The Task

The University of Sheffield participated in Task 1 of the WMT16 Shared Task on Multimodal Machine Translation (MMT):

- Translate an image description from English to German (and vice versa), given the corresponding image.
- Training and testing on Multi30K dataset. A training instance comprises:
  - An image
  - A textual description in a source language
  - A textual description in a target language, professionally translated from the source language description.

Our submissions

Our submissions use:

- Standard phrase-based SMT system based on the Moses decoder, trained only on the text portion of the provided data.
- Image features to re-rank n-best lists produced by Moses.

Our submissions outperform the strong (text-only) Moses baseline for both EN–DE and DE–EN directions.

Image features

Each image represented as a CNN feature:

- VGG-16 FC8 layer (1,000 dimensions)
- Pre-trained on ImageNet
- Represents the posterior probability estimates for 1,000 WordNet synsets
  - e.g. likelihood that ‘cat’ is depicted in the image
- Each vector sums to 1
- Image classification errors:
  - 7.3% for ILSVRC2014, if correct category is in top 5 predictions, but...
  - Not all 1,000 categories appear in Multi30K.
  - Not all object categories in Multi30K are covered by the 1,000 categories.
  - Errors probably higher for Multi30K.
- Errors likely to propagate to the translation task.

System description

Step 1: Moses decoder (development set)

- Run the Moses decoder on the development set
- Use the text-portion only
- Generate 100-best lists for each translation

Step 2: Extract image features (development set)

- Use as additional features to re-rank the n-best list from Step 1
- Explore two variants:
  - vec: The original FC8 vector
  - prob:
    - likelihood that entities mentioned in the translation are also depicted in image
    - DE–EN direction only
    - Match terms in candidate EN translations to the 1,000 categories
    - Average the ‘visual’ probabilities (from FC8) of all matches per translation
    - vec performs better (official submission)
    - Add vec vector to existing Moses (text-only) features

Step 3: Optimise weights (development set)

- Run the optimiser K-best MIRA (Cherry and Foster, 2012)
- MIRA is known to perform better than MERT for larger feature sets in terms of efficiency and performance
- Compute new weights for each combined feature

Step 4: Moses decoder (test set)

- Repeat Step 1 on the test set, using the original configuration

Step 5: Extract image features (test set)

- Extract FC8 features for each test image (as in Step 2)
- Add to existing Moses features from Step 4

Step 6: Re-rank n-best translations (test set)

- Re-score the 100-best list from Step 4, using re-scoring weights from Step 3 and combined features from Step 5.
- Extract the top scoring translation for each test image as output.

Results

Official Meteor scores on test set (1,000 descriptions)

<table>
<thead>
<tr>
<th>Direction</th>
<th>System</th>
<th>Meteor</th>
<th>Meteor-norm</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN–DE</td>
<td>Moses Baseline</td>
<td>0.525</td>
<td>0.573</td>
</tr>
<tr>
<td></td>
<td>After re-ranking</td>
<td>0.526</td>
<td>0.574</td>
</tr>
<tr>
<td>DE–EN</td>
<td>Moses Baseline</td>
<td>0.363</td>
<td>0.398</td>
</tr>
<tr>
<td></td>
<td>After re-ranking</td>
<td>0.365</td>
<td>0.401</td>
</tr>
</tbody>
</table>

Discussion

- 260 out of the 1,000 test segments differ between the Moses baseline and the re-ranking approach.
- Moses baseline does not produce any translation that is exactly the same as the reference.
- After re-ranking, 37 out of 1,000 translations are exactly the same as the reference translations.

Examples

<table>
<thead>
<tr>
<th>EN</th>
<th>DE (Moses Baseline)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>DE (Moses Baseline)</td>
<td>Ein junger brünnete Frau isst und trinkt etwas.</td>
<td></td>
</tr>
<tr>
<td>Reference</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EN</th>
<th>DE (Moses Baseline)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>A black boy is sitting in the sand.</td>
<td>Ein dunkelhäutiger Junge sitzt im Sand.</td>
<td>Ein schwarzer Junge sitzt im Sand.</td>
</tr>
<tr>
<td>DE (Moses Baseline)</td>
<td>Ein schwarzer Junge sitzt im Sand.</td>
<td></td>
</tr>
<tr>
<td>Reference</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EN</th>
<th>DE (Moses Baseline)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>A man with a black vest holding a model airplane</td>
<td>Ein Mann in einer schwarzen Weste und einem Modellflugzeug</td>
<td>Ein Mann mit einer schwarzen Weste hält einen Modellflugzeug</td>
</tr>
<tr>
<td>DE (Moses Baseline)</td>
<td>Ein Mann mit einer schwarzen Weste hält ein Modellflugzeug</td>
<td></td>
</tr>
<tr>
<td>Reference</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Acknowledgements

This work was supported by the QT21 (H2020 No. 645452), Cracker (H2020 No. 645357), and CHIST-ERA VisualSense (ViSen) (EPSRC EP/K019082/1) projects.