The CUED NIST 2008 Arabic-English SMT system

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

Department of Engineering
University of Cambridge

NIST Open Machine Translation Evaluation Workshop

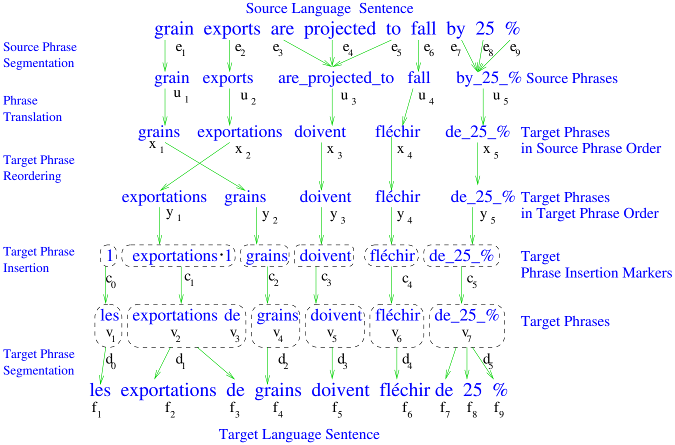
27-28 Mar 2008

CUED general system overview

- The CUED is a phrase-based SMT system following the Transducer Translation Model (TTM)
- Generative model of translation
- Implemented with Weighted Finite State Transducers (WFST)
  - WFSTs used for word alignment, language model, word-to-phrase segmentation, phrase translation and reordering
  - Translation is performed using libraries of standard FST operations
  - No special-purpose decoder required
  - Modularity. Easy to work on translation components in isolation
  - Open Source WFST Toolkit [1] – www.openfst.org/

[1] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut , and M. Mohri (2007), OpenFst: A General and Ecient Weighted Finite-State Transducer Library. CIAA.

Transducer Translation Model (TTM)



- Transformations via stochastic models implemented as WFSTs
- Built with standard WFST operations such as composition and best-path search

TTM Component Models

Basic models:

- Source first-pass language model G
- Source phrase segmentation (unweighted) W
- Phrase translation and reordering R
- Target phrase insertion Φ
- Target phrase segmentation (unweighted) Ω
- Word penalty and phrase penalty

$$\mathcal{T} = G \circ W \circ R \circ \Phi \circ \Omega$$

Additional models for MET:

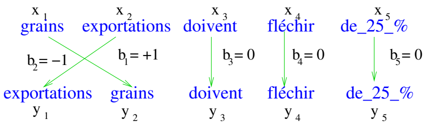
- Inverse phrase translation
- 3 phrase pair count features [2]

$\Rightarrow$ Minimum Error Training to find optimal model weights (10 factors)

- weights are assigned to WFST likelihoods

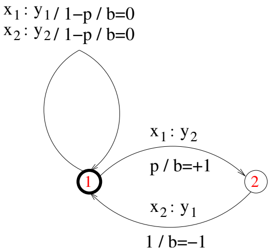
[2] O. Bender, E. Matusov, S. Hahn, S. Hasan, S. Khadivi, and H. Ney (2007). The RWTH Arabic-to-English Spoken Language Translation System. ASRU.

Phrase Swapping by WFSTs [3]



Associate a jump sequence b_1^K with each sequence y_1^K

$$P(b_1^K | x_1^K, u_1^K, K, e_1^I) = \prod_{k=1}^K \underbrace{P(b_k | b_{k-1}, x_{k-1}, x_k, u_{k-1}, u_k)}_{\text{orientation prob., estimated from alignments}}$$



b_k specify relative offsets

MJ-1 : maximum jump of 1

$$b \in \{0, +1, -1\}$$

Extremely simple, but

→ Properly parameterized

→ Not degenerate

[3] Kumar , Byrne 2005. Local phrase reordering models for statistical machine translation. HLT-EMNLP.

Data Preprocessing and Word Alignment

- All allowed Arabic-English Parallel corpus
- All allowed English LM data
- Arabic morphological word decomposition:
  - Split prefixes with MADA Toolkit [4] $\longrightarrow$ 30% vocabulary reduction
  - Remain as separate tokens in input
- Word Alignment using MTTK Toolkit [5]. Supports:
  - IBM Model-1 and Model-2
  - Word-to-Word HMMs
  - Word-to-Phrase HMMs, with bigram translation probabilities
- Standard phrase extraction from union alignments

[4] N. Habash and F. Sadat (2006). Arabic Preprocessing Schemes for Statistical Machine Translation. HLT/NAACL

[5] Y. Deng and B. Byrne. Available at http://mi.eng.cam.ac.uk/∼wjb31/distrib/mttkv1/

Lattice Rescoring with Large Monolingual Models

Stupid backoff zero cut-off 5 gram language model [6]

- Counts are extracted beforehand from all monolingual English data
- 5-grams are extracted from first-pass lattices

$$S(e_i|e_{i-n+1}^{i-1}) = \begin{cases} \frac{\#(e_{i-k+1}^{i})}{\#(e_{i-k+1}^{i-1})} & \text{if } \#(e_{i-k+1}^{i}) > 0 \\ \alpha\, S(e_i|e_{i-k+2}^{i-1}) & \text{otherwise} \end{cases}$$

- exact search with OpenFST libraries in a second translation pass

Phrase Segmentation Transducers

- assign probability to sequences of English phrases
- complements word-based N-grams
- estimated from a subset of LM training data
- implemented as a WFST
- Source phrase segmentation transducer assigns first-order predictors:

$$P(u_1^K|e_1^I) = \prod_k P(u_k \mid u_{k-1}\,,\, e_1^K)$$

[6] T. Brants et al. 2007. Large Language Models in Machine Translation. EMNLP

Minimum Bayes Risk Decoding [7]

Taking the goal as BLEU maximization

- A baseline translation model to give the probabilities over translations: $P(E|F)$
- A set $\mathcal{E}$ of N-Best Translations of F
- A Loss function $L(E, E')$ that measures the the quality of E' relative to E

MBR Decoder

$$\hat{E} = \operatorname*{argmin}_{E' \in \mathcal{E}} \sum_{E \in \mathcal{E}} -L_{BLEU}(E, E')\ P(E|F)$$

$\hat{E}$ is sometimes called the 'consensus hypothesis'

- picks from the middle of the similar, relatively likely translation hypotheses
- must be done over an N-Best list

Rational is to balance estimation criteria (e.g. MLE) with translation criteria (e.g. BLEU)

[7] S. Kumar W. Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. HLT-NAACL

Translation Performance

NIST 2008 Arabic-English MT evaluation development [8]

Lowercase BLEU scores over three test sets from 2002 through 2006:

Method	mt02_05_test	mt06-nist-newswire	mt06-nist-newsgroup
	50.26	48.10	36.78
+ 5B Word SB LM	52.41	49.60	37.23
+ Phrase Seg Trans	53.32	50.07	37.37
+ MBR	53.70	50.99	37.84

- Important gains from lattice rescoring (improved fluency)

[8] nist.gov/speech/tests/mt/2006 nist.gov/speech/tests/mt/2008

Conclusion and further work

Summary (strong points):

- Phrase-based SMT system implemented with WFSTs
- Relatively good performance with models that are really quite simple
- Easy to learn, easy to modify (modularity)
- Can easily generate translation lattices and N-best lists
- Easy to apply to translation of ASR lattices

Known problems (room for improvement):

- Long Arabic phrases wrongly deleted (insertion model needs to be reviewed)
- MJ1 Reordering model does not allow long-range reordering
- Wrong capitalization for all newswire headlines
- Model 1 rescoring should be incorporated into MET

Thanks!
Questions and comments welcome.

Department of Engineering
University of Cambridge