Better Punctuation Prediction with Hierarchical Phrase-Based Translation

Stephan Peitz, Markus Freitag and Hermann Ney
peitz@cs.rwth-aachen.de

IWSLT 2014, Lake Tahoe, CA
December 4th, 2014

Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik 6
Computer Science Department
RWTH Aachen University, Germany
Introduction

- Spoken language translation (SLT)
  - Automatic speech recognition (ASR)
  - Machine translation (MT)

- In speech punctuation is not made explicit
  - ASR systems provide an output without punctuation marks only
  - MT systems however are trained on data with proper punctuation

- Reintroduce punctuation marks with monolingual translation
  - Translate from unpunctuated text to text with punctuation
  - Based on phrase-based translation

- In this work
  - Use hierarchical system rather than phrase-based
  - Investigation of the optimization criterion
Motivation

► Monolingual translation system
  ▶ More features besides the language model
  ▶ Scaling factors can be tuned

► Phrase-based translation
  ▶ Sequence of words are translated at once
  ▶ Local contextual information is preserved
  ▶ Useful to predict punctuation depending of its surrounding words, e.g. commas
  ▶ Limitation: dependencies beyond the local context

► Hierarchical phrase-based translation
  ▶ Discontinuous phrases with “gaps”
  ▶ Capture long-range dependencies between words and punctuation marks
Translation Model

- Extract from a pseudo-bilingual corpus
- Take monolingual corpus as source and target text
- Create monotone alignment
- Remove punctuation marks from the source text
Modeling Punctuation Prediction as Machine Translation

- **Optimization**
  - Remove punctuation marks from a development set
  - Use the original development set as reference
  - Tune scaling factors with MERT [Och 03]

- **Prediction performance is measured with the $F_1$-Score**
  - Use $F_\alpha$-Score rather than BLEU as a more suitable optimization criterion

$$F_\alpha = (1 + \alpha) \cdot \frac{(\text{precision} \cdot \text{recall})}{\alpha \cdot \text{precision} + \text{recall}}$$

- By varying $\alpha$, more emphasis can be put on recall or precision
Modeling Punctuation Prediction as Machine Translation

► Language model
  ▶ Trained on monolingual corpus with proper punctuation

► Decoding
  ▶ Monotone, no reordering model is necessary
  ▶ Translate from unpunctuated text to text with punctuation

► In this work
  ▶ Perform prediction before the actual translation
  ▶ Final machine translation system has not to be retrained
Hierarchical Phrase-based Translation

- Allow discontinuous phrases with “gaps”
- Obtain phrases from word-aligned bilingual training data
  - Sub-phrases within a phrase are replaced by a generic non-terminal $X$
  - Maximum of two gaps per rule

\[
X \rightarrow \langle \text{über } X_0 \text{ hinausgehen } X_1, \text{ go beyond } X_0 \ X_1 \rangle
\]

- Reordering is modelled implicitly
- Formalized as a synchronous context-free grammar (SCFG)
- Speaking of \textit{rules} rather than phrases

[Chiang 05]
S. Peitz et al. Punctuation Prediction with Hiero
Punctuation Prediction based on Hierarchical Translation

Goal: model dependencies between words and punctuation marks

- e.g. dependency between question word (“was”) and question mark

\[ X \rightarrow \langle \text{was } X_0, \text{was } X_0 \ ? \rangle \]
\[ X \rightarrow \langle \text{machst du } X_0, \text{machst du } X_0 \ ? \rangle \]

Restrictions

- Performing monotone translation
- Reordering is not necessary
- Rules with one non-terminal maximum is sufficient
Additional Extraction Heuristic

\[ X \rightarrow \langle \text{was machst du da}, \text{was machst du da} \rangle \]
Additional Extraction Heuristic

\[ X \rightarrow \langle \text{was machst du da}, \text{was machst du da} \rangle \]
Additional Extraction Heuristic

\[ X \rightarrow \langle \text{was machst du da, was machst du da ?} \rangle \]
\[ X \rightarrow \langle \text{machst Du da, machst Du da} \rangle \]
\[ X \rightarrow \langle \text{was } X_0, \text{ was } X_0 ? \rangle \]
Experimental Evaluation

- Evaluated on the IWSLT 2014 translation tasks
  - German → English and English → French
- Monolingual translation systems were trained on indomain data
- Language model were trained on all available data

- Evaluation of prediction performance
  - Removed punctuation from provided development and test sets (manual transcriptions, no ASR errors)
  - Measurement: Precision, Recall and $F_1$-Score
  - Comparison against HIDDEN-NGRAM [Stolcke 02]
Prediction Results

- From unpunctuated German text to German with punctuation marks
- Tune on $F_\alpha$-Score rather than BLEU
- Replace PBT by HPBT

<table>
<thead>
<tr>
<th>system</th>
<th>tuned on</th>
<th>Prec.</th>
<th>Rec.</th>
<th>$F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBT w/o heuristic</td>
<td>BLEU</td>
<td>86.7</td>
<td>24.4</td>
<td>43.9</td>
</tr>
<tr>
<td>PBT</td>
<td>BLEU</td>
<td>82.7</td>
<td>67.5</td>
<td>74.3</td>
</tr>
<tr>
<td></td>
<td>$F_1$</td>
<td>82.6</td>
<td>67.5</td>
<td>74.3</td>
</tr>
<tr>
<td></td>
<td>$F_2$</td>
<td>78.3</td>
<td>71.4</td>
<td>74.7</td>
</tr>
<tr>
<td></td>
<td>$F_3$</td>
<td>76.6</td>
<td>72.2</td>
<td>74.4</td>
</tr>
<tr>
<td></td>
<td>$F_4$</td>
<td>72.5</td>
<td>73.6</td>
<td>73.0</td>
</tr>
<tr>
<td>HPBT</td>
<td>BLEU</td>
<td>86.4</td>
<td>65.5</td>
<td>74.7</td>
</tr>
<tr>
<td></td>
<td>$F_1$</td>
<td>81.8</td>
<td>71.0</td>
<td>76.0</td>
</tr>
<tr>
<td></td>
<td>$F_2$</td>
<td>77.0</td>
<td>75.4</td>
<td>76.2</td>
</tr>
<tr>
<td></td>
<td>$F_3$</td>
<td>75.9</td>
<td>75.2</td>
<td>75.6</td>
</tr>
<tr>
<td></td>
<td>$F_4$</td>
<td>71.8</td>
<td>73.7</td>
<td>74.2</td>
</tr>
<tr>
<td>HIDDEN-NGRAM</td>
<td>-</td>
<td>82.7</td>
<td>69.5</td>
<td>75.5</td>
</tr>
</tbody>
</table>
Were hierarchical rules used in the decoding process?

<table>
<thead>
<tr>
<th>system</th>
<th>tuned on</th>
<th>lexical rules</th>
<th>hierarchical rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBT</td>
<td>BLEU</td>
<td>2313</td>
<td>-</td>
</tr>
<tr>
<td>PBT</td>
<td>$F_2$</td>
<td>2549</td>
<td>-</td>
</tr>
<tr>
<td>HPBT</td>
<td>$F_2$</td>
<td>2234</td>
<td>442</td>
</tr>
</tbody>
</table>

All applied hierarchical rules introduced punctuation marks
Input "was machst du nur"

▶ PBT "was machst du nur ."

▶ Applied phrases
  ▶ \langle\text{was machst du, was machst du}\rangle
  ▶ \langle\text{nur, nur .}\rangle

▶ HPBT "was machst du nur ?"

▶ Applied rules
  ▶ \text{X} \rightarrow \langle\text{was, was}\rangle
  ▶ \text{X} \rightarrow \langle\text{machst du X}^{\sim 0}, \text{machst du X}^{\sim 0} ?\rangle
  ▶ \text{X} \rightarrow \langle\text{nur, nur}\rangle
Impact on Translation Quality

- **German→English**
- **WER of automatic transcription: 21.6%**

<table>
<thead>
<tr>
<th>system</th>
<th>tuned on</th>
<th>Prec.</th>
<th>Rec.</th>
<th>$F_1$</th>
<th>transcription manual</th>
<th>transcription automatic</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBT</td>
<td>BLEU</td>
<td>82.7</td>
<td>67.5</td>
<td>74.3</td>
<td>27.3</td>
<td>18.7</td>
</tr>
<tr>
<td>PBT</td>
<td>$F_2$</td>
<td>78.3</td>
<td>71.4</td>
<td>74.7</td>
<td>27.5</td>
<td>18.6</td>
</tr>
<tr>
<td>HPBT</td>
<td>$F_2$</td>
<td>77.0</td>
<td>75.4</td>
<td>76.2</td>
<td>27.7</td>
<td>19.1</td>
</tr>
<tr>
<td>HIDDEN-NGRAM</td>
<td>-</td>
<td>82.7</td>
<td>69.5</td>
<td>75.5</td>
<td>27.2</td>
<td>19.0</td>
</tr>
<tr>
<td>correct punctuation</td>
<td></td>
<td>29.4</td>
<td>51.3</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>
Impact on Translation Quality

- **English→French**
- **WER of automatic transcription:** 16.7%

<table>
<thead>
<tr>
<th>system</th>
<th>tuned on</th>
<th>Prec.</th>
<th>Rec.</th>
<th>$F_1$</th>
<th>transcription</th>
<th>manual</th>
<th>automatic</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBT</td>
<td>BLEU</td>
<td>81.2</td>
<td>67.6</td>
<td>73.7</td>
<td>28.4</td>
<td>54.5</td>
<td>22.6</td>
</tr>
<tr>
<td>PBT</td>
<td>$F_2$</td>
<td>72.2</td>
<td>75.0</td>
<td>73.6</td>
<td>28.6</td>
<td>55.2</td>
<td>22.8</td>
</tr>
<tr>
<td>HPBT</td>
<td>$F_2$</td>
<td>74.8</td>
<td>77.1</td>
<td>75.9</td>
<td>28.9</td>
<td>54.7</td>
<td>22.7</td>
</tr>
<tr>
<td>HIDDEN-NGRAM</td>
<td>-</td>
<td>82.0</td>
<td>60.2</td>
<td>69.4</td>
<td>27.0</td>
<td>55.4</td>
<td>21.7</td>
</tr>
<tr>
<td>correct punctuation</td>
<td></td>
<td>31.9</td>
<td>50.1</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Conclusion

► Punctuation prediction based hierarchical translation
  ▶ Capture long-range dependencies between words punctuation marks
  ▶ Improvements in terms of Precision, Recall and $F_1$-Score
  ▶ Small impact on translation quality
► Use $F_\alpha$-Score as optimization criterion
► Future work
  ▶ Investigate features operating on sentence level
  ▶ Enrich grammar with syntactical information
Thank you for your attention

Stephan Peitz

peitz@cs.rwth-aachen.de

http://www-i6.informatik.rwth-aachen.de/~peitz
References


