UNIVERSITY OF CAMBRIDGE

The CUED OpenMT12 Arabic-English and Chinese-English SMT Systems

Bill Byrne

Adrià de Gispert, Gonzalo Iglesias, Juan Pino, Rory Waite

Department of Engineering

UCAM SMT Systems - Overview

- Lattice-based Hierarchical SMT system implemented with WFSTs
  - HiFST decoder -- based on the Google OpenFST toolkit
- Alignments with the MTTK HMM toolkit
  - All allowable parallel text
- Hybrid systems for Ar-EN and Zh-En
  - Multiple source-side segmentations
  - Genre-specific lexical and translation features
  - MERT & Pro tuning
- Zero-cutoff stupid-backoff 5-gram LMs
  - Google n-gram data
- Lattice Minimum Bayes Risk Decoding

Source Language Text Processing

- Zh-En: 3 alternative segmentations
  - Stanford PKU and CTB [1]
  - Joint Word Segmenter / POS-tagger trained for low OOVs [2]
- Ar-En: 2 alternative segmentations
  - Both provided by MADA [3]

Word-to-Phrase HMM models used for alignment [4]

[1] http://nlp.stanford.edu/software/segmenter.shmtl

[2] Y.Zhang, S. Clark. A Fast Decoder for Joint Word Segmentation and POS-tagging Using a Single Discriminative Model. EMNLP 2010.

[3] N. Habash, F. Sadat. Arabic Preprocessing Schemes for Statistical Machine Translation. HLT-NAACL, 2006.

[4] Y. Deng, W. Byrne. HMM Word and Phrase Alignment for Statistical Machine Translation. IEEE T-ASLP 2008

Hiero Translation Grammars

- Separate grammar extracted for each source segmentation
  - Ar-En : Shallow-1 grammar with long-distance verb movement [5]
  - Zh-En : Full Hiero grammars
- Hadoop/Hfile framework for grammar extraction and retrieval
- Some monotonic rules excluded by pattern
  - Zh-EN:, e.g. <X1 w X2, w X1 w X2> , <X w, X w>
  - Ar-En: <w X, w X>
  - 10 instances for hierarchical rules, and rule filtering by probability

[5] de Gispert et al. Hierarchical Phrase-Based Translation with Weighted Finite-State Transducers and Shallow-N Grammars. Computational Linguistics, 2010.

Features and Feature Weight Tuning

- The usual collection of translation rule features (next slide)
- Provenance Features
  - Translation rule probabilities and lexical features estimated on genre-specific portions of the parallel texts
    - Essentially one translation grammar for each portion of the parallel text
- Feature weights optimized towards BLEU with MERT [6] and Pro [7]
  - Pro was particularly good with sparse features, but favored shorter hyps
  - Translation length models were used to tune towards longer hypotheses

[6] FJ Och. Minimum Error Rate Training in Statistical Machine Translation. ACL 2003

[7] M. Hopkins and J. May. Tuning as Ranking. EMNLP 2011 .

Description	# AR-EN	# ZH-EN
Source-to-target probability	1	1
Target-to-source probability	1	1
Word penalty	1	1
Rule Penalty	1	1
Rule Count = 1,2,>=3	3	3
Deletion and OOV rule	1	1
Is Glue ?	1	1
Source-to-target lexical probability	1	1
Target-to-source lexical probability	1	1
Provenance source-to-target probability	31	9
Provenance target-to-source probability	31	9
Provenance source-to-target lexical probability	31	9
Provenance target-to-source lexical probability	31	9
Is source length = 1..7 ?	7	Not used
Is target length = 1..7 ?	7	Not used
Is source length = 1..7 and target length = 1..7 ?	49	Not used
Does rule have pattern P ?	65	Not used

- AR-EN: provenance features split by genre and LDC collection ID
  - genres: un, nw, ng, bn, bc, wl, treebank
- ZH-EN: provenance features split by genre:
  - genres: bc, bn, ng, nw, lexicon, treebank, web, un, null

English Language Models

- LMs used all allowable English text
- English language model is the equal interpolation of
  - KN smoothed 4-gram estimated over the parallel text excluding UN data and monolingual data from the English Gigaword Fourth Edition
  - Zero-cutoff stupid-backoff 5-gram LM estimated over all text [8]
    - Google n-grams were included -- helpful for web text
    - Hadoop/Hfile used for n-gram count retrieval & extraction

[8] Brants et al. Large Language Models in Machine Translation. EMNLP07

HiFST Decoder / LMBR decoding

- HiFST decoder [5] based on the Google OpenFST toolkit [7]
- Each grammar (for each source segmentation) was used to generate a separate translation lattice (with the common English LM)
- Lattice MBR [8] can be used to merge hypothesis from multiple analyses
  - Each lattice represents a posterior distribution over translation hypotheses for that translation grammar
  - Posteriors are interpolated efficiently via WFST operations [9]

[5] de Gispert et al. Hierarchical Phrase-Based Translation with Weighted Finite-State Transducers and Shallow-N Grammars. Computational Linguistics, 2010.

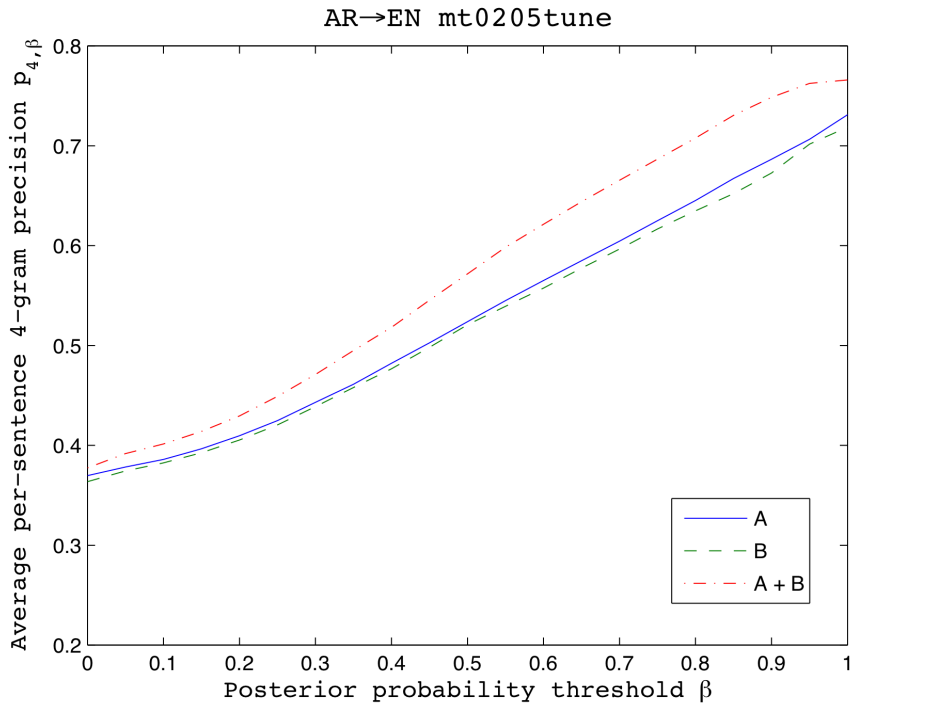
[7] Allauzen et al. OpenFst: A General and Efficient Weighted Finite-State Transducer Library. CIAA 2007

[8] Tromble et al. Lattice Minimum Bayes-Risk Decoding for Statistical Machine Translation. EMNLP 2008

[9] Blackwood et al. Efficient path counting transducers for minimum Bayes risk decoding of statistical machine translation lattices. ACL'10

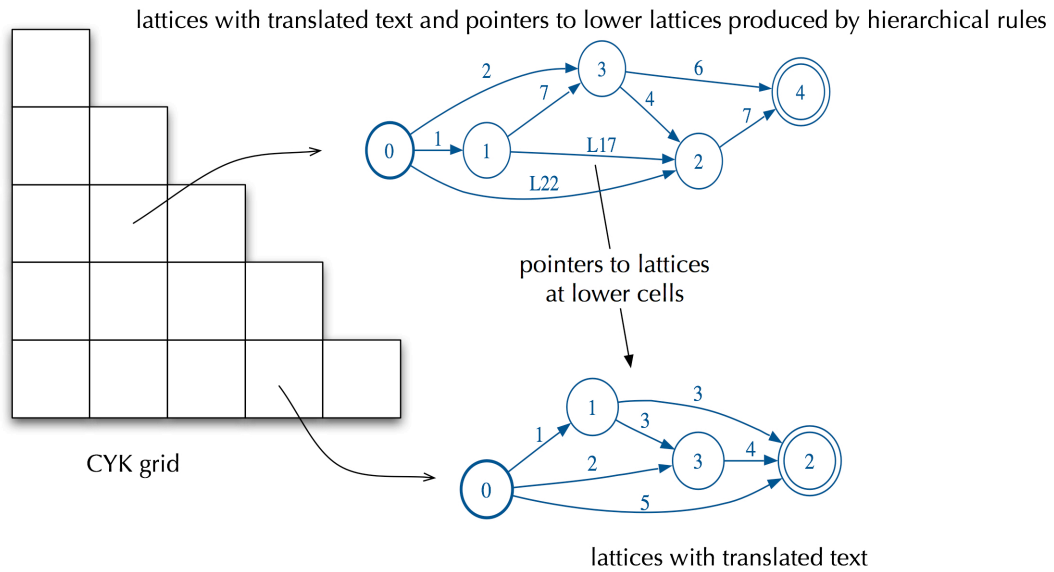
HiFST as a generator of statistics for LMBR

- LMBR consistently gives +0.5-1.0 BLEU over shortest-path hypothesis
  - Gains are even greater for lattices from multiple grammars
- Why: Lattice posteriors can predict translation quality of n-grams in hypotheses



What didn't work (yet): Challenges in using more powerful grammars

- HiFST works very well for grammars with relatively few non-terminals
  - Richer grammars such as SAMT, GHKM, make it difficult to determinise the hypothesis space during search, which is crucial for good performance



  - Alternative decoder architectures are showing some promise [10]

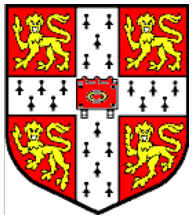
[10] Iglesias et al. Hierarchical Phrase-based Translation Representations. EMNLP'11

Summary

- Flexible grammar configuration strategies
  - Full Hiero for Zh-En
  - Shallow-N for Ar-En
  - Many features, including provenance
- Hybrid translation systems
  - Good gains from lattice MBR based on multiple source language analyses
    - Adds robustness for noisy source text
- Emphasis on efficient construction of translation search spaces
  - Minimal pruning
  - Goal: few search errors
- Lattice MBR provides a straightforward and robust method to combine hypotheses from multiple translation grammars

Acknowledgements

- Graeme Blackwood: initial Pro implementation and the LMBR tools
- Stanford CoreNLP suite, SRI LM Tools, MADA
- The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7-ICT-2009-4) under grant agreement number 247762 (FAUST project)
- R. Waite, J. Pino supported by EPSRC(UK) Doctoral Training Awards
- DARPA GALE and BOLT programs



Department of Engineering
University of Cambridge