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/

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 $\Phi$
- Target phrase segmentation (unweighted) $\Omega$
- Word penalty and phrase penalty

$$\tau = G \circ W \circ R \circ \Phi \circ \Omega$$

Additional models for MET:

- Inverse phrase translation
- 3 phrase pair count features

$\Rightarrow$ Minimum Error Training to find optimal model weights (10 factors)
- weights are assigned to WFST likelihoods

---

Phrase Swapping by WFSTs

$$\begin{align*}
x_1 & : \text{grains} & y_1 & : \text{exportations} \\
x_2 & : \text{exportations} & y_2 & : \text{grains} \\
x_3 & : \text{doivent} & y_3 & : \text{fléchir} \\
x_4 & : \text{de}_25\% & y_4 & : \text{doivent} \\
x_5 & : \text{fléchir} & y_5 & : \text{de}_25\%
\end{align*}$$

$\begin{align*}
b_2 &= -1 \\
b_1 &= +1 \\
b_3 &= 0 \\
b_4 &= 0 \\
b_5 &= 0
\end{align*}$

Associate a jump sequence $b^K_1$ with each sequence $y^K_1$

$$P(b^K_1 \mid x^K_1, u^K_1, K, e^K_1) = \prod_{k=1}^{K} P(b_k \mid b_{k-1}, x_{k-1}, x_k, u_{k-1}, u_k)$$

orientation prob., estimated from alignments

$\begin{align*}
x_1 : y_1 / 1-p / b=0 \\
x_2 : y_2 / 1-p / b=0 \\
x_1 : y_2 \\
p / b=+1 \\
x_2 : y_1 \\
1 / b=-1
\end{align*}$

$b_k$ specify relative offsets

**MJ-1**: maximum jump of 1

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

Extremely simple, but

→ Properly parameterized

→ Not degenerate

---

3Kumar, 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 \(^4\) → 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

---


\(^5\) 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

- Counts are extracted beforehand from all monolingual English data
- 5-grams are extracted from first-pass lattices
  \[
  S(e_i | e_{i-n+1}^{i-1}) = \begin{cases} 
  \frac{#(e_i^{i-k+1})}{#(e_i^{i-k+1})} & \text{if } #(e_i^{i-k+1}) > 0 \\
  \alpha S(e_i | e_{i-k+2}^{i-1}) & \text{otherwise}
  \end{cases}
  \]
- 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:
  \[
  P(u_1^K | e_1^i) = \prod_k P(u_k | u_{k-1}, e_1^K)
  \]

---

6 T. Brants et al. 2007. Large Language Models in Machine Translation. EMNLP
Minimum Bayes Risk Decoding

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 quality of \( E' \) relative to \( E \)

MBR Decoder

\[
\hat{E} = \arg\min_{E' \in \mathcal{E}} \sum_{E \in \mathcal{E}} -L_{\text{BLEU}}(E, E') P(E|F)
\]

\( \hat{E} \) 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)

---

Translation Performance

NIST 2008 Arabic-English MT evaluation development

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

<table>
<thead>
<tr>
<th>Method</th>
<th>mt02_05_test</th>
<th>mt06-nist-newswire</th>
<th>mt06-nist-newsgroup</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>50.26</td>
<td>48.10</td>
<td>36.78</td>
</tr>
<tr>
<td>+ 5B Word SB LM</td>
<td>52.41</td>
<td>49.60</td>
<td>37.23</td>
</tr>
<tr>
<td>+ Phrase Seg Trans</td>
<td>53.32</td>
<td>50.07</td>
<td>37.37</td>
</tr>
<tr>
<td>+ MBR</td>
<td>53.70</td>
<td>50.99</td>
<td>37.84</td>
</tr>
</tbody>
</table>

► Important gains from lattice rescoring (improved fluency)

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