A Study of Style in Machine Translation: Controlling the Formality of Machine Translation Output

Xing Niu, Marianna Martindale, and Marine Carpuat
CLIP Lab, University of Maryland
• "Anybody hurt?"
• "Is someone wounded?"
A Human Translation Service*
Motivation

• "Anybody hurt?"
• "Is someone wounded?"
  • same literal meaning
  • for different audience or environment

• Goal: controlling formality in machine translation
  • ... by asking what is the expected level of formality

• Prior work looked at other aspects of style
  • politeness in German (Sennrich et al. 2016), gender traits (Rabinovich et al. 2017)
A Study of Style in Machine Translation

• How to control formality in machine translation?
  • Re-ranking-based Formality-Sensitive Machine Translation (FSMT)
Formality-Sensitive MT
Formality-Sensitive MT

- *n*-best list re-ranking with a new feature: \( f(h; \ell) = |\text{Formality}(h) - \ell| \)
  - \( \text{Formality}(h) \): the *sentence-level formality* score \([-1,1]\) of translation hypothesis \( h \).
  - \( \ell \): the desired formality level.
A Study of Style in Machine Translation

• How to control formality in machine translation?
  • Re-ranking-based Formality-Sensitive Machine Translation (FSMT)

• How to score sentential formality?
  • Evaluating existing formality modeling methods
Formality Modeling

• Lexical formality models based on vector space models and formal/informal seed words:
  • SimDiff (Brooke et al. 2010) compares words to formal vs. informal seeds.
  • Support Vector Machine (SVM) finds a hyperplane that separates seeds.
  • Projecting a word in the high-dimensional space to a one-dimensional score using Principal Component Analysis (PCA) or Densifier (Rothe et al. 2016).
Formality Modeling – Intrinsic Evaluation

• sentential formality = weighted average of lexical formality
  • Comparing sentential scores with human annotations (11,263 sentences)
    • Lahiri (2015); Pavlick and Tetreault (2016)
  • Metrics: Spearman correlation.
Formality Modeling – Intrinsic Evaluation

Models are close in performance.

Densifier-LSA is selected as a representative for our FSMT system.

<table>
<thead>
<tr>
<th>Latent Semantic Analysis (LSA)</th>
<th>Word2vec</th>
</tr>
</thead>
<tbody>
<tr>
<td>SimDiff</td>
<td>0.660</td>
</tr>
<tr>
<td>SVM</td>
<td>0.657</td>
</tr>
<tr>
<td>PCA</td>
<td>0.656</td>
</tr>
<tr>
<td>Densifier</td>
<td>0.664</td>
</tr>
</tbody>
</table>

Training data for vector space models: ICWSM 2009 Spinn3r
1.6 billion words from blogs
A Study of Style in Machine Translation

• How to control formality in machine translation?
  • Re-ranking-based Formality-Sensitive Machine Translation (FSMT)

• How to score sentential formality?
  • Evaluating existing formality modeling methods

• How effective is the FSMT system?
  • Automatic + Human evaluations
Formality-Sensitive MT – Evaluation

• Data: MultiUN + OpenSubtitles2016 (French->English)

• 3 FSMT systems, with different desired formality
  • Low ($\ell$=-0.4) | Neutral ($\ell$=0) | High ($\ell$=+0.4)
FSMT – Automatic Evaluation (BLEU)

<table>
<thead>
<tr>
<th>Desired formality</th>
<th>Informal test set</th>
<th>Neutral test set</th>
<th>Formal test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>None (baseline)</td>
<td>39.74</td>
<td>40.17</td>
<td>47.97</td>
</tr>
<tr>
<td>Low</td>
<td><strong>40.27</strong></td>
<td>39.65</td>
<td>47.76</td>
</tr>
<tr>
<td>Neutral</td>
<td>38.70</td>
<td><strong>40.46</strong></td>
<td>47.84</td>
</tr>
<tr>
<td>High</td>
<td>37.58</td>
<td>39.53</td>
<td><strong>47.97</strong></td>
</tr>
</tbody>
</table>

- Shorter sentence → larger impact.
- Formal sentences (MultiUN) are sufficiently different.

- Best: when desired formality level matches reference.
FSMT – Automatic Evaluation (BLEU)

<table>
<thead>
<tr>
<th>Desired formality</th>
<th>Informal test set</th>
<th>Neutral test set</th>
<th>Formal test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>None (baseline)</td>
<td>39.74</td>
<td>40.17</td>
<td>47.97</td>
</tr>
<tr>
<td>Low</td>
<td>40.27</td>
<td>39.65</td>
<td>47.76</td>
</tr>
<tr>
<td>Neutral</td>
<td>38.70</td>
<td>40.46</td>
<td>47.84</td>
</tr>
<tr>
<td>High</td>
<td>37.58</td>
<td>39.53</td>
<td>47.97</td>
</tr>
</tbody>
</table>

• $\Delta$BLEU $\approx 3 \iff$ translation quality difference is large.
• BLEU scores conflate translation errors and stylistic mismatch.
FSMT – Human Assessment

• 42 random affected translation pairs for $\ell=\pm0.4$
• 15 volunteers

• Changes in formality:
  • not impact on adequacy
  • small impact on fluency

• According to formality judgment:
  • FSMT impacts 22/42 examples.
  • FSMT correctly yields 21/22 examples w.r.t. formality.
    • more formal output for $\ell=+0.4$ than $\ell=-0.4$. 
### FSMT – Human Assessment (Examples)

<table>
<thead>
<tr>
<th>$\ell$</th>
<th>Examples</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.4</td>
<td>anybody hurt ?</td>
<td></td>
</tr>
<tr>
<td>+0.4</td>
<td>is someone wounded ?</td>
<td>annotated as more formal</td>
</tr>
<tr>
<td>-0.4</td>
<td>... and then he <strong>ran away</strong> .</td>
<td></td>
</tr>
<tr>
<td>+0.4</td>
<td>... and then he <strong>escaped</strong> .</td>
<td>annotated as more formal</td>
</tr>
<tr>
<td>-0.4</td>
<td>he <strong>shot himself</strong> in the middle of it .</td>
<td></td>
</tr>
<tr>
<td>+0.4</td>
<td>he <strong>committed suicide</strong> in the middle of it .</td>
<td>annotated as more formal</td>
</tr>
<tr>
<td>-0.4</td>
<td>to move <strong>things</strong> forward .</td>
<td></td>
</tr>
<tr>
<td>+0.4</td>
<td><strong>in order</strong> to move <strong>the process</strong> forward.</td>
<td>annotated as more formal</td>
</tr>
<tr>
<td>-0.4</td>
<td>how do you do ?</td>
<td>annotated as more formal</td>
</tr>
<tr>
<td>+0.4</td>
<td>how are you?</td>
<td></td>
</tr>
</tbody>
</table>
A Study of Style in Machine Translation

• How to control formality in machine translation?
  • Re-ranking-based Formality-Sensitive Machine Translation (FSMT)

• How to score sentential formality?
  • Evaluating existing formality modeling methods
  • Empirical comparison → similar performance

• How effective is the FSMT system?
  • Effective in controlling language formality without loss in translation quality
  • Based on automatic evaluation and human assessment
Code: https://github.com/xingniu/computational-stylistic-variations