Online Language Model Adaptation via N-gram Mixtures for Statistical Machine Translation

Germán Sanchis-Trilles
Instituto Tecnológico de Informática, Universidad Politécnica de Valencia, Spain

Mauro Cettolo
FBK - Ricerca Scientifica e Tecnologica, Trento, Italy

Saint-Raphaël, May 27-28, 2010
Outline

• Introduction
• Model adaptation
• Experiments
• Future work
• Conclusions
Introduction

- Aimed towards introducing more context in the system
- Key idea: enhance target LM by introducing parameters that are adapted to the input text
- LM is implemented as mixture of sub LMs
- Experiments on Europarl v2 task (WMT06)
Model adaptation

• Most usual translation rule:

\[ e^* = \underset{e}{\arg\max} \sum_{r=1}^{R} \lambda_r h_r(e, f) \]

• LM can be computed either as a single LM or as a mixture of LMs, i.e.:

\[ p(e) = \sum_{i=1}^{M} w_i p_i(e) \]
→ Assume a partition of the parallel training data into M bilingual clusters
→ Train specific source/target LMs for each partition
→ Before translation, estimate the optimal weights of source LMs via EM
→ Transfer the resulting weights to the target LM mixture
Assume a partition of the parallel training data into M bilingual clusters
→ Train specific source/target LMs for each partition
→ Before translation, estimate the optimal weights of source LMs via EM
→ Transfer the resulting weights to the target LM mixture
Model adaptation: clustering

• Goal: group similar sentences from the lexical point of view
• Sentence pair represented as bag of source and target words
• CLUTO package used, direct $k$-way partitioning and cosine distance
• Number of clusters set to 4 according to preliminary investigation
• Additional LM built on the whole training data

⇒ First clustering approach: direct clustering of training data
Clustering: Development-induced

- Adaptation: cover mismatches between training and development/test
  \[\rightarrow\] direct clustering may not be the best choice

\[\Rightarrow\] Cluster development set and mirror it on training data
  1. Cluster bilingual development set
  2. Estimate source and target LMs for each cluster
  3. For each training sentence:
     - Compute best interpolation of cluster-LMs, in source and target sides
     - Classify it according to most-weighted LMs

- Intuitively:
  - LM is a compact representation of the cluster
  - weights in the optimization provide a measure of similarity
Clustering: Development-induced
Clustering: Test-induced

- Test data can be used to induce the clusterings
  - Target side is not available
  - Only relies on source data, but used to classify both sides!
  - May not lead to reliable benefits
  - Take advantage of information of the actual test
  - Clustering performed only on source data, analogously as for dev-induced
→ Assume a partition of the parallel training data into $M$ bilingual clusters
→ Train specific source/target LMs for each partition
→ **Before translation, estimate the optimal weights of source LMs via EM**
→ Transfer the resulting weights to the target LM mixture
On-line weight optimization

Three different approaches:

a) Set specific weights

b) Sentence specific weights

c) Two-steps weight estimation
On-line weight optimization

Three different approaches:

a) Set specific weights:
   ∗ LM weights estimated on the source side of the complete test set
     + Straightforward
     − Does not consider differences between sentences
       ⇒ benefit of approach may fade
On-line weight optimization

Three different approaches:

b) Sentence specific weights:
   ∗ One set of weights for each sentence in the test set

   + EM procedure allowed complete freedom
   − Weights estimated on few data
   ⇒ possibly, not very reliable weights
On-line weight optimization

Three different approaches:

c) Two-step weight estimation:
   1. Estimate sentence-specific weights
   2. Assign each source sentence to the cluster with the most weighted LM
   3. Re-estimate one single set of weights for each of such clusters
   + Mirror the clustering of the training data into the test set
   + Avoid possible data sparseness issues

Sanchis-Trilles and Cettolo
Online LM adaptation
Saint-Raphaël, May 27-28, 2010
Experiments: Corpora

- Experiments conducted on the Europarl corpus (setup of WMT06)
- Consists of transcription of European Parliament speeches
- Experiments conducted on De–En, Es–En and Fr–En, both directions

<table>
<thead>
<tr>
<th></th>
<th>De</th>
<th>En</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentences</td>
<td>751K</td>
<td>2000</td>
</tr>
<tr>
<td>Run. words</td>
<td>15.3M</td>
<td>55K</td>
</tr>
<tr>
<td>Voc.</td>
<td>195K</td>
<td>432</td>
</tr>
<tr>
<td></td>
<td>16.1M</td>
<td>59K</td>
</tr>
<tr>
<td></td>
<td>66K</td>
<td>125</td>
</tr>
<tr>
<td>Dev.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentences</td>
<td>2000</td>
<td></td>
</tr>
<tr>
<td>Run. words</td>
<td>55K</td>
<td></td>
</tr>
<tr>
<td>OoV</td>
<td>432</td>
<td>125</td>
</tr>
<tr>
<td>Test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentences</td>
<td>2000</td>
<td></td>
</tr>
<tr>
<td>Run. words</td>
<td>54K</td>
<td></td>
</tr>
<tr>
<td>OoV</td>
<td>377</td>
<td>127</td>
</tr>
</tbody>
</table>
Experiments: Baseline system

- Built upon Moses SMT toolkit. Log-linear model with
  - Phrase-based translation model
  - Language model
  - Word and phrase penalties
  - Distortion model
- Weights of the log-linear combination optimized with MERT
- Language model: 5-gram with KN smoothing
- Distortion model: "orientation-bidirectional-fe"
# Experiments

- 10K bootstrap repetitions, 95% confidence level pairwise improvement

<table>
<thead>
<tr>
<th>Clustering method</th>
<th>Weight optimization</th>
<th>BLEU</th>
<th>TER</th>
<th>Signif BLEU/TER</th>
</tr>
</thead>
<tbody>
<tr>
<td>—</td>
<td>baseline</td>
<td>19.0</td>
<td>67.4</td>
<td>—</td>
</tr>
<tr>
<td>direct</td>
<td>sentence</td>
<td>18.2</td>
<td>67.4</td>
<td>yes/no</td>
</tr>
<tr>
<td></td>
<td>two-steps</td>
<td>18.1</td>
<td>67.4</td>
<td>yes/no</td>
</tr>
<tr>
<td></td>
<td>test set</td>
<td>18.0</td>
<td>67.6</td>
<td>yes/no</td>
</tr>
<tr>
<td>dev-induced</td>
<td>sentence</td>
<td>19.2</td>
<td>66.7</td>
<td>yes/yes</td>
</tr>
<tr>
<td></td>
<td>two-steps</td>
<td>19.2</td>
<td>66.7</td>
<td>yes/yes</td>
</tr>
<tr>
<td></td>
<td>test set</td>
<td>18.7</td>
<td>67.2</td>
<td>yes/no</td>
</tr>
<tr>
<td>test-induced</td>
<td>sentence</td>
<td>18.9</td>
<td>67.3</td>
<td>no/no</td>
</tr>
<tr>
<td></td>
<td>two-steps</td>
<td>18.9</td>
<td>67.3</td>
<td>no/no</td>
</tr>
<tr>
<td></td>
<td>test set</td>
<td>18.9</td>
<td>67.1</td>
<td>no/yes</td>
</tr>
</tbody>
</table>
General remarks

- Best results achieved when using:
  - development-induced clustering
  - two-steps (or sentence-based) weight optimization
- Results found to be statistically significant and coherent
- sentence and two-steps weighting schemes yield similar results
  → For long sentences, sentence is best (cheaper)
- Test and development sets are extracted from a narrow time frame
  → development-induced clustering exploits un-even distribution of data better
- Test clustering relies on monolingual data
  → Much less information for clustering (less than half of it!)
Conclusions

- Technique for adapting the LM of SMT systems to actual input
- LM is assumed to be provided as a linear interpolation of sub-LMs
- Weights are estimated dynamically on the text to be translated
- Best results by:
  - Exploiting both source and target of the development set
  - Weight estimation at sentence level or two-steps approach
- Such results yield consistent improvements over the reference baseline
Future work

- Results achieved depend on the clustering technique employed
  → Clustering based on $n$-grams or PoS-tag information
- Supervised clustering
  → Detailed supervision is available only for limited amount of data
- Learn source-to-target weight mapping schemes from parallel data
Questions? Comments? Suggestions?