Machine Translation Panel

Nadir Durrani, University of Edinburgh: Operation Sequence Model

Chris Dyer, CMU: Word Classes

Spence Green, Stanford: Sparse Feature Training

Kenneth Heafield, Stanford: Huge Language Models

Stephan Peitz, RWTH Aachen: Leave One Out Training

Philip Williams, University of Edinburgh: String-to-Tree Syntax
Nadir Durrani
University of Edinburgh

Operation Sequence Model
Introduction

- A model that
  - combines benefits from Phrase-based and N-gram-based SMT
  - is based on minimal translation but memorizes like phrases
  - considers source and target contextual information across phrases
  - integrates translation and reordering into a single model

- Convert a bilingual sentence to a sequence of operations
  - Translate (Generate a minimal translation unit)
  - Reordering (Insert a gap or Jump)

- \( P(e,f,a) = \text{N-gram model over resulting operation sequences} \)
Example

Sie würden gegen Sie stimmen

Operations

- $o_1$ Generate (Sie, They)
- $o_2$ Generate (würden, would)
- $o_3$ Insert Gap
- $o_4$ Generate (stimmen, vote)
- $o_5$ Jump Back (1)
- $o_6$ Generate (gegen, against)
- $o_7$ Generate (Sie, you)

Context Window

Model:

$$p_{osm}(F, E, A) = p(o_1 \ldots, o_N) = \prod_i p(o_i | o_{i-n+1} \ldots o_{i-1})$$
How does it improve Phrase-based SMT?

• Overcomes phrasal independence assumption
  – Considers source and target contextual information across phrases

• Better reordering model
  – Translation and reordering decisions influence each
  – Handles local and long distance reorderings in a unified manner

• No spurious phrasal segmentation problem

• Average gain of +0.40 on news-test2013 across 10 pairs

Thank You !!!
Chris Dyer
CMU

Word Classes
Using Word Clusters

1. Cluster monolingual data
2. 500-1000 clusters
3. Use for: LMs, features
Using Word Clusters

Rule “shape” features (prefix length = 6)

\[ X \rightarrow (\text{dass } X \text{ angekommen ist, that } X \text{ arrived}) \]

7-gram class-based LM

Absolute discounting (\(d=0.5\))
Separate features for transitions and emissions

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>MET</th>
<th>TER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>25.3</td>
<td>30.4</td>
<td>52.6</td>
</tr>
<tr>
<td>+Rule shape</td>
<td>25.5</td>
<td>30.5</td>
<td>52.4</td>
</tr>
<tr>
<td>+7gm LM</td>
<td>26.4</td>
<td>31.0</td>
<td>51.9</td>
</tr>
</tbody>
</table>
Sparse Feature Training
Large-scale Discriminative Tuning

#1: 2010s ML in MT tuning
  Online convex optimization
  Arbitrary, overlapping features

#2: Large tuning sets
  Fast decoding and updating
  Bitext tuning...

See our poster and talk for details
WMT14 Shared Task Results

Uncased BLEU results

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>dense–dev</th>
<th>features–dev</th>
<th>2014 rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fr–En</td>
<td>19.6</td>
<td>20.0</td>
<td>1</td>
</tr>
<tr>
<td>En–De</td>
<td>32.0</td>
<td>32.5</td>
<td>1</td>
</tr>
</tbody>
</table>

Tune: 13.5k sentences (2008–2012)

Models have 200–300k features
Kenneth Heafield
Stanford

Huge Language Models
## Impact of Big Language Models

<table>
<thead>
<tr>
<th>Target</th>
<th>Base Rank</th>
<th>+LM Rank</th>
<th>ΔBLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Czech</td>
<td>5–6</td>
<td>1–3</td>
<td>+0.6</td>
</tr>
<tr>
<td>Hindi</td>
<td>4–5</td>
<td>3</td>
<td>+1.4</td>
</tr>
<tr>
<td>Russian</td>
<td>6–7</td>
<td>4–5</td>
<td>+1.2</td>
</tr>
<tr>
<td>German</td>
<td>8–10</td>
<td>3–6</td>
<td>+0.5</td>
</tr>
</tbody>
</table>

After the evaluation: Hindi–English +0.9 BLEU
Download multiple LMs and training data from statmt.org/ngrams

English: 1.8 trillion tokens

<table>
<thead>
<tr>
<th>n</th>
<th>Unique n-grams</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2,640,258,088</td>
</tr>
<tr>
<td>2</td>
<td>15,297,753,348</td>
</tr>
<tr>
<td>3</td>
<td>61,858,786,129</td>
</tr>
<tr>
<td>4</td>
<td>156,775,272,110</td>
</tr>
<tr>
<td>5</td>
<td>263,690,452,834</td>
</tr>
</tbody>
</table>

Current work: approximate LM storage.
Stephan Peitz
RWTH Aachen

Leave One Out Training
Consistent Phrase Training

State of the art

- Heuristic extraction of phrases using word alignments
- Compute translation probabilities as relative frequencies

Issues of this heuristic

- Extract from likely alignment?
- Models used in decoding are not considered $\Rightarrow$ inconsistency

Forced decoding

- Run decoder on training data
- Count used phrases, recompute probabilities
- Apply leave-one-out to counteract overfitting
Occurrences of a phrase in a sentence pair \((f_n, e_n)\) are subtracted from the phrase counts obtained from the full training data:

\[
\begin{align*}
\ell_{10,n}(\tilde{f}|\tilde{e}) &= \frac{C(\tilde{f}, \tilde{e}) - C_n(\tilde{f}, \tilde{e})}{\sum_{\tilde{f}'} C(\tilde{f}', \tilde{e}) - C_n(\tilde{f}', \tilde{e})}
\end{align*}
\]

Singleton phrases get a low probability.
Consistent Phrase Training using Leave-One-Out

▶ Publications
  ▶ Phrase-based [Wuebker & Mauser+ 10, Wuebker & Ney 13]
  ▶ Hierarchical [Peitz & Mauser+ 12, Peitz & Vilar+ 14]

▶ Improvements: 0.5-1.5 BLEU

▶ Reducing phrase-table size to 5-20% of the original size

▶ Systems using phrase/rule training:
  ▶ WMT 2011 (RWTH, German→English, phrase-based)
  ▶ IWSLT 2011 (RWTH, German→English, phrase-based)
  ▶ IWSLT 2012 (RWTH, German→English, hierarchical)
  ▶ BOLT 2012 (RWTH, Chinese→English, hierarchical)
  ▶ OpenMT 2012 (NRC, Chinese→English, phrase-based)

▶ Implemented in RWTH’s translation toolkit Jane
http://www.hltpr.rwth-aachen.de/jane
References


Philip Williams
University of Edinburgh

String-to-Tree Syntax
das Protokoll der letzten Sitzung wurde verteilt.

the minutes of yesterday's sitting have been distributed.

TOP → the X₁ X₂ have been X₃ X₄ | das NN₁ NP-AG₂ wurde VP-OC₃ PUNC₄
Use syntactic structure to help model other aspects of target-side grammar.

Example 1. Agreement

\[
\text{TOP} \rightarrow \text{the } X_1 X_2 \text{ have been } X_3 X_4 \mid \text{das } \langle \text{NN}_1 \text{ NP-AG}_2 \text{ wurde VP-OC}_3 \text{ PUNC}_4 \rangle
\]

\[
\langle \text{NN}_1 \text{ AGR} \rangle = \langle \text{wurde AGR} \rangle
\]

Example 2. Compound Splitting

Best constrained system for:
English-German
German-English
Hindi-English (tied with CMU)