Online Language Model Adaptation for Spoken Dialog Translation

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

Mauro Cettolo, Nicola Bertoldi, Marcello Federico
FBK - Ricerca Scientifica e Tecnologica, Trento, Italy

Tokyo, Dec 1-2, 2009
Outline

- Introduction
- Model adaptation
- Experiments
- Future work
- Conclusions
Introduction

- Spoken language translation
- 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 IWSLT 2009 CT task, CRR conditions
Model adaptation

• Most usual translation rule:

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

• 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 the source LMs via EM
→ Transfer the resulting weights to the target LM mixture
IWSLT Data

- Experiments carried out on the CT task (both CE and EC)
- We considered the use of Agent, Customer and Interpreter annotations
- We also considered the use of the Dialog tags

**Speaker-based statistics of the CT data**

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Training</th>
<th>Development</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>agent</td>
<td></td>
<td></td>
</tr>
<tr>
<td>native</td>
<td>46.7K</td>
<td>2240</td>
</tr>
<tr>
<td>interpreter</td>
<td>26.8K</td>
<td>1626</td>
</tr>
<tr>
<td>customer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>native</td>
<td>33.3K</td>
<td>2082</td>
</tr>
<tr>
<td>interpreter</td>
<td>33.8K</td>
<td>1878</td>
</tr>
</tbody>
</table>
Nespole! data

- NEgotiating through SPOken Language in E-commerce
- Collected involving Italian speakers, translated into English

Statistics of the Nespole! dialogs.

| #turns | | |  
|---|---|---|---|
| 2522 | 15335 | 1344 | 6.1 |

Most frequent Nespole! dialog acts.

<table>
<thead>
<tr>
<th>label</th>
<th>counter</th>
</tr>
</thead>
<tbody>
<tr>
<td>give-information</td>
<td>963</td>
</tr>
<tr>
<td>affirm</td>
<td>408</td>
</tr>
<tr>
<td>descriptive</td>
<td>285</td>
</tr>
<tr>
<td>request-information</td>
<td>199</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>2522</td>
</tr>
</tbody>
</table>
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"
Model adaptation

TRAINING PARALLEL TEXTS

SRC TGT

CLUSTERING

CLSTR₁ CLSTR₂ CLSTRₙ

OPTIMIZATION of SRC LMs

INTERPOLATION of TGT LMs

INTERPOLATION of TGT LMs

LM ESTIMATION

SRC TGT

LM₁ LM₂ ⋮ LMₙ

LM₁ LM₂ ⋮ LMₙ

OFF-LINE

ON-LINE

TRANSITION

Sanchis-Trilles et. al. Online LM adaptation Tokyo, Dec 1-2, 2009
Clustering: IWSLT

- Dialog based
  - Consider each dialog as a bag of source and target words
  - Compute 2, 4, 6 and 8 clusters by means of CLUTO
    * direct clustering algorithm
    * cosine distance
  - Additional LM for BTEC+CT data

- Speaker based
  - Specific clusters for native agent/customer, and interpreter agent/customer
  - Additional LMs for BTEC and BTEC+CT data
Clustering: Nespole!

- Three LMs estimated on (English) Nespole! data:
  - give-information
  - request-information
  - other
- Such LMs are used to partition the IWSLT data on the basis of perplexity
- The clusters are mirrored on the Chinese side
- New LMs were trained on the IWSLT clusters
- Additional LM for all the BTEC+CT data
Model adaptation

TRAINING PARALLEL TEXTS

SRC   TGT

CLUSTERING

CLSTR₁   CLSTR₂   ...   CLSTRₘ

SRC   TGT

LM ESTIMATION

LM₁   LM₂   ...   LMₘ

SRC   TGT

ON-LINE

OFF-LINE

OPTIMIZATION of SRC LMs

INTERPOLATION of TGT LMs

SMT

TRANSLATION

Sanchis-Trilles et. al.  Online LM adaptation  Tokyo, Dec 1-2, 2009
On-line weight optimization

Four different approaches:

- 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

Four different approaches:

- **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, less reliable weights
On-line weight optimization

Four different approaches:

- **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
On-line weight optimization

Four different approaches:

- **Oracle weight estimation:**
  - Estimate weights at sentence level on the reference texts (i.e. target side)
  - Provides a sort of upper bound
  - Not fair
Results

Results for sentence-based weight estimation

Sanchis-Trilles et. al.  
Online LM adaptation  
Tokyo, Dec 1-2, 2009
Results

Results for two-step weight estimation

en-zh TEST: DEV2

BLEU (%) vs. number of classes

Baseline
Dialog
nespole
ACI
Oracle

PP vs. number of classes

Baseline
Dialog
nespole
ACI
Oracle
Analysis

- Significant improvements are achieved in terms of perplexity for every setup
- Improvements in perplexity are not always mirrored by BLEU
- Oracle curves are unimodal with peak at six clusters
- Oracle setup confirms that the approach is appealing, room for improvement
- Two-step: does not improve sentence-based, but curves are unimodal
  → more predictable
- Dialog clustering improves or is as good as baseline:
  + two-step: seems to guarantee stable improvements
- Nespole! guided clustering does not seem to be effective
- Clustering according to ACI labels works well for EC (not for CE)
Analysis

- Training/development and test conditions are quite different

<table>
<thead>
<tr>
<th></th>
<th>test on</th>
<th>mert on</th>
<th>Δ BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEV1</td>
<td>DEV2</td>
<td>-0.19</td>
<td>+3.39</td>
</tr>
<tr>
<td>DEV2</td>
<td>DEV1</td>
<td>-0.67</td>
<td>-1.12</td>
</tr>
</tbody>
</table>

- Clustering according to ACI labels produces speaker-specific LMs.
  → According to training!
  → This is bound to have an important effect
Future work

- Obtain data partitioning in an unsupervised manner
  - Surface form
  - PoS
  - ...
- Perform development/test-driven partitioning of the training data
- Source-to-target weight mapping
- Assess these techniques on larger tasks such as Europarl or NIST
Questions? Comments? Suggestions?