

# THE CUED NIST 2009 ARABIC-ENGLISH SMT SYSTEM

Adrià de Gispert, Gonzalo Iglesias, Graeme Blackwood, Jamie Brunning, Bill Byrne



Department of Engineering  
University of Cambridge

NIST Open MT 2009 Evaluation Workshop  
Ottawa, 31 August 2009

# The CUED SMT system

- ▶ **Lattice-based Hierarchical SMT system implemented with weighted finite state transducers (WFSTs)**
  - HiFST decoder
  - Based on the Google OpenFST toolkit
- ▶ **Hierarchical translation rules extracted from MTTK-aligned parallel text**
  - Training data: identical to NIST MT 2008
  - Arabic-to-English : ~6M sentences, ~150M words
- ▶ **Hybrid system based on three Arabic morphological decompositions**
  - MADA-D2, MADA-D3, Sakhr
- ▶ MET parameter optimization for BLEU, separately for newswire and web text
  - MT06 newswire and web sets for tuning
- ▶ Two English language models
  - Top cell pruning: 4gram Knesser-Ney estimated over 0.8B words
  - Lattice generation: zero cut-off, stupid backoff 5gram estimated over 4.7B words
  - Exact back-off implementations with failure transitions
- ▶ **Lattice-based MBR rescoring**
- ▶ Casing with SRILM

# HiFST. Hierarchical Translation with WFSTs <sup>1</sup>

- ▶ Goal: Keep all possible derivations in each cell

Efficiently explore largest  $\mathcal{T}$  in

$$\operatorname{argmax}_{t \in \mathcal{T}} P(s|t) P(t)$$

S	X		
x8420	x20		
x420	x20		
x20	x20	x20	x20
	$s_1$	$s_2$	$s_3$

- ▶ CYK grid constructed in the usual way
- ▶ **Build a WFSA in each cell for target side of translation rules active in each cell**
  - ▶ Avoids cube pruning
  - ▶ Faster and more efficient to work with sublattices containing many hypotheses than to work individually with distinct hypotheses

In each cell, do:

For each rule in the cell:

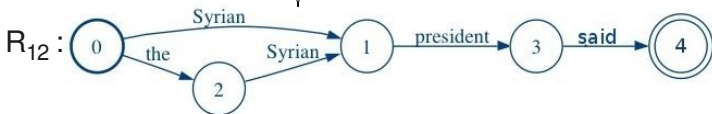
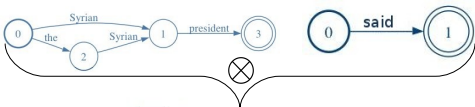
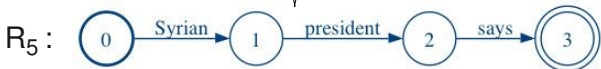
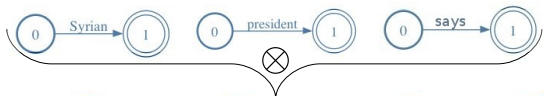
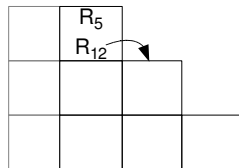
Build Rule WFSA by **Concatenating** target elements (  $\otimes$  )

Build Cell WFSA by **Unioning** Rule FSAs (  $\oplus$  )

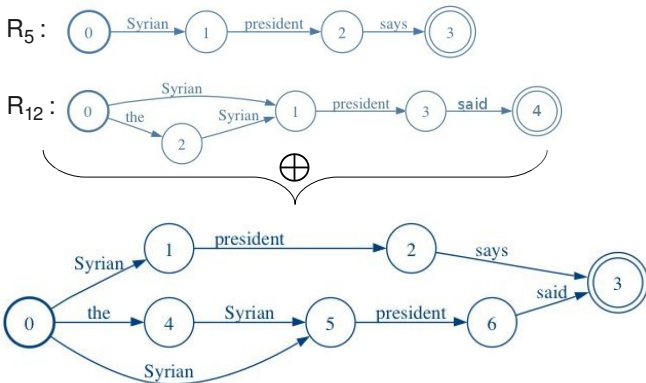
<sup>1</sup>G. Iglesias, A. de Gispert, E. R. Banga, and W. Byrne. 2009. Hierarchical Phrase-Based Translation with Weighted Finite State Transducers. Proc. of NAACL-HLT.

## Building Rule WFSTs by Concatenation

$R_5: X \rightarrow \langle s_1 s_2 s_3, \text{Syrian president says} \rangle$   
 $R_{12}: X \rightarrow \langle s_1 X, X \text{ said} \rangle$

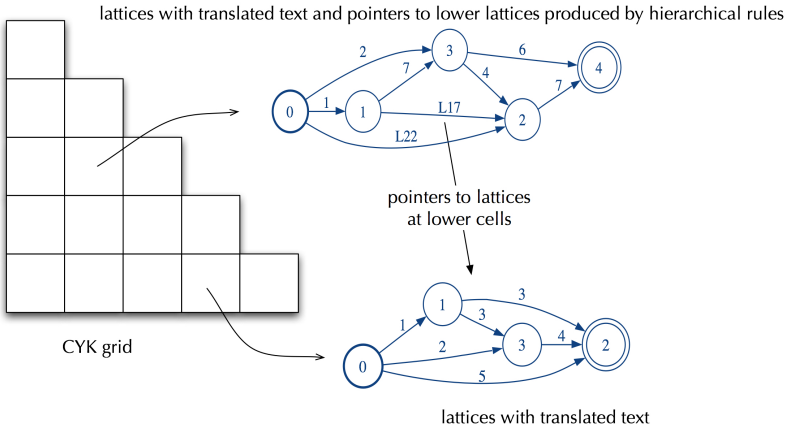


## Building Cell WFSA by Union



- ▶ Can be made compact
- ▶ Target language model can be applied, if pruning during search is required
- ▶ Translation hypothesis space can be built incrementally and efficiently

## Delayed Translation



- ✓ Avoids expanding and replicating hypotheses until needed
- ✓ Easy implementation with FST Replace operation
- ✓ Usual FST operations can be applied to skeleton → lattice size reduction

## Building the Rule Set

Hierarchical rules extracted from MTTK-aligned parallel text.

- ▶ Standard constraints <sup>2</sup>:
  - ▶ maximum number of non-terminals is two
  - ▶ disallow adjacent non-terminals in the source language
  - ▶ unaligned words are not allowed at the edges of the rule
  - ▶ require at least one pair of aligned words per rule
- ▶ Rule filtering <sup>3</sup>
  - ▶ by pattern: discard monotonic patterns
    - $\langle wX, wX \rangle$  and  $\langle Xw, Xw \rangle$
    - $\langle wX_1wX_2, wX_1wX_2 \rangle$  and  $\langle X_1wX_2w, X_1wX_2w \rangle$
    - $\langle wX_1wX_2w, wX_1wX_2w \rangle$
  - ▶ by number of translations (20)

<sup>2</sup>D. Chiang, 2005. A Hierarchical Phrase-Based Model for Statistical Machine Translation. Proc. ACL.

<sup>3</sup>G. Iglesias, A. de Gispert, E. R. Banga, and W. Byrne. 2009. Rule Filtering by Pattern for Efficient Hierarchical Translation. Proc. of EACL.

## Grammar configurations: Shallow grammar

Goal: avoid nested hierarchical rules of non-terminals

<b>full hierarchical grammar</b>	
$S \rightarrow \langle X, X \rangle$	glue rule
$S \rightarrow \langle S X, S X \rangle$	glue rule
$X \rightarrow \langle \gamma, \alpha, \sim \rangle, \gamma, \alpha \in \{X \cup \mathbf{T}\}^+$	hiero rules of any level

- ▶ For Arabic-to-English, shallow-1 grammar performs as full hiero
- ▶ Constrained search space, but can be built exactly and quickly - no pruning required

<b>shallow-1 grammar</b>	
$S \rightarrow \langle X^1, X^1 \rangle$	glue rule
$S \rightarrow \langle S X^1, S X^1 \rangle$	glue rule
$X^1 \rightarrow \langle \gamma^0, \alpha^0, \sim \rangle, \gamma^0, \alpha^0 \in \{\{X^0\} \cup \mathbf{T}\}^+$	hiero rules level 1
$X^0 \rightarrow \langle \gamma^p, \alpha^p \rangle, \gamma^p, \alpha^p \in \mathbf{T}^+$	regular phrases



## Grammar configurations: Low-level rule concatenation for long-range verb movement

- ▶ Allow for VSO to SVO
- ▶ Grammar allows long-range movement if the target-side contains a verb
- ▶ Otherwise, allow only monotonic rule concatenation (includes local reordering)
- ▶ Enforced by adding non-terminals:
  - ▶ additional non-terminal  $M^{n,k}$  allows monotonic concatenation of  $k$  rules  $X^n$
  - ▶  $M^{n,k}$  only used in rule  $X^{n+1}$  if  $\exists s \in \mathbf{T}, s \subset \gamma^n / s$  is verb
- ▶ +0.4 BLEU in newswire

rules included	
$S \rightarrow \langle X^2, X^2 \rangle$	glue rule
$S \rightarrow \langle S X^2, S X^2 \rangle$	glue rule
$X^2 \rightarrow \langle \gamma^1, \alpha^1, \sim \rangle, \gamma^1, \alpha^1 \in \{ \{X^1, M^{1,2}, M^{1,3}\} \cup \mathbf{T} \}^+,$	hiero rules level 2
$X^1 \rightarrow \langle \gamma^0, \alpha^0, \sim \rangle, \gamma^0, \alpha^0 \in \{ \{X^0\} \cup \mathbf{T} \}^+$	hiero rules level 1
$X^0 \rightarrow \langle \gamma^p, \alpha^p \rangle, \gamma^p, \alpha^p \in \mathbf{T}^+$	regular phrases
$M^{1,2} \rightarrow \langle X^1 X^1, X^1 X^1 \rangle$	catenation of 2 $X^1$
$M^{1,3} \rightarrow \langle X^1 M^{1,2}, X^1 M^{1,2} \rangle$	catenation of 3 $X^1$

## Automatic Genre Detection

- ▶ Goal: Train genre-specific Arabic LMs to classify input documents by perplexity
  - ▶ LM data is tokenized Arabic portion of parallel text for NIST 09 evaluation task
  - ▶ LM vocabulary is the set of all Arabic words in the respective evaluation set
  - ▶ Train Witten-Bell 3-gram NW and NG LMs from genre-specific subsets of parallel text
- ▶ Classify Arabic input document as newswire if  $PP_{nw} < PP_{ng}$  and newsgroup otherwise

		Classification	
		nw	ng
Reference	nw	73	1
	ng	14	36

**Table:** Classification results for the 74 newswire (nw) and 50 newsgroup (ng) documents of the NIST mt08 evaluation set. 14 newsgroup documents are incorrectly labeled as newswire by the classifier.

- ▶ Translation performance not greatly affected by incorrectly classified documents:

	mt08-nw			mt08-ng		
	BLEU	TER	BP	BLEU	TER	BP
Reference	49.3	43.7	0.979	34.4	55.2	0.977
Genre Classifier	49.3	43.7	0.979	34.3	54.9	0.969

**Table:** BLEU score, TER and brevity penalty (BP) for first pass translation using reference labels and perplexity-based classifier labels for newswire (mt08-nw) and newsgroup (mt08-ng) evaluation subsets.

- ▶ If all documents are misclassified, the overall score degrades by approximately -1.0 BLEU

## Arabic Morphological Decompositions

- ▶ Three decompositions considered: MADA-D2, MADA-D3 and SAKHR
- ▶ Build independent rule sets for each decomposition
- ✗ Individual decompositions are not robust across genres
- ✓ Motivation for developing a **hybrid system**: consider all decompositions simultaneously

BLEU — 4-gram LM <b>decomposition</b>	<b>newswire</b>		<b>web</b>	
	<b>mt06-nist</b>	<b>mt08</b>	<b>mt06-nist</b>	<b>mt08</b>
MD2	51.1	51.4	37.8	36.3
MD3	51.0	51.4	38.0	36.2
Sakhr	51.4	51.5	36.9	35.5

# Lattice Minimum Bayes Risk for Multiple Source Language Analyses

- ▶ Lattice-based Minimum Bayes Risk has gains over N-best MBR <sup>4</sup>
- ▶ MBR can be used to merge hypothesis from multiple analyses <sup>5</sup>
  - ▶ Hierarchical decoder with one rule set for each analysis and a common target LM
  - ▶ Scores (posterior distributions over ngrams) derived from each rule set are interpolated to form a single distribution for MBR
- ▶ An alternative to packing multiple analyses into a lattice for lattice-based translation
  - ▶ quite easy if an MBR implementation is available

		<b>mt02-05-tune</b>		<b>mt02-05-test</b>		<b>mt08</b>	
		<b>BLEU</b>	<b>TER</b>	<b>BLEU</b>	<b>TER</b>	<b>BLEU</b>	<b>TER</b>
MD2	5gram LM, 1-best	54.2	40.5	53.8	41.0	44.9	48.5
MD3		53.8	41.2	53.6	41.4	45.0	48.3
Sakhr		54.1	40.7	53.8	40.7	44.7	48.7
MD2+MD3	N-Best MBR	55.1	40.0	54.7	40.3	46.1	47.7
	LMBR	55.7	40.1	55.4	40.2	46.7	47.8
MD2+Sakhr	N-Best MBR	55.4	39.7	54.9	39.9	46.5	47.7
	LMBR	56.0	39.5	55.9	39.6	46.9	47.4
MD2+Sakhr+MD3	N-Best MBR	55.3	39.7	54.9	40.0	46.5	47.8
	LMBR	56.0	39.6	55.7	39.7	47.3	46.9

<sup>4</sup>R. Tromble, S. Kumar, F. Och, and W. Macherey. 2008. Lattice Minimum Bayes-Risk Decoding for Statistical Machine Translation. Proceedings of the EMNLP, pp. 620-629.

<sup>5</sup>A. de Gispert, S. Virpioja, M. Kurimo, and W. Byrne. Minimum Bayes risk combination of translation hypotheses from alternative morphological decompositions. Proceedings of NAACL-HLT, 2009.

## Conclusion

### New this year: Syntax

- ▶ HiFST: WFST implementation of Hiero
- ▶ Grammar configuration strategies:
  - ▶ Shallow translation
  - ▶ Modeling of verb movement for Arabic to English MT → good gains in newswire

### Hybrid translation system

- ▶ Good gains from lattice MBR translation based on multiple morphological analyses
- ▶ Particularly effective for web translation

### Emphasis on efficient construction of translation search spaces

- ▶ Minimal pruning
- ▶ Aim for few (i.e. no) search errors under the translation grammar
- ▶ Why: Much richer search spaces leads to greater gains in subsequent decoding steps

### Entire translation process is essentially a front-end for MBR decoding

- ▶ HiFST produces (a) very large translation lattices and (b) n-gram posterior distributions
- ▶ No translation hypotheses are produced until the final LMBR decoding step

### N.B. Our Chinese-English SMT system is pretty good, too

- ▶ Since Chinese-English was offered only as a progress set, we focused on Arabic-English

## Acknowledgments

- ▶ This work was supported in part by the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022
- ▶ G. Iglesias was supported by a Spanish Government research grant BES-2007-15956 (project TEC2006-13694-C03-03)
- ▶ Thanks to Sakhr Inc for use of Arabic text processed by their morphological analyzer



Department of Engineering  
University of Cambridge