UEdin: Translating L1 Phrases in L2 Context using Context-Sensitive SMT

Eva Hasler

School of Informatics
University of Edinburgh

SemEval 2014
SemEval Task 5: L2 writing assistant

Language pairs

- English-Spanish, French-English, Dutch-English

Proposed system

- Phrase-based SMT: L1 → L2
- Language model scoring of translated fragments in target context
- Context similarity feature adapts to topic of each context
- No explicit modelling of grammatical correctness
Language model scoring of L2 context

- Provide target context as identity translation of source text around L1 fragment
Language model scoring of L2 context

- Provide target context as identity translation of source text around L1 fragment

Example test instance (L1: French, L2: English)

\(<f>\text{les manifesteurs}\</f>\) want to replace the government.

Input to MT system

\(<\text{wall/}>\)

\(\text{les manifesteurs}\)

\(<\text{wall/}>\)

\(<t \text{ translation="want to replace the government."}>\)

want to replace the government.

\(</t>\)

\(<\text{wall/}>\)
Context similarity feature

- Resolve *lexical ambiguities* using the topical context
- Derived from the **Phrase Pair Topic Model** [Hasler et al., 2014]
- Model learns *semantic representations* for all phrase pairs in the phrase table, using a variant of Latent Dirichlet Allocation
Learning semantic representations

- Represent each phrase pair as a pseudo document containing context words
- Standard MT: source context (French)
- Here: target context (English)
Phrase Pair Topic Model

Learning semantic representations

- Represent each phrase pair as a pseudo document containing context words
- Standard MT: source context (French)
- Here: target context (English)
Phrase Pair Topic Model

Context similarity feature

- compute for each test sentence and phrase pairs
Variations of Context Similarity Feature

- **50-topics**
  Cosine similarity according to model with 50 topics

- **mixture:geoAvg**
  Geometric average of cosine similarities of models with 20, 50 and 100 topics

- **mixture:max**
  Choose model (20, 50, 100 topics) that yields lowest entropy of distribution over translations
Training Data

<table>
<thead>
<tr>
<th>Training data</th>
<th>En-Es</th>
<th>Fr-En</th>
<th>Nl-En</th>
</tr>
</thead>
<tbody>
<tr>
<td>Europarl</td>
<td>1.92M</td>
<td>1.96M</td>
<td>1.95M</td>
</tr>
<tr>
<td>News Commentary</td>
<td>192K</td>
<td>181K</td>
<td>n/a</td>
</tr>
<tr>
<td>TED</td>
<td>157K</td>
<td>159K</td>
<td>145K</td>
</tr>
<tr>
<td>News</td>
<td>2.1G</td>
<td>2.1G</td>
<td>2.1G</td>
</tr>
<tr>
<td>Commoncrawl</td>
<td>50M</td>
<td>82M</td>
<td>-</td>
</tr>
</tbody>
</table>

Phrase-based baselines (trained with Moses)

- Standard feature set + 5-gram LM + context similarity feature
Simulating Ambiguous Development Data

- Trial data contained only few ambiguous examples
- Extracted 1076 development instances of 14 ambiguous French words and English translations (mixed corpus of NewsCom, TED, Commoncrawl)

<table>
<thead>
<tr>
<th>Source words</th>
<th>Translations</th>
</tr>
</thead>
<tbody>
<tr>
<td>chaîne</td>
<td>chain, string, channel, station</td>
</tr>
<tr>
<td>matière</td>
<td>matter, material, subject</td>
</tr>
<tr>
<td>flux</td>
<td>stream, flow, feed</td>
</tr>
</tbody>
</table>
Simulating Ambiguous Development Data

Without LM scoring of target context

<table>
<thead>
<tr>
<th>System</th>
<th>French-English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.314</td>
</tr>
<tr>
<td>+ 50-topics</td>
<td>*0.674</td>
</tr>
<tr>
<td>+ mixture:geoAvg</td>
<td>*0.670</td>
</tr>
<tr>
<td>+ mixture:max</td>
<td><strong>0.690</strong></td>
</tr>
</tbody>
</table>

- Score: Average word accuracy
Simulating Ambiguous Development Data

With LM scoring of target context

<table>
<thead>
<tr>
<th>System</th>
<th>French-English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.314</td>
</tr>
<tr>
<td>+ LM context</td>
<td>0.726</td>
</tr>
<tr>
<td>+ 50-topics</td>
<td>*0.886</td>
</tr>
<tr>
<td>+ mixture:geoAvg</td>
<td>*0.883</td>
</tr>
<tr>
<td>+ mixture:max</td>
<td><strong>0.889</strong></td>
</tr>
</tbody>
</table>

• Score: Average word accuracy
Results on Official Test Sets

Without LM scoring of target context

<table>
<thead>
<tr>
<th>System</th>
<th>English-Spanish</th>
<th>French-English</th>
<th>Dutch-English</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>best oof</td>
<td>best oof</td>
<td>best oof</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.674 0.854</td>
<td>0.722 0.884</td>
<td>0.613 0.750</td>
</tr>
<tr>
<td>+50-topics</td>
<td>0.682 0.860</td>
<td>0.719 *0.896</td>
<td>0.616 *0.759</td>
</tr>
<tr>
<td>+geoAvg</td>
<td>0.677 *0.863</td>
<td>0.715 *0.896</td>
<td>0.619 *0.756</td>
</tr>
<tr>
<td>+mixture:max</td>
<td>0.679 0.860</td>
<td>0.712 0.887</td>
<td>0.618 0.753</td>
</tr>
</tbody>
</table>

- Good baseline accuracy
- In most cases, one of the systems including the context similarity feature performs best
## Results on Official Test Sets

With LM scoring of target context (submitted runs)

<table>
<thead>
<tr>
<th>System</th>
<th>English-Spanish</th>
<th>French-English</th>
<th>Dutch-English</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>best oof</td>
<td>best oof</td>
<td>best oof</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.674 0.854</td>
<td>0.722 0.884</td>
<td>0.613 0.750</td>
</tr>
<tr>
<td>+ LM context</td>
<td><strong>0.839</strong> 0.944</td>
<td>0.823 0.934</td>
<td>0.686 0.809</td>
</tr>
<tr>
<td>+ 50-topics</td>
<td>0.827 0.946</td>
<td><strong>0.824</strong> 0.938</td>
<td><strong>0.692</strong> 0.811</td>
</tr>
<tr>
<td>+ geoAvg</td>
<td>0.827 0.944</td>
<td>0.821 *0.939</td>
<td>0.688 0.808</td>
</tr>
<tr>
<td>+ max</td>
<td>0.820 <strong>0.949</strong></td>
<td>0.816 0.937</td>
<td>0.688 0.808</td>
</tr>
</tbody>
</table>

- LM scoring always improves the baseline
- Small gains with context similarity feature (but mostly not significant)
Results on Official Test Sets

Weaker performance of topic feature

- Topic model only helps with lexical ambiguity
- Not all test examples ambiguous without L2 context

Example translations

Input: There are many ways of cooking \(<f>\text{des œufs}</f>\) for breakfast.
Baseline: .. \(<f>\text{eggs}</f>\) ..
50-topics: .. \(<f>\text{eggs}</f>\) ..
Reference: .. \(<f>\text{eggs}</f>\) ..
Results on Official Test Sets

Weaker performance of topic feature

- Topic model only helps with lexical ambiguity
- Not all test examples ambiguous without L2 context

Example translations

Input: This project represents one of the rare advances in strengthening \textit{les liens} between Brazil and the European Union.
Baseline: .. \textit{the links} ..
50-topics: .. \textit{the ties} ..
Reference: .. \textit{the ties}||relations||the bonds ..
Conclusions

L2 writing assistant task

- Phrase-based SMT + LM scoring + context similarity feature
- LM scoring always improves word accuracy
- Topic feature mostly effective for ambiguous source phrases
Thank you!