Unraveling the Contribution of Image Captioning and Neural Machine Translation for Multimodal Machine Translation

Chiraag Lala, Pranava Madhyastha, Josiah Wang, Lucia Specia
University of Sheffield
{clala1, p.madhyastha, j.k.wang, l.specia}@sheffield.ac.uk

Task
We study the Multimodal Machine Translation (MMT) task: given a description in a source language and its corresponding image, translate it into a target language.

Our Contributions
• We isolate two distinct but related components of MMT and analyse their individual contributions:
  – NMT: Machine translation (Neural MT: Nematus [2]) – text-only, bilingual
  – IC: Image caption generation (Multimodal RNN: Show and Tell [4]) – multimodal, monolingual
• We propose a method to combine the output of both components to improve MMT

Experimental Settings

Dataset
The dataset for the WMT16 MMT task [3] is used. Two variants:
• Task 1: 1 English description + 1 professionally translated German description per image
• Task 2: 5 English descriptions + 5 independently crowdsourced German descriptions per image

We concentrate on translating German descriptions to English (DE–EN direction).

Training data
• Parallel: Task 1 corpus. 1 (DE, EN) description pair per image. DE is a direct translation of the EN description.
• Comparable: Task 2 corpus. 5 (DE, EN) description pairs per image. DE is not a direct translation of EN (independently crowdsourced).
• Out Of Domain: Larger corpus.
  – NMT: News, etc. [2]
  – IC: MS COCO [1]
• Cross-comparable (NMT only): Task 2 corpus. Each 5 DE descriptions is paired with each 5 EN descriptions (25 pairs).

Test data
• WMT16 MMT Task 1 test data (1,000 samples)

Analysis
Analysis is performed on NMT and IC models using BLEU, Meteor and four types of Vocabulary Overlap:

\[
\begin{align*}
\mathcal{V}_d(i) &= \frac{|\phi(r_d) \cap \phi(w_i)|}{|\phi(r_d)|} \\
\mathcal{V}_p(i) &= \frac{|\phi(r_p) \cap \phi(w_i)|}{|\phi(r_p)|}
\end{align*}
\]

where \( \phi \) is the set function, \( \oplus \) the concatenation operator, \( \odot \) the intersection operator, \( r \) the cardinality, \( w \) the beam size, \( i \) the test input, \( n \) the n-th best hypothesis for \( r, \odot \) the reference.

Neural MT models

<table>
<thead>
<tr>
<th>Data</th>
<th>Setting</th>
<th>( V_d )</th>
<th>( V_p )</th>
<th>( V_0 )</th>
<th>( \text{BLEU} )</th>
<th>Meteor % len. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>News</td>
<td>Out Of Domain</td>
<td>0.23</td>
<td>0.31</td>
<td>0.40</td>
<td>96.82</td>
<td>95.98</td>
</tr>
<tr>
<td>Task1</td>
<td>Parallel</td>
<td>0.11</td>
<td>0.13</td>
<td>0.19</td>
<td>100.54</td>
<td></td>
</tr>
<tr>
<td>Task2</td>
<td>Comparable</td>
<td>0.10</td>
<td>0.12</td>
<td>0.18</td>
<td>158.07</td>
<td></td>
</tr>
</tbody>
</table>

Neural machine translation performs:
• best when trained on the in-domain parallel Task1 data
• sufficiently well when trained on the out-of-domain parallel News corpus
• very poorly when trained on the remaining comparable data settings

Image captioning models

<table>
<thead>
<tr>
<th>Data</th>
<th>Setting</th>
<th>( V_d )</th>
<th>( V_p )</th>
<th>( \text{BLEU} )</th>
<th>Meteor % len. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSCOCO</td>
<td>Out Of Domain</td>
<td>0.12</td>
<td>0.15</td>
<td>0.19</td>
<td>95.90</td>
</tr>
<tr>
<td>Task1</td>
<td>Parallel</td>
<td>0.11</td>
<td>0.14</td>
<td>0.19</td>
<td>95.90</td>
</tr>
<tr>
<td>Task2</td>
<td>Comparable</td>
<td>0.17</td>
<td>0.26</td>
<td>0.30</td>
<td>12.31</td>
</tr>
</tbody>
</table>

Image captioning performs:
• best when trained on the in-domain Task2 data which has 5 descriptions per image
• poorly when trained on other data settings

Combining NMT and IC for MMT

Main idea: re-rank n-best outputs of NMT models using m-best outputs from IC models.

Scope of Re-ranking: Oracle Experiment

NMT model trained on Task 1 data and IC model trained on Task 2 data.

Re-ranking NMT using IC word probabilities

Re-rank the m-best NMT translations using word probabilities in the m-best IC outputs.

\[ p_{\text{ic}}(w) = \frac{1}{\alpha} \cdot p_{\text{nmt}}(w) + \alpha \cdot p_{\text{ic}}(w) \]

where \( p_{\text{ic}}(w) \) is the new word score, \( p_{\text{nmt}}(w) \) is the word probability from the NMT system, \( p_{\text{ic}}(w) \) is the aggregated word probability from the IC system, by averaging over all occurrences of \( w \) in m-best IC outputs (AVERAGE) \( \alpha \) is tuned on the validation set using grid search.

<table>
<thead>
<tr>
<th>Data</th>
<th>Setting</th>
<th>( V_d )</th>
<th>( V_p )</th>
<th>( \text{BLEU} )</th>
<th>Meteor % len. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td>0.13</td>
<td>0.19</td>
<td>0.25</td>
<td>96.37</td>
</tr>
<tr>
<td>B</td>
<td></td>
<td>0.19</td>
<td>0.26</td>
<td>0.30</td>
<td>95.90</td>
</tr>
<tr>
<td>C</td>
<td></td>
<td>0.22</td>
<td>0.29</td>
<td>0.34</td>
<td>12.31</td>
</tr>
<tr>
<td>D</td>
<td></td>
<td>0.19</td>
<td>0.26</td>
<td>0.30</td>
<td>12.31</td>
</tr>
<tr>
<td>E</td>
<td></td>
<td>0.14</td>
<td>0.18</td>
<td>0.21</td>
<td>158.07</td>
</tr>
</tbody>
</table>

- AVERAGE (39.43 BLEU) outperform text-only NMT baseline (39.13 BLEU)
- Human evaluation: all judges preferred AVERAGE over baseline

IC gave high word probability scores to rocky (0.42) and mountain (0.28) compared to body (0.00) and water (0.00).

Conclusions
• Combining NMT and IC outputs improves MMT performance over NMT system. We confirm that image information definitely has potential to improve MT.
• Future work: Better system combinations/joint models exploiting NMT and IC word probabilities

Acknowledgements

This work was supported by the MultiMT project (H2020 ERC Starting Grant No. 678017).

References