PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 108 JUNE 2017 197–208

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

Abstract

Recent work on multimodal machine translation has attempted to address the problem of producing target language image descriptions based on both the source language description and the corresponding image. However, existing work has not been conclusive on the contribution of visual information. This paper presents an in-depth study of the problem by examining the differences and complementarities of two related but distinct approaches to this task: text-only neural machine translation and image captioning. We analyse the scope for improvement and the effect of different data and settings to build models for these tasks. We also propose ways of combining these two approaches for improved translation quality.

1. Introduction

There has been recent interest among the Machine Translation (MT) community in incorporating different modalities, such as images, to inform and improve machine translation, in contrast to learning from textual data only. For instance, the *Multimodal Machine Translation* (MMT) shared task (Specia et al., 2016) was introduced to investigate if images can potentially help the task of translating an image description (e.g. "*A brown dog is running after the black dog*") to a target language, given the description in a source language and its corresponding image as input (see Figure 1).

In the shared task, the organisers observed that image information is only useful in improving translations when used indirectly (e.g. for re-scoring n-best lists of text-only MT approaches). While this indicates that a text-only MT system is the primary contributor in MMT, it remains inconclusive whether image information can play a

 Corresponding author: clala1@sheffield.ac.uk
Cite as: Chiraag Lala, Pranava Madhyastha, Josiah Wang, Lucia Specia. Unraveling the Contribution of Image Captioning and Neural Machine Translation for Multimodal Machine Translation. The Prague Bulletin of Mathematical Linguistics No. 108, 2017, pp. 197–208. doi: 10.1515/pralin-2017-0020.

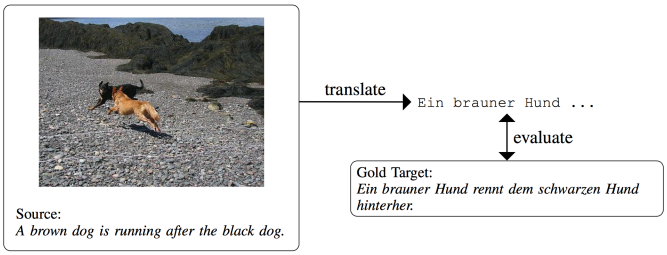


Figure 1: Multimodal Translation Task: source segment (English) and its human translation (German), against which system outputs are evaluated (Specia et al., 2016)

Figure 2: Example of an ambiguous word that could be solved with visual information. The word "hat" in English needs to be disambiguated in order to be translated as "Hut" in German (summer hat), rather than "Mütze" (winter hat)

more significant role. It would be counter-intuitive to simply rule out the contribution of images to the task, particularly when the text is descriptive of the image, which is the case in this dataset. An example (taken from our data) of where visual information can be helpful is shown in Figure 2. We, therefore, posit that visual information is indeed complementary to a text-only MT system for MMT, but the questions are: to what extent and in what way? To our knowledge, no extensive study has been done to understand the role that images play for the MMT task in a systematic manner.

To gain some insight into this matter, in this paper we isolate the text-only MT and the image description generation components of MMT. For the former, we use state-of-the-art Neural MT (NMT) models, which are based on a sequence-to-sequence neural architecture. For image captioning (IC)[1], we use state-of-the-art models based on multimodal recurrent neural networks as described in Vinyals et al. (2015) with default parameter settings. We build models for these two approaches using different datasets (parallel and target language only) and study their complementarities. Additionally, since the decoders of both the approaches perform approximately similar functions, we propose ways of combining the information coming from each model.

Our main contributions, therefore, are (i) an analysis of the individual contributions of a text-only NMT model and a monolingual but multimodal IC model to the MMT task by examining the effect of different data and model settings; and (ii) two

[1]We use the terms "image description" and "image caption" interchangeably.

new approaches for combining the outputs of NMT and IC models. In our experiments, the best-proposed combination approach outperforms the baseline.

2. Background

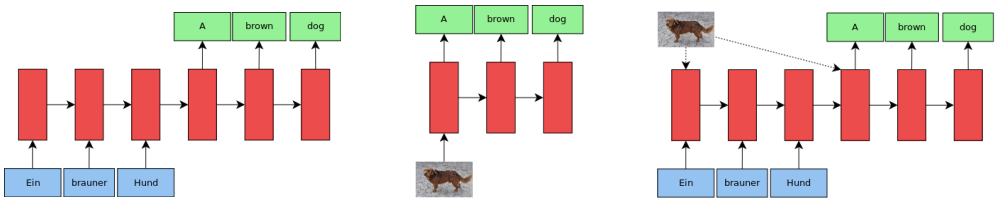
The standard approach in **Neural MT** uses an attention based encoder-decoder model that takes in a source sentence and encodes it using a Recurrent Neural Network (RNN) to produce a sequence of encoded vectors. The approach then decodes it using another RNN in the target language which is conditioned on the sequence of encoded vectors. The model searches through the encoded sequence vectors at each time step and aligns to the corresponding source hidden states adaptively (Bahdanau et al., 2015) (Figure 3a).

Early **Image Captioning** approaches were mainly based on generating a description using explicit visual detector outputs (Yao et al., 2010). We refer readers to Bernardi et al. (2016) for an in-depth discussion on various image captioning approaches. In recent years, multimodal RNN approaches have become dominant, achieving state-of-the-art results on the IC task (Vinyals et al., 2015). Such methods encode an input image as an embedding (e.g. Convolutional Neural Networks (CNN)) and learn an RNN for generating image descriptions conditioned on the image embedding. In this paper, we focus on such state of the art approaches, more specifically the system proposed by Vinyals et al. (2015) which uses a Long Short-Term Memory (LSTM) RNN to model the image descriptions (Figure 3b).

As a first attempt at **Multimodal Machine Translation**, Elliott et al. (2015) added image information at the encoder or the decoder in an NMT setup (Figure 3c) and found marginal improvements from doing so. The systems submitted to the subsequent shared task on Multimodal Machine Translation (Specia et al., 2016) mostly involved a type of NMT, i.e., an encoder-decoder approach, or used a standard phrase-based statistical MT (SMT) system. SMT systems made use of image information mostly during re-ranking, such as Shah et al. (2016). Hitschler et al. (2016) use image information by pivoting it on an external image captioning corpora. Most systems that make use of NMT add the image feature information into either the NMT encoder or decoder (Huang et al., 2016; Hokamp and Calixto, 2016), similar to Elliott et al. (2015) with various enhancements. Marginal improvements according to automatic evaluation metrics were found only for approaches using re-ranking. However, the results of the task do not provide an indication on whether this is inherently because of the task itself (i.e. images cannot help MT) or because of limitations of the methods proposed.

3. Experimental Settings

As Figure 3 shows, IC and NMT models are intrinsically similar from the perspective of decoding, producing the same type of output sequences. The primary differ-



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

Figure 3: Typical architecture of NMT, IC, and MMT systems. In (a), the source sentence is encoded as a sequence of vectors and then decoded using a target language RNN. In (b), the input image is encoded as a vector, and a description is decoded using an RNN. In (c), the source sentence encoding is used as input to the decoder, and the image embedding is used as input to either the source encoder or target decoder

ence is the attention mechanism in NMT. In this section, we analyse the contributions of NMT and IC to a description translation task by studying various aspects of these systems independently and their impact on translation quality.

Dataset: We use the Multi30K dataset (Elliott et al., 2016), an extension of Flickr30K (Young et al., 2014) built for the WMT16 MMT task (Specia et al., 2016). Multi30K contains two variants: (i) one English description and a professionally translated German description per image (used in Task 1: multimodal translation); (ii) five English descriptions and five independently crowdsourced German descriptions per image (used in Task 2: image description generation). See Table 1 for detailed statistics. We use the data in the German–English (DE–EN) direction.

		Train	Val	Test	Tokens	Avg. Length
	Images	29,000	1,014	1,000	–	–
Task1	English	29,000	1,014	1,000	357,172	11.9
	German				333,833	11.1
Task2	English	145,000	5,070	5,000	1,841,159	12.3
	German				1,434,998	9.6

Table 1: Corpus statistics

Data Settings: To analyse the performance of the NMT and IC models with respect to different types of training data, we perform experiments in the following settings:

1. *Parallel*: The corpus for 'Task1' is used. Each image has a corresponding (DE, EN) description pair, where the DE description is a direct (professional) translation of the corresponding EN description.
2. *Comparable*: The corpus for 'Task2' is used. Each image has five independent (DE, EN) description pairs. The DE descriptions are obtained from the image only by crowdsourcing. They are much shorter than the English ones as com-

pared to the Task1 dataset (see Table 1). This is considered a comparable corpus, as the descriptions are not direct translations of each other.

3. *Out of Domain*: Here we train the models on larger datasets of different domains. For NMT, we take (News, etc.) data described in Sennrich et al. (2016), and for IC we take the MSCOCO corpus (Lin et al., 2014). These are large datasets and were not part of the MMT shared task at WMT16.
4. *Cross-comparable* (Only NMT): The corpus of 'Task2' is used to create a new dataset for NMT. Each of the five DE descriptions is randomly paired with each of the five EN descriptions resulting in 25 (DE, EN) description pairs per image. This is similar to the *Comparable* setting except that it is much larger.

All experiments were conducted using the Task1 test set of 1000 samples consisting one reference translation/description for each source sentence/image.

Toolkits: We use state-of-the-art toolkits: Nematus (Sennrich et al., 2016) for NMT and Show and Tell (Vinyals et al., 2015) for IC with default hyperparameters. We experiment with different beam sizes during decoding: 3, 10, 100 and 300. Besides the 1-best output, n-best outputs (where n is the beam size) are also generated from every model to provide a more comprehensive view of what the models can do. For NMT, in order to handle rare words, these are segmented into subwords using the Byte-Pair Encoding Compression Algorithm (Sennrich et al., 2015). We have also tried such a segmentation for IC, but no improvements were observed.

4. Analysis

In the following subsections, the effects of 'Data Setting' and 'Beam Size' on the performance of NMT and IC models are studied using 'Vocabulary Overlap', 'Perplexity', and the MT Metrics 'BLEU' and 'Meteor'. To study the effect of data settings, we fix the beam size to 10 and then train systems on the different training data sets. The data settings that gave the best performing NMT and IC systems are then fixed for the study on the effect of beam size, where we only vary the beam sizes. For a more holistic analysis, both 1-best and n-best outputs are used in our experiments.

4.1. Vocabulary Overlap and Perplexity

The vocabulary overlap between the system-generated outputs and gold standard references helps us to understand the performance of the systems at a very basic level. Given an NMT (or IC) system of beam size n, we denote i to be a test input (a DE sentence for NMT, an image for IC). Let $o_i^1, o_i^2, ..., o_i^n$ be the n-best hypotheses for input i, sorted in descending order by the log probability of o_i^k (i.e., the model score). Let r_i be the reference sequence for input i in the target language (EN). Let ϕ be the set function, $\oplus$ the concatenation operator, $\cap$ the intersection operator, and $|.|$ the cardinality.

We define four types of overlaps as follows:

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

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

$\mathbb{V}_A$ measures the proportion of words in the reference for Task1 captured by the 1-best output, while $\mathbb{V}_B$ measures the proportion of the words in the 1-best output found in the reference. $\mathbb{V}_C$ and $\mathbb{V}_D$ are similar to $\mathbb{V}_A$ and $\mathbb{V}_B$ respectively, except that the 1-best output is replaced by the concatenation of all n-best outputs. $\mathbb{V}_A$ and $\mathbb{V}_C$ correspond to word-overlap recalls, and $\mathbb{V}_B$ and $\mathbb{V}_D$ correspond to word-overlap precisions.

Perplexity scores measure how well the models (NMT and IC) can predict a sample. Given a system that generates a sequence $x_1, \cdots, x_m$ with probabilities $p_1, \cdots, p_m$, perplexity is defined as $\mathbb{P}(x) = 2^{\{-\sum_{i=1}^{m} p_i \log(p_i)\}}$. We use two types of perplexity measures $\mathbb{P}_A, \mathbb{P}_B$ based on whether the 1-best or n-best outputs of our systems are used: a) $\mathbb{P}_A(i) = \mathbb{P}(o_i^1)$ and b) $\mathbb{P}_B(i) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{P}(o_i^k)$

	Data	$\mathbb{V}_A \uparrow$	$\mathbb{V}_B \uparrow$	$\mathbb{V}_C \uparrow$	$\mathbb{V}_D \uparrow$	$\mathbb{P}_A \downarrow$	$\mathbb{P}_B \downarrow$
NMT	News	61.24	63.41	69.83	**37.47**	11.25	12.57
	Task1	**66.11**	**68.27**	**73.02**	36.88	**4.78**	**5.76**
	Cross	26.22	44.23	34.91	19.76	11.16	13.11
	Task2	21.30	15.44	33.45	6.79	49.28	113.57
IC	MSCOCO	12.08	16.45	20.68	**11.16**	**10.22**	**12.38**
	Task1	11.38	14.19	24.76	6.35	19.50	39.59
	Task2	**17.70**	**26.29**	**30.04**	8.46	19.89	35.81

Table 2: Effect of training data studied using Vocabulary Overlaps $\mathbb{V}_A, \mathbb{V}_B, \mathbb{V}_C, \mathbb{V}_D$ (in %), and Perplexity $\mathbb{P}_A, \mathbb{P}_B$. All models are trained with a fixed beam size of 10

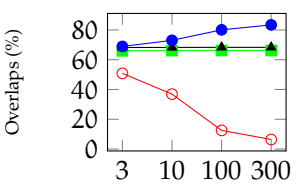
The sentences are pre-processed (removal of symbols and stop words, case-normalisation) to retain only content words. The vocabulary overlap and perplexity scores (averaged over all test inputs) are shown in Table 2 and Figure 4.

4.2. MT Metrics

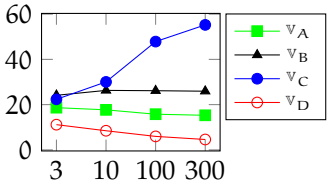
We evaluate the independent NMT and IC systems using BLEU (Papineni et al., 2002) and Meteor (Denkowski and Lavie, 2011). BLEU is computed using the script from Moses suite [2], and Meteor is computed using version 1.5 [3]. In addition, we also measure the ratio between the length of system-generated sequence over the length of reference ('len.'). The scores are tabulated in Tables 3 and 4.

[2] https://github.com/moses-smt

[3] http://www.cs.cmu.edu/~alavie/METEOR



(a) NMT and beam size



(b) IC and beam size

Beam	NMT $\mathbb{P}_A \downarrow$	NMT $\mathbb{P}_B \downarrow$	IC $\mathbb{P}_A \downarrow$	IC $\mathbb{P}_B \downarrow$
3	5.22	5.83	39.19	51.04
10	4.78	5.76	19.89	35.81
100	4.65	5.31	9.49	21.71
300	4.65	5.23	8.52	20.07

(c) Perplexity and beam size

Figure 4: Effect of beam size studied using vocabulary overlap $\mathbb{V}_A, \mathbb{V}_B, \mathbb{V}_C, \mathbb{V}_D$ (in %) and Perplexity $\mathbb{P}_A, \mathbb{P}_B$. Plot (a) shows vocabulary overlap of outputs of NMT system trained on Task1 data. Plot (b) shows vocabulary overlaps of outputs of IC system trained on Task2 data. Table (c) shows perplexity scores.

	Data	BLEU ↑	Meteor ↑	len. (%)
NMT	News	33.89	36.85	96.98
	Task1	**39.13**	**36.87**	**100.54**
	Cross	6.92	14.62	63.06
	Task2	3.08	12.83	158.07
IC	MSCOCO	3.11	9.56	78.45
	Task1	3.91	9.75	**86.37**
	Task2	**5.79**	**12.31**	75.55

Table 3: Effect of training data studied using MT evaluation metrics

	Beam	BLEU ↑	Meteor ↑	len. (%)
NMT	3	39.08	36.81	100.61
	10	39.13	36.87	100.54
	100	39.11	36.89	100.72
	300	39.11	36.89	100.72
IC	3	6.75	12.94	89.63
	10	5.79	12.31	75.55
	100	4.12	10.82	61.13
	300	3.83	10.47	58.73

Table 4: Effect of beam size studied using MT evaluation metrics

4.3. Discussion

Effect of Training Data: We observe that NMT models perform best when trained on the in-domain parallel Task1 data, with overlap $\mathbb{V}_A = 66.11\%$ and $\text{BLEU} = 39.13\%$ as summarised in Tables 2 and 3. We also observe that NMT performs sufficiently well when trained on the Out-of-Domain parallel News corpus with overlap $\mathbb{V}_A = 61.24\%$ and $\text{BLEU} = 33.89\%$. In the remaining comparable data settings (Cross and Task2) it performs very poorly, indicating that NMT system performance generally improves when constrained to parallel corpora and degrades when partially parallel corpora is added. The IC models perform best when trained on the in-domain Task2 data, which has 5 descriptions per image (see Table 1), with overlap $\mathbb{V}_A = 17.70\%$ and $\text{BLEU} = 5.79\%$ (or 20.52% when we use the five references of Task2). It performs poorly in other data settings. When compared to the NMT system, this can be seen as an indication that the ICs are better trained on larger in-domain data having multiple descriptions per image. We also observed that the IC system trained only on MSCOCO produced shorter sentences, resulting in lower perplexity scores.

Effect of Beam Size: By fixing Task1 data for NMT and Task2 data for IC and studying the effect of beam size, we observe that the NMT performance remains largely unchanged as the beam size changes (see Table 4) with $\texttt{BLEU} = 39.1\%$. On the other hand, the IC performance drops as beam size increases. We also observe that IC outputs shorter sentences with larger beam sizes. This is because an end-of-sentence token is more likely to be sampled (and sampled earlier) as beam size increases. Shorter captions are thus ranked higher as they end up having larger model scores (a product of target word probabilities). This may partly explain the performance decrease, although more work is needed to ascertain this. Another interesting observation from this experiment is that the n-best output from both NMT and IC is able to cover more content of the reference as the beam size n increases (See $\mathbb{V}_C, \mathbb{P}_A, \mathbb{P}_B$ in Figure 4). Especially for IC, the overlap $\mathbb{V}_C$ and perplexity measures show large improvements. For instance, $\mathbb{V}_C$ improves from 22.34% (beam 3) to 55.23% (beam 300). This shows that the n-best outputs are able to capture more information content in the reference as the beam size increases. In NMT we see a drastic fall in $\mathbb{V}_D$ from 50.83% (beam 3) to 6.41% (beam 300), which means that as the beam size increases the n-best output of NMT becomes very noisy, with many spurious words. We try to exploit these observations in our system combination strategies in later sections.

5. Combining NMT and IC for MMT

In the previous section, we analysed NMT and IC models independently and observed some important properties. Most notably, for IC the vocabulary overlap $\mathbb{V}_C$ increases drastically for larger beam sizes (see Figure 4) and becomes comparable to NMT models of smaller beam sizes. Recall that $\mathbb{V}_C$ is the overlap of content words in the n-best output (taken collectively) and the reference. This motivates us to explore the possibilities of improving MT by combining the n-best outputs of NMT and IC models of different beam sizes at the word-level.

We approach this task as that of re-ranking the n-best outputs of NMT models using the m-best outputs from IC models. To motivate this, we first explore the scope for improvement with re-ranking through an oracle experiment.

5.1. Scope for Re-ranking: Oracle Experiment

The oracle experiment assumes that we have an 'oracle' that always chooses the best translation out of the n-best outputs generated by the system. We compute an upper bound on the performance of re-ranking approaches using this oracle. For a given MT-metric (we used BLEU) we use the reference translation to obtain the best translation given an n-best list of translation hypotheses.

This experiment was performed on the outputs of NMT systems trained on Task1 for beam sizes 10, 30, 100, and 300. The results are shown in Figure 5. We observe that an ideal re-ranking approach could significantly improve NMT performance. As the

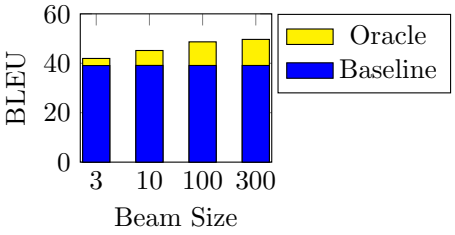


Figure 5: Scope for improvement, as indicated by the yellow bar over the baseline

beam size increases, the scope for obtaining a better translation generally improves. We also performed this experiment for IC systems, but no significant improvements were found. The best translation selected by the oracle is also observed to be usually close to the middle of the system-ranked n-best list. In the following sections, we focus on re-ranking the NMT hypotheses using IC outputs.

5.2. Re-ranking NMT using IC Word Probabilities

We propose to re-rank the n-best NMT translations using image information extracted as word probabilities in the m-best IC outputs. The decoders in both systems produce a word w with a probability $p_{nmt}(w)$ and $p_{ic}(w)$ respectively. We estimate new word scores for each word w by interpolating the information from both systems:

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

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, and α is a hyper-parameter in the range $[0, 1]$ tuned on the validation set using grid search. For a translation hypothesis $(w_1, w_2, ..., w_k)$, its score is computed as a product of these new word-level scores $\prod_{i=1}^{k} p_{new}(w_i)$. We re-rank the n-best NMT hypotheses using the new scores. We propose three ways of aggregating the word probability $\tilde{p}_{ic}(w^t)$ for the t^{th} instance of w in the m-best IC outputs:

1. AVERAGE: $p_{ic}^{avg}(w) = \frac{1}{L} \sum_{t=1}^{L} \tilde{p}_{ic}(w^t)$
2. SUM: $p_{ic}^{sum}(w) = \sum_{t=1}^{L} \tilde{p}_{ic}(w^t)$
3. MAX: $p_{ic}^{max}(w) = \max_{t \in [1,2,...,L]} \tilde{p}_{ic}(w^t)$

where the word w occurs L times in the m-best IC outputs. We set $p_{ic}(w) = 0$ if w does not occur in any of the outputs.

5.3. Re-ranking NMT by similarity with IC Outputs

Here we explore re-ranking NMT hypotheses by their similarity to IC outputs. The motivation is that if we assume the IC outputs accurately describe image content, a

more adequate translation can be selected from the NMT hypotheses if we include the IC outputs in the re-ranking process. We do this by using the BLEU metric as a measure of overlap between an NMT hypothesis and the m-best IC outputs. The NMT hypothesis that has the highest n-gram overlap with the IC outputs should be the most adequate translation. This implies that we are re-ranking the NMT hypotheses based on the information overlap score. For this paper, we use BLEU-4 with smoothing and brevity penalty as the overlap score. We call this approach 'BLEU-rerank'.

5.4. Results and Human Evaluation

For both system combination strategies, the best results are obtained using the NMT system trained on Task1 data and decoded with beam size 10 and the IC system trained on Task2 data with beam size 100 (except for BLEU-rerank where both NMT and IC systems have beam size 3). The highest ranked output after re-ranking is used for evaluation. We report the 1-best output of the same NMT system (before re-ranking) as the baseline. We summarise the results in Table 5. We observe that the method that uses *IC word probabilities* is able to select better sentences. The AVERAGE aggregation works best and gives a small improvement when evaluated with BLEU. Given that the improvement is only observed for BLEU, we resorted to manual evaluation to obtain a better understanding of our re-ranking approaches.

Re-Ranking	α	BLEU↑	Meteor↑
AVERAGE	0.41	**39.43**	36.72
SUM	0.0049	39.34	36.65
MAX	0.26	39.30	36.67
NMT BASELINE	–	39.13	**36.87**
BLEU-rerank	–	36.20	35.30

Table 5: Performance of re-ranking strategies

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%)

Table 6: Human evaluation: NMT vs MMT

Human evaluation: 31% of the 1-best outputs of AVERAGE differ from the baseline after re-ranking. To better understand the differences in these sentences, we asked humans to judge their quality. Five judges (proficient in English) were given 50 samples, each showing the source input image, reference translation, and the translation options from the two systems (without revealing the systems). The judges were asked to decide which option was better in terms of (i) proximity in meaning to the reference and (ii) fluency, giving precedence to the former. They could choose 'Either' when the two translations were equally good or bad. Table 6 summarises the results. All five judges preferred AVERAGE over the text-only baseline.

Figure 6 shows an example output comparing 1-best translation of the text-only baseline and our proposed 'AVERAGE' system combination strategy. The IC system-generated captions give high word probability scores to the words *rocky* and *mountain* compared to the words *body* and *water* [$\mathrm{p}_{\mathrm{ic}}^{\mathrm{avg}}(\mathtt{rocky}) = 0.42$; $\mathrm{p}_{\mathrm{ic}}^{\mathrm{avg}}(\mathtt{mountain}) = 0.28$;

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.

Figure 6: Example output translation for the baseline (text-only NMT) and the best MMT system combination (AVERAGE)

$p_{ic}^{avg}(body) = 0.00; p_{ic}^{avg}(water) = 0.00$]. This is probably because rocky mountain is more prominent in the image. This indicates that there is scope for developing system combination methods and joint models that combine both IC and NMT systems.

6. Conclusions

In this paper, we studied text-only NMT and IC systems independently from each other. The NMT system was found to be better when constrained to an in-domain parallel corpus; its performance degrades when trained on a partly parallel corpus. On the other hand, the IC system was found to be better when trained on a corpus that has multiple descriptions of the same image, enabling the model to capture more information content more reliably from the image. n-best outputs of the IC system are able to capture more information content for higher beam sizes. For NMT, the oracle experiment suggests that there is enormous potential to improve performance for higher beam sizes n if we can re-rank the n-best output wisely. However, we also see the $\mathbb{V}_D$ precision decreases dramatically for NMT with higher beam sizes, suggesting higher chances of spurious re-ranking and, hence, the need to find the right trade-off between more information and spurious information. In our attempt to combine outputs from NMT and IC, we found that system combinations can be helpful if we make use of word probabilities from NMT and IC systems. Our method interpolating these probabilities is able to use image information and outperforms the baseline. This shows evidence that image information has potential to improve MT. Creative and robust system combinations and joint models that exploit NMT and IC word probabilities are promising directions for future work.

Acknowledgements: This work was supported by the MultiMT project (H2020 ERC Starting Grant No. 678017). The authors also thank the anonymous reviewers for their valuable comments.

Bibliography

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR*, 2015.

Bernardi, Raffaella, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures. *Journal of Artificial Intelligence Research*, 55:409–442, 2016.

Denkowski, Michael and Alon Lavie. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *WMT*, 2011.

Elliott, Desmond, Stella Frank, and Eva Hasler. Multi-Language Image Description with Neural Sequence Models. *CoRR*, abs/1510.04709, 2015.

Elliott, D., S. Frank, K. Sima'an, and L. Specia. Multi30K: Multilingual English-German Image Descriptions. In *5th Workshop on Vision and Language*, pages 70–74, 2016.

Hitschler, Julian, Shigehiko Schamoni, and Stefan Riezler. Multimodal Pivots for Image Caption Translation. In *Association for Computational Linguistics*, pages 2399–2409, 2016.

Hokamp, Chris and Iacer Calixto. Multimodal neural machine translation using minimum risk training, 2016. URL `https://www.github.com/chrishokamp/multimodal_nmt`.

Huang, Po-Yao, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. Attention-based Multimodal Neural Machine Translation. In *WMT*, pages 639–645, 2016.

Lin, Tsung-Yi, 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.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In *ACL*, pages 311–318, 2002.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of ACL*, pages 1715–1725, 2015.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. Edinburgh Neural Machine Translation Systems for WMT 16. In *WMT*, pages 371–376, 2016.

Shah, Kashif, Josiah Wang, and Lucia Specia. SHEF-Multimodal: Grounding Machine Translation on Images. In *WMT*, pages 660–665, 2016.

Specia, Lucia, Stella Frank, Khalil Sima'an, and Desmond Elliott. A Shared Task on Multimodal Machine Translation and Crosslingual Image Description. In *WMT*, pages 543–553, 2016.

Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and Tell: A Neural Image Caption Generator. In *IEEE Conference on Computer Vision & Pattern Recognition*, 2015.

Yao, Benjamin Z., Xiong Yang, Liang Lin, Mun Wai Lee, and Song Chun Zhu. I2T: Image Parsing to Text Description. *Proceedings of the IEEE*, 98(8):1485–1508, 2010. ISSN 0018-9219.

Young, Peter, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

Address for correspondence:
Chiraag Lala
`clala1@sheffield.ac.uk`
Department of Computer Science, The University of Sheffield
Regent Court, 211 Portobello, Sheffield, S1 4DP, United Kingdom