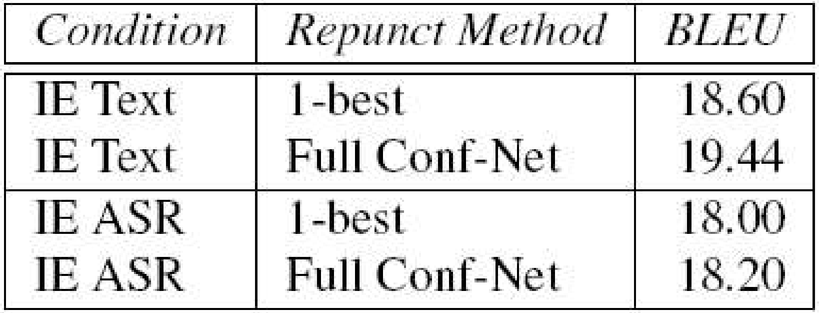
- Light Arabic Morphology
- Improved Confusion Network Decoding
- Direct Lattice Decoding
  - Reordering Constraints
  - Higher-order LMs
- ➡ Experiments
  - Lattice vs. Confusion Network Decoding
  - Arabic Preprocessing
- Summary

ASR Lattice Decoding Results

Language (dev/test)	Condition	BLEU
CE (dev4/dev5)	Fixed ASR Lattice Weights	16.98
CE (dev4/dev5)	+Optimized ASR Lattice Weights	17.23
CE (dev4/dev5)	+Rescoring Features	18.27
AE (dev4/dev5)	Fixed ASR Lattice Weights	21.55
AE (dev4/dev5)	+Optimized ASR Lattice Weights	22.70
AE (dev4/dev5)	+Rescoring Features	23.73
IE (dev4/dev5)	Fixed ASR Lattice Weights	29.45
IE (dev4/dev5)	+Optimized ASR Lattice Weights	29.42
IE (dev4/dev5)	+Rescoring Features	30.15
IE (dev5bp1/dev5bp2)	Fixed ASR Lattice Weights	16.25
IE (dev5bp1/dev5bp2)	+Optimized ASR Lattice Weights	17.06
IE (dev5bp1/dev5bp2)	+Rescoring Features	17.90

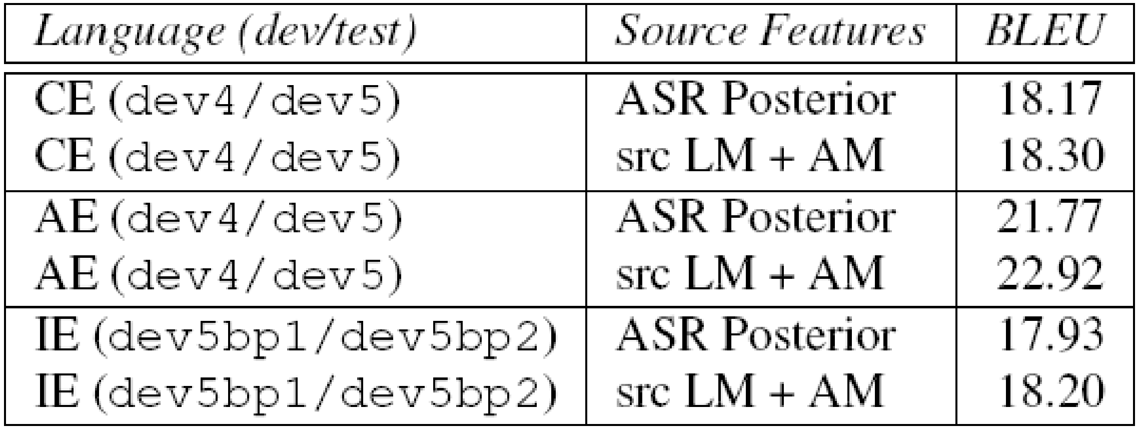
Confusion Network Results

- **Repunctuation**



Condition	*Repunct Method*	*BLEU*
IE Text	1-best	18.60
IE Text	Full Conf-Net	19.44
IE ASR	1-best	18.00
IE ASR	Full Conf-Net	18.20

- **Posterior vs. Separate AM and LM scores**



Language (dev/test)	*Source Features*	*BLEU*
CE (dev4/dev5)	ASR Posterior	18.17
CE (dev4/dev5)	src LM + AM	18.30
AE (dev4/dev5)	ASR Posterior	21.77
AE (dev4/dev5)	src LM + AM	22.92
IE (dev5bp1/dev5bp2)	ASR Posterior	17.93
IE (dev5bp1/dev5bp2)	src LM + AM	18.20

Confusion Network vs. Lattice Decoding

- **All configurations use rescoring features**

- **Different models for distortion:** phrase vs. word

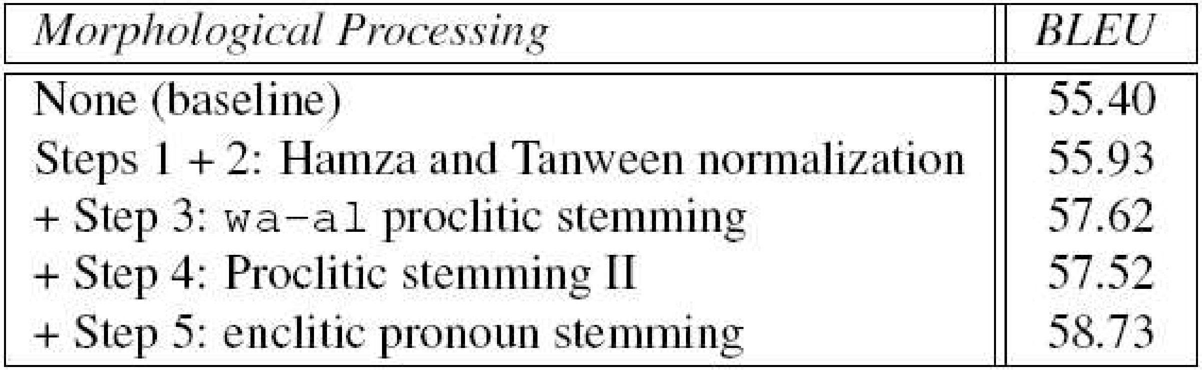
Language (dev/test)	Condition	BLEU
CE (dev4/dev5)	Confusion Network	18.30
CE (dev4/dev5)	Lattice Decoding	18.27
AE (dev4/dev5)	Confusion Network	22.92
AE (dev4/dev5)	Lattice Decoding	23.73
IE (dev5bp1/dev5bp2)	Confusion Network	18.20
IE (dev5bp1/dev5bp2)	Lattice Decoding	17.90

- **Similar performance for CE and IE**

- **Arabic improvement with lattice decoding**
  - ConfNet posterior issues due to ASR mismatch (?)

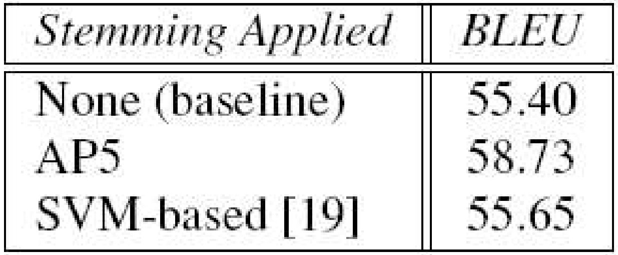
Arabic Morphology

- **Improvement from each AP5 Processing Step**



Morphological Processing	*BLEU*
None (baseline)	55.40
Steps 1 + 2: Hamza and Tanween normalization	55.93
+ Step 3: `wa-al` proclitic stemming	57.62
+ Step 4: Proclitic stemming II	57.52
+ Step 5: enclitic pronoun stemming	58.73

- **Compare with ASVM and no morphology**



Stemming Applied	*BLEU*
None (baseline)	55.40
AP5	58.73
SVM-based [19]	55.65

The MITLL/AFRL MT System

Overview

- Light Arabic Morphology
- Improved Confusion Network Decoding
- Direct Lattice Decoding
  - Reordering Constraints
  - Higher-order LMs
- Experiments
  - Lattice vs. Confusion Network Decoding
  - Arabic Preprocessing
- Summary

Summary

- Significant improvement for Arabic with light morphology
  - Five Deterministic Rules
- Improved Confusion Network decoding with separate source LM and ASR scores
- Lattice-based Decoding comparable to Confusion Network decoding
  - Improvement for Arabic Task