Domain Adaptation in SMT using Factored Translation Models

Jan Niehues and Alex Waibel
Overview

- Motivation

- Related Work

- Factored Domain Model
  - Domain Factors Translation Model
  - Domain Factors Sequence Model

- Evaluation
  - News Task
  - Lecture Task

- Conclusion
Motivation

- Large amounts of training data are needed for SMT systems

- Best to have data from similar topics and genre
  - Possible only for few scenarios
    - European Parliament
  - Not possible for many real-world scenarios
  - Example:
    - Lecture Translation
    - Even News for some languages

- Common technique:
  - Use all available data to build a baseline system
  - Adapt system using in-domain data
Motivation

- How to adapt the system?

- State-of-the-art SMT Systems:
  - Assumption: All training sentences are equally important
  - No longer holds if we have in-domain and out-of-domain data
  - Leads especially to many errors if in-domain data is small

- In-domain data should be more important
  - Introduce sentence weights into SMT model
Motivation

- Model domain of the training data explicitly
  - Integrate corpus identifier into the translation model

- Prefer phrase pairs learned from in-domain data
  - Weights can be tuned automatically

- Integration using Factored Translation Models (Koehn and Hoang (2007))
  - Easy to integrate into state-of-the-art SMT systems
Related Work

- Only monolingual in-domain data
  - Language Model Adaption
    - Inspired by work, that was done for ASR
  - Creating synthetic parallel text
    - Translate monolingual text using a baseline system
    - Use translated text as additional training data
    - Ueffing et al. 2007, Schwenk and Senellart 2009
Related Work

- Only Monolingual in-domain data

- Parallel in-domain data
  - Combine translation models using alternate decoding paths
    - (Koehn and Schroeder (2007))
  - Adapt translation models using mixture models
    - (Foster and Kuhn (2007))
    - Linear and log-linear combination
    - Different methods to set weights for domain
  - Discriminative weights for sentences of parallel corpus
    - (Matsoukas et al. 2009)
Related Work

- Only Monolingual in-domain data
- Parallel in-domain data

Data selection
- Select similar sentences using (cross-lingual) information retrieval techniques
  - Hildebrand et al. 2005
  - Snover et al. 2008
Factored Translation Model

- Framework to integrate corpus id

- Represent words by vector of factors instead of token
  - Integrate additional annotation into SMT

- mainly used to incorporate additional linguistic knowledge

<table>
<thead>
<tr>
<th>Source</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word</td>
<td>Word</td>
</tr>
<tr>
<td>Part-of-speech</td>
<td>Part-of-speech</td>
</tr>
<tr>
<td>lemma</td>
<td>lemma</td>
</tr>
</tbody>
</table>
Factored Domain Model

- Model Domain of data explicitly
  - Directly model influence of in-domain and out-of-domain data
  - Optimize weights on development data

- Representation of domain
  - Introduce corpus identifier for different training sources

- Assumption: Phrase pairs extracted from in-domain data are more important
Factored Domain Model

Integration into SMT system:
- Use corpus id as an additional target factor

Translation differ by generated target words and domains of these words
Factored Domain Model

- Domain representation during translation
  - Sequence of corpus ids
  - Example:

<table>
<thead>
<tr>
<th>Ein</th>
<th>blauer</th>
<th>Bogen</th>
<th>( demokratischer )</th>
<th>Staaten im Osten</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>blue</td>
<td>arc</td>
<td>( democratic )</td>
<td>states in the east</td>
</tr>
<tr>
<td>IN</td>
<td>OUT</td>
<td>IN</td>
<td>IN IN IN</td>
<td>IN IN IN IN</td>
</tr>
</tbody>
</table>
Factored Domain Model

- Domain representation during translation
  - Sequence of corpus ids
  - Example:

  Assumption: Prefer translation with more corpus ids from text similar to test domain
Factored Domain Model

- Phrase pairs that occur in both corpus
  - Example:
    - Bogen # arc # IN
    - Bogen # arc # OUT

- Lead to different phrase pair
  - Existing phrase pair scores are the same
  - New model scores are different

- Select best one according to current weights
Factored Domain Model

- Describe probability of the domain by two additional models

  Domain Factors Translation Model:
  - Probability of generating a sequence of corpus id tags given the sentence
  - Example:
    - im Osten # in Eastern Europe # IN -> quite low probability
    - im Osten # in Eastern Europe # OUT -> higher probability

  Domain Factors Sequence Model:
  - How probable is a sequence of corpus id tags
  - Example:
    - IN OUT IN IN IN IN IN IN IN IN -> high probability
    - IN OUT OUT IN IN OUT OUT OUT OUT OUT OUT -> lower probability
Domain Factors Translation Model

- Probability of generating a sequence of corpus id tags given the sentence
  - Similar to phrase translation model in state-of-the-art SMT approach
  - Modeled using 2 two scores

\[
P(t \mid s) = \frac{\text{cooc}(s,t)}{\text{cooc}(s,*)}
\]

\[
P(s \mid t) = \frac{\text{cooc}(s,t)}{\text{cooc}(*,t)}
\]

- Estimated using cooccurrence counts \( \text{cooc}(s,t) \)
Domain Factors Translation Model

- Cooccurrence count depending on three parameters
  - Use $cooc(s,t,d)$ instead of $cooc(s,t)$

- Leads to 3 different probabilities

- Domain Frequency:
  - Probability of the domain tags given the phrase pair
  - Can be approximated by:

$$P(d \mid s,t) = \frac{cooc(s,t,d)}{cooc(s,t,*)}$$
Domain Factors Translation Model

- **Target Frequency:**
  - Probability of the target phrase given source and domain sequence

\[ P(t \mid s,d) = \frac{cooc(s,t,d)}{cooc(s,*,d)} \]

- Equal to:
  - Target Translation Probability restricted to phrase pairs extracted only from the one domain

- **Source Frequency:**
  - Probability of the source phrase given target and domain sequence

\[ P(s \mid t,d) = \frac{cooc(s,t,d)}{cooc(*,t,d)} \]
Domain Factors Sequence Model

- Describe probability of a sequence of corpus id tags

- Similar to language model

- Problem: cannot be trained on training data
  - Training sentences are always from one domain

- Use discriminative uni-gram model
Domain Factors Sequence Model

- **Word Count model**
  - Number of target words translated by in-domain phrase pairs

- **Phrase Count model**
  - Number of phrase pairs extracted from the in-domain corpus

- Several different corpora:
  - One feature for every corpus id
  - Example:
    - ID1 ID2 ID1 ID3 ID1 ID3 -> (3 1 2)
Evaluation

Two task for translating from German to English

- Translation of News-commentary texts
  - Test set: News test set of WMT 2007
  - Out-of-domain Data: European Parliament (ca. 33 M Words)
  - In-domain Data: News commentary (ca. 1 M Words)

- Translation of university Lectures
  - Test set: Different lectures
  - Data: European Parliament, BTEC and News Commentary
  - In-domain Data: Lecture 200K Words
Evaluation

- Preprocessing:
  - Normalization of different German writing systems
  - Compound Splitting

- Discriminative Word Alignment

- Phrase extraction according to moses scripts
  - Additional smoothing of relative frequencies

- POS-based reordering for short and long-range reorderings
  - Rules learned from word-aligned corpus
  - Different reorderings encoded in lattice
Domain Factors Sequence Models

- News translation task
- Domain Factors Translation Model:
  - Domain Relative Frequencies

<table>
<thead>
<tr>
<th>System</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25.90</td>
<td>29.03</td>
</tr>
<tr>
<td>2</td>
<td>26.68</td>
<td>29.24</td>
</tr>
<tr>
<td>3</td>
<td>26.80</td>
<td>29.21</td>
</tr>
<tr>
<td>4</td>
<td>27.03</td>
<td>29.63</td>
</tr>
<tr>
<td>5</td>
<td>27.09</td>
<td>29.54</td>
</tr>
</tbody>
</table>
Domain Factors Translation Models

- News translation task
- Domain Factors Sequence Model:
  - Word Count Model

<table>
<thead>
<tr>
<th>System</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Baseline</td>
<td>25.90</td>
<td>29.03</td>
</tr>
<tr>
<td>2 (1) + LM Adaptation</td>
<td>26.68</td>
<td>29.24</td>
</tr>
<tr>
<td>3 (2) + Word Count Model</td>
<td>26.13</td>
<td>29.17</td>
</tr>
<tr>
<td>4 (3) + Domain Frequency</td>
<td>27.03</td>
<td>29.63</td>
</tr>
<tr>
<td>5 (3) + Target Frequency</td>
<td>27.00</td>
<td>29.51</td>
</tr>
<tr>
<td>6 (3) + Source Frequency</td>
<td>26.95</td>
<td>29.84</td>
</tr>
<tr>
<td>7 (3) + All</td>
<td>27.07</td>
<td>29.69</td>
</tr>
</tbody>
</table>
Lecture Task

- Domain Factors Sequence Model:
  - Word Count Model

<table>
<thead>
<tr>
<th>System</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 LM Adaptation</td>
<td>36.93</td>
<td>29.84</td>
</tr>
<tr>
<td>2 (1) + Source Frequency</td>
<td>37.90</td>
<td>31.12</td>
</tr>
<tr>
<td>3 (1) + Target Frequency</td>
<td>37.63</td>
<td>30.73</td>
</tr>
<tr>
<td>4 (1) + Domain Frequency</td>
<td>37.28</td>
<td>30.16</td>
</tr>
<tr>
<td>5 (1) + All</td>
<td>37.74</td>
<td>31.53</td>
</tr>
<tr>
<td>6 (2) + All Sep. corpus ids</td>
<td>38.01</td>
<td>31.51</td>
</tr>
</tbody>
</table>
Example Translations

- **Input:**
  - Ein blauer Bogen (demokratischer) Staaten im Osten, ...

- **Reference:**
  - An arc of blue (Democratic) states in the East, ...

- **Baseline:**
  - A blue sheet (democratic) countries in Eastern Europe, ...

- **Adapted:**
  - A blue arc (democratic) states in the east, …
Conclusion

- New approach to adapt phrase-based SMT systems using Factored Translation Models
  - Easy to integrate

- Model domain of training corpus explicitly
  - Introduce corpus id
  - Add two types of features to log-linear model
  - Weights can be optimized using MERT

- Translation performance could be improved by up to 1 BLEU point