SHEF-Multimodal: Grounding Machine Translation on Images

Kashif Shah, Josiah Wang, Lucia Specia
University of Sheffield, UK

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)

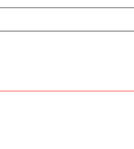
Direction	System	Meteor	Meteor-norm
EN–DE	Moses Baseline	0.525	0.573
	After re-ranking	0.526	0.574
DE–EN	Moses Baseline	0.363	0.398
	After re-ranking	0.365	0.401

Discussion

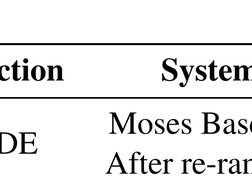
- 260 out of the 1,000 test segments differ between the Moses baseline and the re-ranking approach
- Moses baseline does not produce any translation that is exactly the same as the reference
- After re-ranking, 37 out of 1,000 translations are exactly the same as the reference translations.

Examples

EN	A young brunette woman eating and drinking something.
DE (Moses Baseline)	Eine junge Frau mit braunen Haaren und isst und trinkt etwas.
DE (After re-ranking)	Ein junger brünette Frau isst und trinkt etwas.
Reference	Eine junge brünette Frau isst und trinkt etwas.



EN	A black boy is sitting in the sand.
DE (Moses Baseline)	Ein dunkelhäutiger Junge sitzt im Sand.
DE (After re-ranking)	Ein schwarzer Junge sitzt im Sand.
Reference	Ein schwarzer Junge sitzt im Sand.



EN	A man with a black vest holding a model airplane
DE (Moses Baseline)	Ein Mann in einer schwarzen Weste und einem Modellflugzeug
DE (After re-ranking)	Ein Mann mit einer schwarzen Weste hält einem Modellflugzeug
Reference	Ein Mann mit einer schwarzen Weste hält ein Modellflugzeug

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.