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

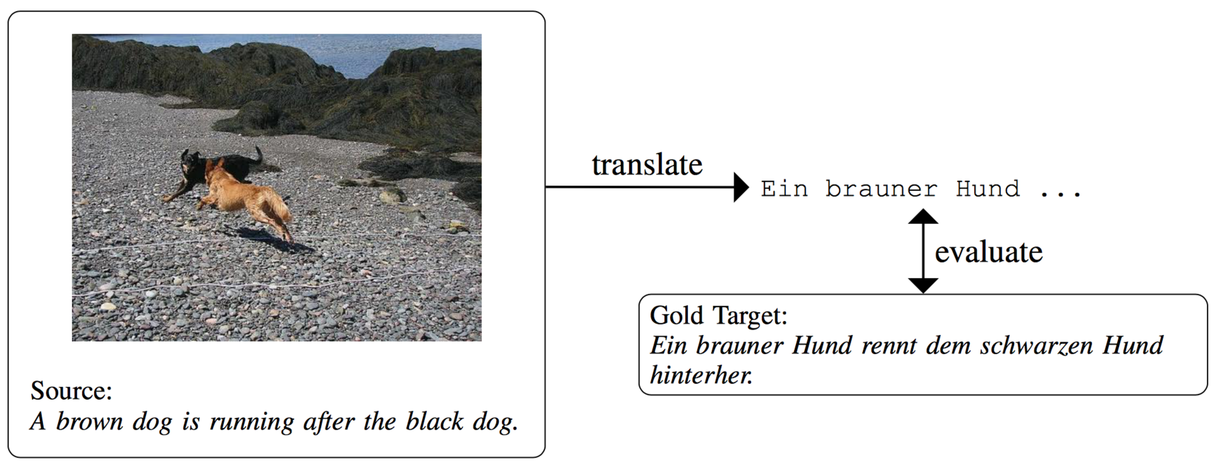
Chiraag Lala, Pranava Madhyastha, Josiah Wang, Lucia Specia

University of Sheffield

{clala1, p.madhyastha, j.k.wang, l.specia}@sheffield.ac.uk

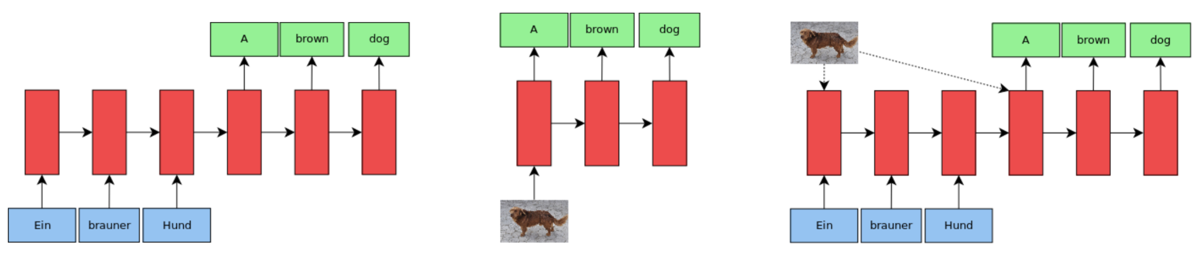
Task

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



(a) Neural MT (b) IC (c) Multimodal NMT

Experimental Settings

Dataset

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**:

$$\mathbb{V}_A(i) = \frac{|\phi(r_i) \cap \phi(o_i^1)|}{|\phi(r_i)|} \qquad \mathbb{V}_B(i) = \frac{|\phi(r_i) \cap \phi(o_i^1)|}{|\phi(o_i^1)|}$$

$$\mathbb{V}_C(i) = \frac{|\phi(r_i) \cap \phi(o_i^1 \oplus o_i^2 \oplus ... \oplus o_i^n)|}{|\phi(r_i)|} \qquad \mathbb{V}_D(i) = \frac{|\phi(r_i) \cap \phi(o_i^1 \oplus o_i^2 \oplus ... \oplus o_i^n)|}{|\phi(o_i^1 \oplus o_i^2 \oplus ... \oplus o_i^n)|}$$

where ϕ is the set function, $\oplus$ the concatenation operator, $\cap$ the intersection operator, $|.|$ the cardinality, n the beam size, i the test input, o_i^n the n-th best hypothesis for i, r_i the reference.

Neural MT models

Data	Setting	$\mathbb{V}_A\uparrow$	$\mathbb{V}_B\uparrow$	$\mathbb{V}_C\uparrow$	$\mathbb{V}_D\uparrow$	BLEU ↑	Meteor ↑	len. (%)
News	*Out Of Domain*	61.24	63.41	69.83	**37.47**	33.89	36.85	96.98
Task1	***Parallel***	**66.11**	**68.27**	**73.02**	36.88	**39.13**	**36.87**	**100.54**
Cross	*Cross-comparable*	26.22	44.23	34.91	19.76	6.92	14.62	63.06
Task2	*Comparable*	21.30	15.44	33.45	6.79	3.08	12.83	158.07

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

Data	Setting	$\mathbb{V}_A\uparrow$	$\mathbb{V}_B\uparrow$	$\mathbb{V}_C\uparrow$	$\mathbb{V}_D\uparrow$	BLEU ↑	Meteor ↑	len. (%)
MSCOCO	*Out Of Domain*	12.08	16.45	20.68	**11.16**	3.11	9.56	78.45
Task1	*Parallel*	11.38	14.19	24.76	6.35	3.91	9.75	**86.37**
Task2	***Comparable***	**17.70**	**26.29**	**30.04**	8.46	**5.79**	**12.31**	75.55

Image captioning performs:

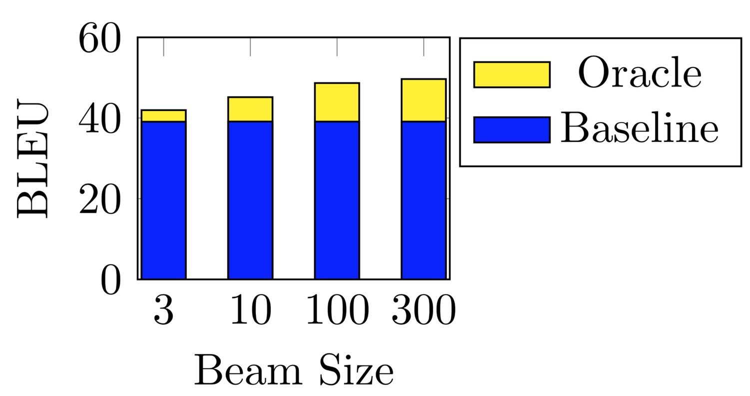
- **best** when trained on the in-domain Task2 data which has 5 descriptions per image
- **poorly** when trained on other data settings

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

$$p_{new}(w) = (1 - \alpha) * p_{nmt}(w) + \alpha * p_{ic}(w)$$

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

Judge	Either	Baseline	AVERAGE
A	17	15	18
B	5	19	26
C	22	9	19
D	19	11	20
E	27	9	14
Total	90 (36%)	63 (25%)	97(39%)

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

Reference a dog treads through a shallow area of water located on a rocky mountainside.

Baseline a dog walks through a body of water, with a body of water in it.

AVERAGE a dog walks through a body of water, looking at a rocky mountain.

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

[1] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[2] Rico Sennrich, Barry Haddow, and Alexandra Birch. Edinburgh neural machine translation systems for wmt 16. In *First Conference on Machine Translation*, pages 371–376, 2016.

[3] Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. A shared task on multimodal machine translation and crosslingual image description. In *First Conference on Machine Translation*, pages 543–553, 2016.

[4] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision & Pattern Recognition*, 2015.