Findings of the Second Shared Task on Multimodal Machine Translation and Multilingual Image Description

Desmond Elliott
School of Informatics
University of Edinburgh
d.elliott@ed.ac.uk

Stella Frank
Centre for Language Evolution
University of Edinburgh
stella.frank@ed.ac.uk

Loïc Barrault and **Fethi Bougares**
LIUM
University of Le Mans
first.last@univ-lemans.fr

Lucia Specia
Department of Computer Science
University of Sheffield
l.specia@sheffield.ac.uk

Abstract

We present the results from the second shared task on multimodal machine translation and multilingual image description. Nine teams submitted 19 systems to two tasks. The multimodal translation task, in which the source sentence is supplemented by an image, was extended with a new language (French) and two new test sets. The multilingual image description task was changed such that at test time, only the image is given. Compared to last year, multimodal systems improved, but text-only systems remain competitive.

1 Introduction

The Shared Task on Multimodal Translation and Multilingual Image Description tackles the problem of generating descriptions of images for languages other than English. The vast majority of image description research has focused on English-language description due to the abundance of crowdsourced resources (Bernardi et al., 2016). However, there has been a significant amount of recent work on creating multilingual image description datasets in German (Elliott et al., 2016; Hitschler et al., 2016; Rajendran et al., 2016), Turkish (Unal et al., 2016), Chinese (Li et al., 2016), Japanese (Miyazaki and Shimizu, 2016; Yoshikawa et al., 2017), and Dutch (van Miltenburg et al., 2017). Progress on this problem will be useful for native-language image search, multilingual e-commerce, and audio-described video for visually impaired viewers.

The first empirical results for multimodal translation showed the potential for visual context to improve translation quality (Elliott et al., 2015; Hitschler et al., 2016). This was quickly followed by a wider range of work in the first shared task at WMT 2016 (Specia et al., 2016). The current shared task consists of two subtasks:

- Task 1: **Multimodal translation** takes an image with a source language description that is then translated into a target language. The training data consists of parallel sentences with images.
- Task 2: **Multilingual image description** takes an image and generates a description in the target language without additional source language information at test time. The training data, however, consists of images with independent descriptions in both source and target languages.

The translation task has been extended to include a new language, French. This extension means the Multi30K dataset (Elliott et al., 2016) is now triple aligned, with English descriptions translated into both German and French.

The description generation task has substantially changed since last year. The main difference is that source language descriptions are no longer observed for test images. This mirrors the real-world scenario in which a target-language speaker wants a description of image that does not already have source language descriptions associated with it. The two subtasks are now more distinct because multilingual image description requires the use of the image (no text-only system is possible because the input contains no text).

Another change for this year is the introduction of two new evaluation datasets: an extension of the

Proceedings of the Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 215–233
Copenhagen, Denmark, September 7–11, 2017.

existing Multi30K dataset, and a "teaser" evaluation dataset with images carefully chosen to contain ambiguities in the source language.

This year we encouraged participants to submit systems using unconstrained data for both tasks. Training on additional out-of-domain data is underexplored for these tasks. We believe this setting will be critical for future real-world improvements, given that the current training datasets are small and expensive to construct.

2 Tasks & Datasets

2.1 Tasks

The Multimodal Translation task (Task 1) follows the format of the 2016 Shared Task (Specia et al., 2016). The Multilingual Image Description Task (Task 2) is new this year but it is related to the Crosslingual Image Description task from 2016. The main difference between the Crosslingual Image Description task and the Multilingual Image Description task is the presence of source language descriptions. In last year's Crosslingual Image Description task, the aim was to produce a single target language description, given five source language descriptions and the image. In this year's Multilingual Image Description task, participants received only an unseen image at test time, without source language descriptions.

2.2 Datasets

The Multi30K dataset (Elliott et al., 2016) is the primary dataset for the shared task. It contains 31K images originally described in English (Young et al., 2014) with two types of multilingual data: a collection of professionally translated German sentences, and a collection of independently crowdsourced German descriptions.

This year the Multi30K dataset has been extended with new evaluation data for the Translation and Image Description tasks, and an additional language for the Translation task. In addition, we released a new evaluation dataset featuring ambiguities that we expected would benefit from visual context. Table 1 presents an overview of the new evaluation datasets. Figure 1 shows an example of an image with an aligned English-German-French description.

In addition to releasing the parallel text, we also distributed two types of ResNet-50 visual features (He et al., 2016) for all of the images, namely the 'res4_relu' convolutional features (which preserve the spatial location of a feature in the original image) and averaged pooled features.

En: A group of people are eating noddles.
De: Eine Gruppe von Leuten isst Nudeln.
Fr: Un groupe de gens mangent des nouilles.

Figure 1: Example of an image with a source description in English, together with German and French translations.

Multi30K French Translations

We extended the translation data in Multi30K dataset with crowdsourced French translations. The crowdsourced translations were collected from 12 workers using an internal platform. We estimate the translation work had a monetary value of €9,700. The translators had access to the source segment, the image and an automatic translation created with a standard phrase-based system (Koehn et al., 2007) trained on WMT'15 parallel text. The automatic translations were presented to the crowdworkers to further simplify the crowdsourcing task. We note that this did not end up being a post-editing task, that is, the translators did not simply copy and paste the suggested translations. To demonstrate this, we calculated text-similarity metric scores between the phrase-based system outputs and the human translations on the training corpus, resulting in 0.41 edit distance (measured using the TER metric), meaning that more than 40% of the words between these two versions do not match.

Multi30K 2017 test data

We collected new evaluation data for the Multi30K dataset. We sampled new images from five of the six Flickr groups used to create the original Flickr30K dataset using MMFeat (Kiela, 2016)[1]. We sampled additional images from two thematically related groups (Everything Outdoor and

[1] Strangers!, Wild Child, Dogs in Action, Action Photography, and Outdoor Activities.

	Training set		Development set	
	Images	Sentences	Images	Sentences
Translation	29,000	29,000	1,014	1,014
Description	29,000	145,000	1,014	5,070
	2017 test		COCO	
	Images	Sentences	Images	Sentences
Translation	1,000	1,000	461	461
Description	1,071	5,355	—	

Table 1: Overview of the Multi30K training, development, 2017 test, and Ambiguous COCO datasets.

Group	Task 1	Task 2
Strangers!	150	154
Wild Child	83	83
Dogs in Action	78	92
Action Photography	238	259
Flickr Social Club	241	263
Everything Outdoor	206	214
Outdoor Activities	4	6

Table 2: Distribution of images in the Multi30K 2017 test data by Flickr group.

Flickr Social Club) because Outdoor Activities only returned 10 new CC-licensed images and Flickr-Social no longer exists. Table 2 shows the distribution of images across the groups and tasks. We initially downloaded 2,000 images per Flickr group, which were then manually filtered by three of the authors. The filtering was done to remove (near) duplicate images, clearly watermarked images, and images with dubious content. This process resulted in a total of 2,071 images.

We crowdsourced five English descriptions of each image from Crowdflower[2] using the same process as Elliott et al. (2016). One of the authors selected 1,000 images from the collection to form the dataset for the Multimodal Translation task based on a manual inspection of the English descriptions. Professional German translations were collected for those 1,000 English-described images. The remaining 1,071 images were used for the Multilingual Image Description task. We collected five additional independent German descriptions of those images from Crowdflower.

Ambiguous COCO

As a secondary evaluation dataset for the Multimodal Translation task, we collected and translated a set of image descriptions that potentially contain ambiguous verbs. We based our selection on the VerSe dataset (Gella et al., 2016), which annotates a subset of the COCO (Lin et al., 2014) and TUHOI (Le et al., 2014) images with OntoNotes senses for 90 verbs which are ambiguous, e.g. *play*. Their goals were to test the feasibility of annotating images with the word sense of a given verb (rather than verbs themselves) and to provide a gold-labelled dataset for evaluating automatic visual sense disambiguation methods.

Altogether, the VerSe dataset contains 3,518 images, but we limited ourselves to its COCO section, since for our purposes we also need the image descriptions, which are not available in TUHOI. The COCO portion covers 82 verbs; we further discarded verbs that are unambiguous in the dataset, i.e. although some verbs have multiple senses in OntoNotes, they all occur with one sense in VerSe (e.g. *gather* is used in all instances to describe the 'people gathering' sense), resulting in 57 ambiguous verbs (2,699 images). The actual descriptions of the images were not distributed with the VerSe dataset. However, given that the ambiguous verbs were selected based on the image descriptions, we assumed that in all cases at least one of the original COCO description (out of the five per image) should contain the ambiguous verb. In cases where more than one description contained the verb, we randomly selected one such description to be part of the dataset of descriptions containing ambiguous

[2] http://www.crowdflower.com

En: A man on a motorcycle is passing another vehicle.
De: Ein Mann auf einem Motorrad fährt an einem anderen Fahrzeug vorbei.
Fr: Un homme sur une moto dépasse un autre véhicule.

En: A red train is passing over the water on a bridge
De: Ein roter Zug fährt auf einer Brücke über das Wasser
Fr: Un train rouge traverse l'eau sur un pont.

Figure 2: Two senses of the English verb "to pass" in their visual contexts, with the original English and the translations into German and French. The verb and its translations are underlined.

verbs. This resulted in 2,699 descriptions.

As a consequence of the original goals of the VerSe dataset, each sense of each ambiguous verb was used multiple times in the dataset, which resulted in many descriptions with the same sense, for example, 85 images (and descriptions) were available for the verb *show*, but they referred to a small set of senses of the verb.

The number of images (and therefore descriptions) per ambiguous verb varied from 6 (*stir*) to 100 (*pull*, *serve*). Since our intention was to have a small but varied dataset, we selected a subset of a subset of descriptions per ambiguous verb, aiming at keeping 1-3 instances per sense per verb. This resulted in 461 descriptions for 56 verbs in total, ranging from 3 (e.g. *shake*, *carry*) to 26 (*reach*) (the verb *lay/lie* was excluded as it had only one sense). We note that the descriptions include the use of the verbs in phrasal verbs. Two examples of the English verb "to pass" are shown in Figure 2. In the German translations, the source language verb did not require disambiguation (both German translations use the verb "fährt"), whereas in the French translations, the verb was disambiguated into "dépasse" and "traverse", respectively.

3 Participants

This year we attracted submissions from nine different groups. Table 3 presents an overview of the groups and their submission identifiers.

AFRL-OHIOSTATE (Task 1) The AFRL-OHIOSTATE system submission is an atypical Machine Translation (MT) system in that the image is the catalyst for the MT results, and not the textual content. This system architecture assumes an image caption engine can be trained in a target language to give meaningful output in the form of a set of the most probable n target language candidate captions. A learned mapping function of the encoded source language caption to the corresponding encoded target language captions is then employed. Finally, a distance function is applied to retrieve the "nearest" candidate caption to be the translation of the source caption.

CMU (Task 2) The CMU submission uses a multi-task learning technique, extending the baseline so that it generates both a German caption and an English caption. First, a German caption is generated using the baseline method. After the LSTM for the baseline model finishes producing a German caption, it has some final hidden state. Decoding is simply resumed starting from that final state with an independent decoder, separate vocabulary, and this time without any direct access to the image. The goal is to encourage the model to keep information about the image in the hidden state throughout the decoding process, hopefully improving the model output. Although the model is trained to produce both German and English captions, at evaluation time the English component of the model is ignored and only German captions are

ID	Participating team
AFRL-OHIOSTATE	Air Force Research Laboratory & Ohio State University (Duselis et al., 2017)
CMU	Carnegie Melon University (Jaffe, 2017)
CUNI	Univerzita Karlova v Praze (Helcl and Libovický, 2017)
DCU-ADAPT	Dublin City University (Calixto et al., 2017a)
LIUMCVC	Laboratoire d'Informatique de l'Université du Maine & Universitat Autonoma de Barcelona Computer Vision Center (Caglayan et al., 2017a)
NICT	National Institute of Information and Communications Technology & Nara Institute of Science and Technology (Zhang et al., 2017)
OREGONSTATE	Oregon State University (Ma et al., 2017)
SHEF	University of Sheffield (Madhyastha et al., 2017)
UvA-TiCC	Universiteit van Amsterdam & Tilburg University (Elliott and Kádár, 2017)

Table 3: Participants in the WMT17 multimodal machine translation shared task.

generated.

CUNI (Tasks 1 and 2) For Task 1, the submissions employ the standard neural MT (NMT) scheme enriched with another attentive encoder for the input image. It uses a hierarchical attention combination in the decoder (Libovický and Helcl, 2017). The best system was trained with additional data obtained from selecting similar sentences from parallel corpora and by back-translation of similar sentences found in the SDEWAC corpus (Faaß and Eckart, 2013).

The submission to Task 2 is a combination of two neural models. The first model generates an English caption from the image. The second model is a text-only NMT model that translates the English caption to German.

DCU-ADAPT (Task 1) This submission evaluates ensembles of up to four different multimodal NMT models. All models use global image features obtained with the pre-trained CNN VGG19, and are either incorporated in the encoder or the decoder. These models are described in detail in (Calixto et al., 2017b). They are model IMG_W, in which image features are used as words in the source-language encoder; model IMG_E, where image features are used to initialise the hidden states of the forward and backward encoder RNNs; and model IMG_D, where the image features are used as additional signals to initialise the decoder hidden state. Each image has one corresponding feature vector, obtained from the activations of the FC7 layer of the VGG19 network, and consist of a 4096D real-valued vector that encode information about the entire image.

LIUMCVC (Task 1) LIUMCVC experiment with two approaches: a multimodal attentive NMT with separate attention (Caglayan et al., 2016) over source text and convolutional image features, and an NMT where global visual features (2048-dimensional pool5 features from ResNet-50) are multiplicatively interacted with word embeddings. More specifically, each target word embedding is multiplied with global visual features in an element-wise fashion in order to visually contextualize word representations. With 128-dimensional embeddings and 256-dimensional recurrent layers, the resulting models have around 5M parameters.

NICT (Task 1) These are constrained submissions for both language pairs. First, a hierarchical phrase-based (HPB) translation system s built using Moses (Koehn et al., 2007) with standard features. Then, an attentional encoder-decoder network (Bahdanau et al., 2015) is trained and used as an additional feature to rerank the n-best output of the HPB system. A unimodal NMT model is also trained to integrate visual information. Instead of integrating visual features into the NMT model directly, image retrieval methods are employed to obtain target language descriptions of images that are similar to the image described by the source sentence, and this target description information is integrated into the NMT model. A multimodal

NMT model is also used to rerank the HPB output. All feature weights (including the standard features, the NMT feature and the multimodal NMT feature) were tuned by MERT (Och, 2003). On the development set, the NMT feature improved the HPB system significantly. However, the multimodal NMT feature did not further improve the HPB system that had integrated the NMT feature.

OREGONSTATE (Task 1) The OREGONSTATE system uses a very simple but effective model which feeds the image information to both encoder and decoder. On the encoder side, the image representation was used as an initialization information to generate the source words' representations. This step strengthens the relatedness between image's and source words' representations. Additionally, the decoder uses alignment to source words by a global attention mechanism. In this way, the decoder benefits from both image and source language information and generates more accurate target side sentence.

UvA-TiCC (Task 1) The submitted systems are Imagination models (Elliott and Kádár, 2017), which are trained to perform two tasks in a multitask learning framework: a) produce the target sentence, and b) predict the visual feature vector of the corresponding image. The constrained models are trained over only the 29,000 training examples in the Multi30K dataset with a source-side vocabulary of 10,214 types and a target-side vocabulary of 16,022 types. The unconstrained models are trained over a concatenation of the Multi30K, News Commentary (Tiedemann, 2012) parallel texts, and MS COCO (Chen et al., 2015) dataset with a joint source-target vocabulary of 17,597 word pieces (Schuster and Nakajima, 2012). In both constrained and unconstrained submissions, the models were trained to predict the 2048D GoogleLeNetV3 feature vector (Szegedy et al., 2015) of an image associated with a source language sentence. The output of an ensemble of the three best randomly initialized models - as measured by BLEU on the Multi30K development set - was used for both the constrained and unconstrained submissions.

SHEF (Task 1) The SHEF systems utilize the predicted posterior probability distribution over the image object classes as image features. To do so, they make use of the pre-trained ResNet-152 (He et al., 2016), a deep CNN based image network that is trained over the 1,000 object categories on the Imagenet dataset (Deng et al., 2009) to obtain the posterior distribution. The model follows a standard encoder-decoder NMT approach using *softdot* attention as described in (Luong et al., 2015). It explores image information in three ways: a) to initialize the encoder; b) to initialize the decoder; c) to condition each source word with the image class posteriors. In all these three ways, non-linear affine transformations over the posteriors are used as image features.

Baseline — Task 1 The baseline system for the multimodal translation task is a text-only neural machine translation system built with the Nematus toolkit (Sennrich et al., 2017). Most settings and hyperparameters were kept as default, with a few exceptions: batch size of 40 (instead of 80 due to memory constraints) and ADAM as optimizer. In order to handle rare and OOV words, we used the Byte Pair Encoding Compression Algorithm to segment words (Sennrich et al., 2016b). The merge operations for word segmentation were learned using training data in both source and target languages. These were then applied to all training, validation and test sets in both source and target languages. In post-processing, the original words were restored by concatenating the subwords.

Baseline — Task 2 The baseline for the multilingual image description task is an attention-based image description system trained over only the German image descriptions (Caglayan et al., 2017b). The visual representation are extracted from the so-called *res4f_relu* layer from a ResNet-50 (He et al., 2016) convolutional neural network trained on the ImageNet dataset (Russakovsky et al., 2015). Those feature maps provide spatial information on which the model focuses through the attention mechanism.

4 Text-similarity Metric Results

The submissions were evaluated against either professional or crowd-sourced references. All submissions and references were pre-processed to lowercase, normalise punctuation, and tokenise the sentences using the Moses scripts.[3] The evaluation was performed using `MultEval` (Clark et al., 2011) with the primary metric of Meteor 1.5 (Denkowski and Lavie, 2014). We also report the results using BLEU (Papineni et al., 2002) and

[3] `https://github.com/moses-smt/mosesdecoder/blob/master/scripts/`

TER (Snover et al., 2006) metrics. The winning submissions are indicated by •. These are the top-scoring submissions and those that are not significantly different (based on Meteor scores) according the approximate randomisation test (with p-value ≤ 0.05) provided by `MultEval`. Submissions marked with * are not significantly different from the Baseline according to the same test.

4.1 Task 1: English → German

4.1.1 Multi30K 2017 test data

Table 4 shows the results on the Multi30K 2017 test data with a German target language. It interesting to note that the metrics do not fully agree on the ranking of systems, although the four best (statistically indistinguishable) systems win by all metrics. All-but-one submission outperformed the text-only NMT baseline. This year, the best performing systems include both multimodal (LIUMCVC_MNMT_C and UvA-TiCC_IMAGINATION_U) and text-only (NICT_NMTrerank_C and LIUMCVC_MNMT_C) submissions. (Strictly speaking, the UvA-TiCC_IMAGINATION_U system is incomparable because it is an unconstrained system, but all unconstrained systems perform in the same range as the constrained systems.)

4.1.2 Ambiguous COCO

Table 5 shows the results for the out-of-domain ambiguous COCO dataset with a German target language. Once again the evaluation metrics do not fully agree on the ranking of the submissions.

It is interesting to note that the metric scores are lower for the out-of-domain Ambiguous COCO data compared to the in-domain Multi30K 2017 test data. However, we cannot make definitive claims about the difficulty of the dataset because the Ambiguous COCO dataset contains fewer sentences than the Multi30K 2017 test data (461 compared to 1,000).

The systems are mostly in the same order as on the Multi30K 2017 test data, with the same four systems performing best. However, two systems (DCU-ADAPT_MultiMT_C and OREGON-STATE_1NeuralTranslation_C) are ranked higher on this test set than on the in-domain Flickr dataset, indicating that they are relatively more robust and possibly better at resolving the ambiguities found in the Ambiguous COCO dataset.

4.2 Task 1: English → French

4.2.1 Multi30K Test 2017

Table 6 shows the results for the Multi30K 2017 test data with French as target language. A reduced number of submissions were received for this new language pair, with no unconstrained systems. In contrast to the English→German results, the evaluation metrics are in better agreement about the ranking of the submissions.

Translating from English→French is an easier task than English→German systems, as reflected in the higher metric scores. This also includes the baseline systems where English→French results in 63.1 Meteor compared to 41.9 for English→German.

Eight out of the ten submissions outperformed the English→French baseline system. Two of the best submissions for English→German remain the best for English→French (LIUMCVC_MNMT_C and NICT_NMTrerank_C), the text-only system (LIUMCVC_NMT_C) decreased in performance, and no UvA-TiCC_IMAGINATION_U system was submitted for French.

An interesting observation is the difference of the Meteor scores between text-only NMT system (LIUMCVC_NMT_C) and Moses hierarchical phrase-based system with reranking (NICT_NMTrerank_C). While the two systems are very close for the English→German direction, the hierarchical system is better than the text-only NMT systems in the English→French direction. This pattern holds for both the Multi30K 2017 test data and Ambiguous COCO test data.

4.2.2 Ambiguous COCO

Table 7 shows the results for the out-of-domain Ambiguous COCO dataset with the French target language. Once again, in contrast to the English→German results, the evaluation metrics are in better agreement about the ranking of the submissions. The performance of all the models is once again in mostly agreement with the Multi30K 2017 test data, albeit lower. Both DCU-ADAPT_MultiMT_C and OREGON-STATE_2NeuralTranslation_C again perform relatively better on this dataset.

4.3 Task 2: English → German

The description generation task, in which systems must generate target-language (German) captions for a test image, has substantially changed since

	BLEU ↑	Meteor ↑	TER ↓
•LIUMCVC_MNMT_C	33.4	54.0	48.5
•NICT_NMTrerank_C	31.9	53.9	48.1
•LIUMCVC_NMT_C	33.2	53.8	48.2
•UvA-TiCC_IMAGINATION_U	33.3	53.5	47.5
UvA-TiCC_IMAGINATION_C	30.2	51.2	50.8
CUNI_NeuralMonkeyTextualMT_U	31.1	51.0	50.7
OREGONSTATE_2NeuralTranslation_C	31.0	50.6	50.7
DCU-ADAPT_MultiMT_C	29.8	50.5	52.3
CUNI_NeuralMonkeyMultimodalMT_U	29.5	50.2	52.5
CUNI_NeuralMonkeyTextualMT_C	28.5	49.2	54.3
OREGONSTATE_1NeuralTranslation_C	29.7	48.9	51.6
CUNI_NeuralMonkeyMultimodalMT_C	25.8	47.1	56.3
SHEF_ShefClassInitDec_C	25.0	44.5	53.8
SHEF_ShefClassProj_C	24.2	43.4	55.9
Baseline (text-only NMT)	19.3	41.9	72.2
AFRL-OHIOSTATE-MULTIMODAL_U	6.5	20.2	87.4

Table 4: Official results for the WMT17 Multimodal Machine Translation task on the English-German Multi30K 2017 test data. Systems with grey background indicate use of resources that fall outside the constraints provided for the shared task.

	BLEU ↑	Meteor ↑	TER ↓
•LIUMCVC_NMT_C	28.7	48.9	52.5
•LIUMCVC_MNMT_C	28.5	48.8	53.4
•NICT_1_NMTrerank_C	28.1	48.5	52.9
•UvA-TiCC_IMAGINATION_U	28.0	48.1	52.4
DCU-ADAPT_MultiMT_C	26.4	46.8	54.5
OREGONSTATE_1NeuralTranslation_C	27.4	46.5	52.3
CUNI_NeuralMonkeyTextualMT_U	26.6	46.0	54.8
UvA-TiCC_IMAGINATION_C	26.4	45.8	55.4
OREGONSTATE_2NeuralTranslation_C	26.1	45.7	55.9
CUNI_NeuralMonkeyMultimodalMT_U	25.7	45.6	55.7
CUNI_NeuralMonkeyTextualMT_C	23.2	43.8	59.8
CUNI_NeuralMonkeyMultimodalMT_C	22.4	42.7	60.1
SHEF_ShefClassInitDec_C	21.4	40.7	56.5
SHEF_ShefClassProj_C	21.0	40.0	57.8
Baseline (text-only NMT)	18.7	37.6	66.1

Table 5: Results for the Multimodal Translation task on the English-German Ambiguous COCO dataset. Systems with grey background indicate use of resources that fall outside the constraints provided for the shared task.

	BLEU ↑	Meteor ↑	TER ↓
•LIUMCVC_MNMT_C	55.9	72.1	28.4
•NICT_NMTrerank_C	55.3	72.0	28.4
DCU-ADAPT_MultiMT_C	54.1	70.1	30.0
LIUMCVC_NMT_C	53.3	70.1	31.7
OREGONSTATE_2NeuralTranslation_C	51.9	68.3	32.7
OREGONSTATE_1NeuralTranslation_C	51.0	67.2	33.6
CUNI_NeuralMonkeyMultimodalMT_C	49.9	67.2	34.3
CUNI_NeuralMonkeyTextualMT_C	50.3	67.0	33.6
Baseline (text-only NMT)	44.3	63.1	39.6
*SHEF_ShefClassInitDec_C	45.0	62.8	38.4
SHEF_ShefClassProj_C	43.6	61.5	40.5

Table 6: Results for the Multimodal Translation task on the English-French Multi30K Test 2017 data.

	BLEU ↑	Meteor ↑	TER ↓
•LIUMCVC_MNMT_C	45.9	65.9	34.2
•NICT_NMTrerank_C	45.1	65.6	34.7
•DCU-ADAPT_MultiMT_C	44.5	64.1	35.2
OREGONSTATE_2NeuralTranslation_C	44.1	63.8	36.7
LIUMCVC_NMT_C	43.6	63.4	37.4
CUNI_NeuralMonkeyTexutalMT_C	43.0	62.5	38.2
CUNI_NeuralMonkeyMultimodalMT_C	42.9	62.5	38.2
OREGONSTATE_1NeuralTranslation_C	41.2	61.6	37.8
SHEF_ShefClassInitDec_C	37.2	57.3	42.4
*SHEF_ShefClassProj_C	36.8	57.0	44.5
Baseline (text-only NMT)	35.1	55.8	45.8

Table 7: Results for the Multimodal Translation task on the English-French Ambiguous COCO dataset.

	BLEU ↑	Meteor ↑	TER ↓
Baseline (target monolingual)	9.1	23.4	91.4
CUNI_NeuralMonkeyCaptionAndMT_C	4.2	22.1	133.6
CUNI_NeuralMonkeyCaptionAndMT_U	6.5	20.6	91.7
CMU_NeuralEncoderDecoder_C	9.1	19.8	63.3
CUNI_NeuralMonkeyBilingual_C	2.3	17.6	112.6

Table 8: Results for the Multilingual Image Description task on the English-German Multi30K 2017 test data.

last year. The main difference is that source language descriptions are no longer observed for images at test time. The training data remains the same and contains images with both source and target language descriptions. The aim is thus to leverage multilingual training data to improve a monolingual task.

Table 8 shows the results for the Multilingual image description task. This task attracted fewer submissions than last year, which may be because it was no longer possible to re-use a model designed for Multimodal Translation. The evaluation metrics do not agree on the ranking of the submissions, with major differences in the ranking using either BLEU or TER instead of Meteor.

The main result is that none of the submissions outperform the monolingual German baseline according to Meteor. All of the submissions are statistically significantly different compared to the baseline. However, the CMU_NeuralEncoderDecoder_C submission marginally outperformed the baseline according to TER and equalled its BLEU score.

5 Human Judgement Results

This year, we conducted a human evaluation in addition to the text-similarity metrics to assess the translation quality of the submissions. This evaluation was undertaken for the Task 1 German and French outputs for the Multi30K 2017 test data.

This section describes how we collected the human assessments and computed the results. We would like to gratefully thank all assessors.

5.1 Methodology

The system outputs were manually evaluated by bilingual Direct Assessment (DA) (Graham et al., 2015) using the Appraise platform (Federmann, 2012). The annotators (mostly researchers) were asked to evaluate the semantic relatedness between the source sentence in English and the target sentence in German or French. The image was shown along with the source sentence and the candidate translation and evaluators were told to rely on the image when necessary to obtain a better understanding of the source sentence (e.g. in cases where the text was ambiguous). Note that the reference sentence is not displayed during the evaluation, in order to avoid influencing the assessor. Figure 3 shows an example of the direct assessment interface used in the evaluation. The score of each translation candidate ranges from 0 (meaning that the meaning of the source is not preserved in the target language sentence) to 100 (meaning the meaning of the source is "perfectly" preserved). The human assessment scores are standardized according to each individual assessor's overall mean and standard deviation score. The overall score of a given system (z) corresponds to the mean standardized score of its translations.

5.2 Results

The French outputs were evaluated by seven assessors, who conducted a total of 2,521 DAs, resulting in a minimum of 319 and a maximum of 368 direct assessments per system submission, respectively. The German outputs were evaluated by 25 assessors, who conducted a total of 3,485 DAs, resulting in a minimum of 291 and a maximum of 357 direct assessments per system submission, respectively. This is somewhat less than the recommended number of 500, so the results should be considered preliminary.

Tables 9 and 10 show the results of the human evaluation for the English to German and the English to French Multimodal Translation task (Multi30K 2017 test data). The systems are ordered by standardized mean DA scores and clustered ac-

Figure 3: Example of the human direct assessment evaluation interface.

cording to the Wilcoxon signed-rank test at p-level $p \leq 0.05$. Systems within a cluster are considered tied. The Wilcoxon signed-rank scores can be found in Tables 11 and 12 in Appendix A.

When comparing automatic and human evaluations, we can observe that they globally agree with each other, as shown in Figures 4 and 5, with German showing better agreement than French. We point out two interesting disagreements: First, in the English→French language pair, CUNI_NeuralMonkeyMultimodalMT_C and DCU-ADAPT_MultiMT_C are significantly better than LIUMCVC_MNMT_C, despite the fact that the latter system achieves much higher metric scores. Secondly, across both languages, the text-only LIUMCVC_NMT_C system performs well on metrics but does relatively poorly on human judgements, especially as compared to the multimodal version of the same system.

6 Discussion

Visual Features: do they help? Three teams provided text-only counterparts to their multimodal systems for Task 1 (CUNI, LIUMCVC, and OREGONSTATE), which enables us to evaluate the contribution of visual features. For many systems, visual features did not seem to help reliably, at least as measured by metric evaluations: in German, the CUNI and OREGONSTATE text-only systems outperformed the counterparts, while in French, there were small improvements for the CUNI multimodal system. However, the LIUMCVC multimodal system outperformed their text-only system across both languages.

The human evaluation results are perhaps more promising: nearly all the highest ranked systems (with the exception of NICT) are multimodal. An intruiging result was the text-only LIUMCVC_NMT_C, which ranked highly on metrics but poorly in the human evaluation. The LIUMCVC systems were indistinguishable from each other in terms of Meteor scores but the standardized mean direct assessment score showed a significant difference in performance (see Tables 11 and 12): further analysis of the reasons for humans disliking the text-only translations will be necessary.

The multimodal Task 1 submissions can be broadly categorised into three groups based on how they use the images: approaches useing double-attention mechanisms, initialising the hidden state of the encoder and/or decoder networks with the global image feature vector, and alternative uses of image features. The double-attention models calculate context vectors over the source language hidden states and location-preserving feature vectors over the image; these vectors are used as inputs to the translation decoder (CUNI_NeuralMonkeyMultimodalMT). Encoder and/or decoder initialisation involves initialising the recurrent neural network with an affine transformation of a global image feature vector (DCU-ADAPT_MultiMT, OREGONSTATE_1NeuralTranslation) or initialising the encoder and decoder with the 1000 dimension softmax probability vector over the object classes in ImageNet object recognition challenge

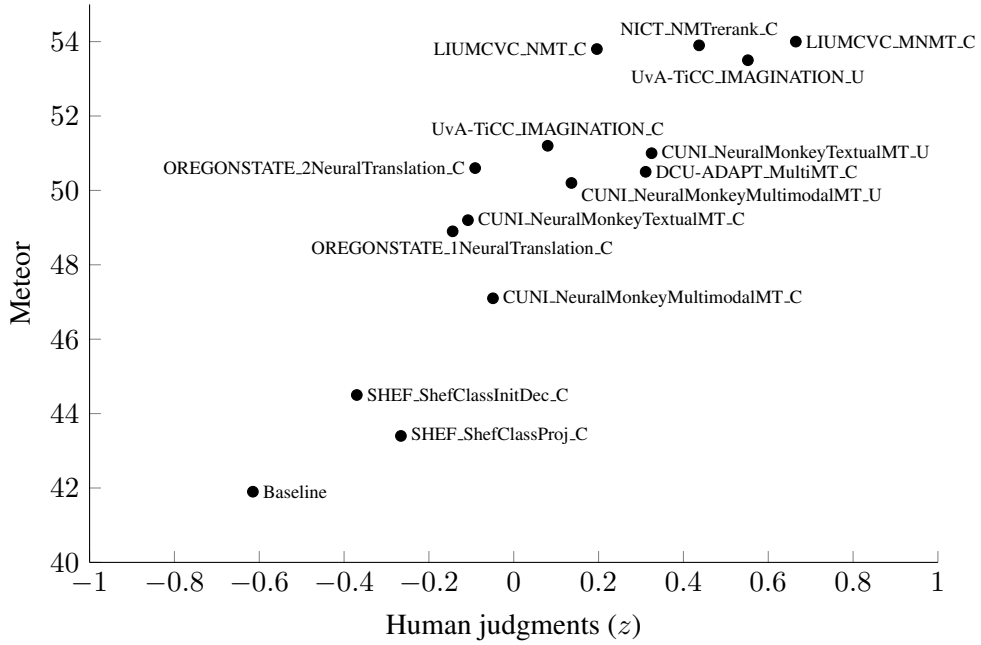

Figure 4: System performance on the English→German Multi30K 2017 test data as measured by human evaluation against Meteor scores. The AFRL-OHIOSTATE-MULTIMODAL_U system has been ommitted for readability.

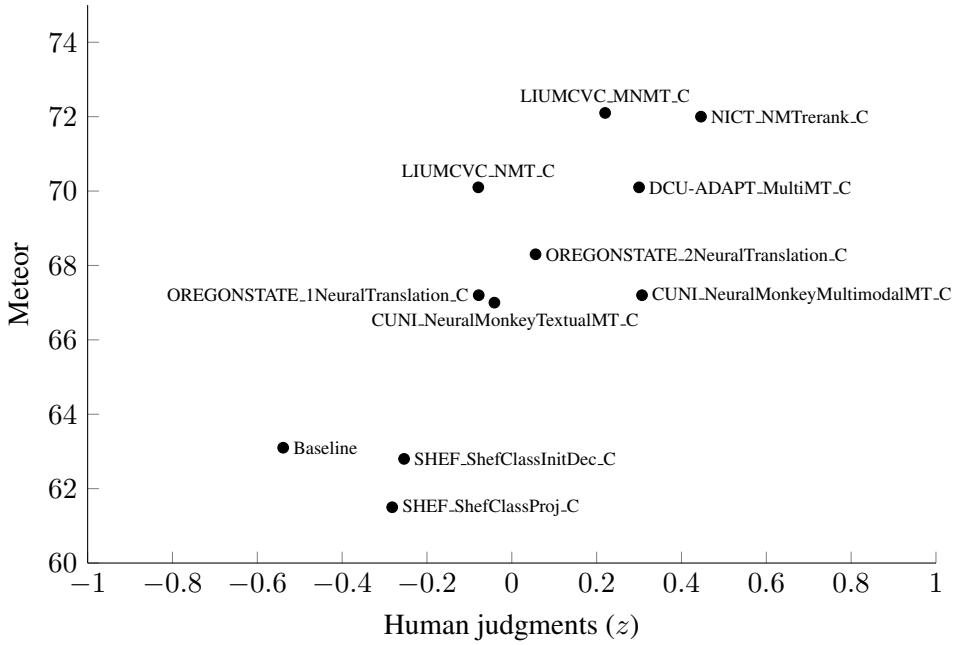

Figure 5: System performance on the English→French Multi30K 2017 test data as measured by human evaluation against Meteor scores.

			English→German
#	Raw	z	System
1	77.8	0.665	LIUMCVC_MNMT_C
2	74.1	0.552	UvA-TiCC_IMAGINATION_U
3	70.3	0.437	NICT_NMTrerank_C
	68.1	0.325	CUNI_NeuralMonkeyTextualMT_U
	68.1	0.311	DCU-ADAPT_MultiMT_C
	65.1	0.196	LIUMCVC_NMT_C
	60.6	0.136	CUNI_NeuralMonkeyMultimodalMT_U
	59.7	0.08	UvA-TiCC_IMAGINATION_C
	55.9	-0.049	CUNI_NeuralMonkeyMultimodalMT_C
	54.4	-0.091	OREGONSTATE_2NeuralTranslation_C
	54.2	-0.108	CUNI_NeuralMonkeyTextualMT_C
	53.3	-0.144	OREGONSTATE_1NeuralTranslation_C
	49.4	-0.266	SHEF_ShefClassProj_C
	46.6	-0.37	SHEF_ShefClassInitDec_C
15	39.0	-0.615	Baseline (text-only NMT)
	36.6	-0.674	AFRL-OHIOSTATE_MULTIMODAL_U

Table 9: Results of the human evaluation of the WMT17 English-German Multimodal Translation task (Multi30K 2017 test data). Systems are ordered by standardized mean DA scores (z) and clustered according to Wilcoxon signed-rank test at p-level $p \leq 0.05$. Systems within a cluster are considered tied, although systems within a cluster may be statistically significantly different from each other (see Table 11). Systems using unconstrained data are identified with a gray background.

(SHEF_ShefClassInitDec). The alternative uses of the image features include element-wise multiplication of the target language embeddings with an affine transformation of a global image feature vector (LIUMCVC_MNMT), summing the source language word embeddings with affine-transformed 1000 dimension softmax probability vector (SHEF_ShefClassProj), using the visual features in a retrieval framework (AFRL-OHIOSTATE_MULTIMODAL), and learning visually-grounded encoder representations by learning to predict the global image feature vector from the source language hidden states (UvA-TiCC_IMAGINATION).

Overall, the metric and human judgement results in Sections 4 and 5 indicate that there is still a wide scope for exploration of the best way to integrate visual and textual information. In particular, the alternative approaches proposed in the LIUMCVC_MNMT and UvA-TiCC_IMAGINATION submissions led to strong performance in both the metric and human judgement results, surpassing the more common approaches using initialisation and double attention.

Finally, the text-only NICT system ranks highly across both languages. This system uses hierarchical phrase-based MT with a reranking step based on a neural text-only system, since their multimodal system never outperformed the text-only variant in development (Zhang et al., 2017). This is in line with last year's results and the strong Moses baseline (Specia et al., 2016), and suggests a continuing role for phrase-based MT for small homogeneous datasets.

Unconstrained systems The Multi30k dataset is relatively small, so unconstrained systems use more data to complement the image description translations. Three groups submitted systems using external resources: UvA-TiCC, CUNI, and AFRL-OHIOSTATE. The unconstrained UvA-TiCC and CUNI submissions always outperformed their respective constrained variants by 2–3 Meteor points and achieved higher standardized mean DA scores. These results suggest that external parallel text corpora (UvA-TiCC and CUNI) and external monolingual image description datasets (UvA-TiCC) can usefully improve the quality of multimodal translation models.

However, tuning to the target domain remains important, even for relatively simple image captions.

			English→French
#	Raw	z	System
1	79.4	0.446	NICT_NMTrerank_C
	74.2	0.307	CUNI_NeuralMonkeyMultimodalMT_C
	74.1	0.3	DCU-ADAPT_MultiMT_C
4	71.2	0.22	LIUMCVC_MNMT_C
	65.4	0.056	OREGONSTATE_2NeuralTranslation_C
	61.9	-0.041	CUNI_NeuralMonkeyTextualMT_C
	60.8	-0.078	OREGONSTATE_1NeuralTranslation_C
	60.5	-0.079	LIUMCVC_NMT_C
9	54.7	-0.254	SHEF_ShefClassInitDec_C
	54.0	-0.282	SHEF_ShefClassProj_C
11	44.1	-0.539	Baseline (text-only NMT)

Table 10: Results of the human evaluation of the WMT17 English-French Multimodal Translation task (Multi30K 2017 test data). Systems are ordered by standardized mean DA score (z) and clustered according to Wilcoxon signed-rank test at p-level $p \leq 0.05$. Systems within a cluster are considered tied, although systems within a cluster may be statistically significantly different from each other (see Table 12).

We ran the best-performing English→German WMT'16 news translation system (Sennrich et al., 2016a) on the English→German Multi30K 2017 test data to gauge the performance of a state-of-the-art text-only translation system trained on only out-of-domain resources[4]. It ranked 10th in terms of Meteor (49.9) and 11th in terms of BLEU (29.0), placing it firmly in the middle of the pack, and below nearly all the text-only submissions trained on the in-domain Multi30K dataset.

The effect of OOV words The Multi30k translation training and test data are very similar, with a low OOV rate in the Flickr test set (1.7%). In the 2017 test set, 16% of English test sentences include a OOV word. Human evaluation gave the impression that these often led to errors propagated throughout the whole sentence. Unconstrained systems may perform better by having larger vocabularies, as well as more robust statistics. When we evaluate the English→German systems over only the 161 OOV-containing test sentences, the highest ranked submission by all metrics is the unconstrained UvA-TiCC_IMAGINATION submission, with +2.5 Metor and +2.2 BLEU over the second best system (LIUMCVC_NMT; 45.6 vs 43.1 Meteor and 24.0 vs 21.8 BLEU).

The difference over non-OOV-containing sentences is not nearly as stark, with constrained systems all performing best (both LIUMCVC systems, MNMT and NMT, with 56.6 and 56.3 Meteor, respectively) but unconstrained systems following close behind (UvA-TiCC with 55.4 Meteor, CUNI with 53.4 Meteor).

Ambiguous COCO dataset We introduced a new evaluation dataset this year with the aim of testing systems' ability to use visual features to identify word senses.

However, it is unclear whether visual features improve performance on this test set. The text-only NICT_NMTrerank system performs competitively, ranking in the top three submissions for both languages. We find mixed results for submissions with text-only and multimodal counterparts (CUNI, LIUMCVC, OREGONSTATE): LIUMCVC's multimodal system improves over the text-only system for French but not German, while the visual features help for German but not French in the CUNI and OREGONSTATE systems.

We plan to perform a further analysis on the extent of translation ambiguity in this dataset. We will also continue to work on other methods for constructing datasets in which textual ambiguity can be disambiguated by visual information.

Multilingual Image Description It proved difficult for Task 2 systems to use the English data to improve over the monolingual German baseline.

[4] http://data.statmt.org/rsennrich/wmt16_systems/en-de/

In future iterations of the task, we will consider a lopsided data setting, in which there is much more English data than target language data. This setting is more realistic and will push the use of multilingual data. We also hope to conduct human evaluation to better assess performance because automatic metrics are problematic for this task (Elliott and Keller, 2014; Kilickaya et al., 2017).

7 Conclusions

We presented the results of the second shared task on multimodal translation and multilingual image description. The shared task attracted submissions from nine groups, who submitted a total of 19 systems across the tasks. The Multimodal Translation task attracted the majority of the submissions. Human judgements for the translation task were collected for the first time this year and ranked systems broadly in line with the automatic metrics.

The main findings of the shared task are:

(i) There is still scope for novel approaches to integrating visual and linguistic features in multilingual multimodal models, as demonstrated by the winning systems.

(ii) External resources have an important role to play in improving the performance of multimodal translation models beyond what can be learned from limited training data.

(iii) The differences between text-only and multimodal systems are being obfuscated by the well-known shortcomings of text-similarity metrics. Multimodal systems often seem to be prefered by humans but not rewarded by metrics. Future research on this topic, encompassing both multimodal translation and multilingual image description, should be evaluated using human judgements.

In future editions of the task, we will encourage participants to submit the output of single decoder systems to better understand the empirical differences between approaches. We are also considering a Multilingual Multimodal Translation challenge, where the systems can observe two language inputs alongside the image to encourage the development of multi-source multimodal models.

Acknowledgements

This work was supported by the CHIST-ERA M2CR project (French National Research Agency No. ANR-15-CHR2-0006-01 – Loïc Barrault and Fethi Bougares), and by the MultiMT project (EU H2020 ERC Starting Grant No. 678017 – Lucia Specia). Desmond Elliott acknowledges the support of NWO Vici Grant No. 277-89-002 awarded to K. Sima'an, and an Amazon Academic Research Award. We thank Josiah Wang for his help in selecting the Ambiguous COCO dataset.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations*.

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *J. Artif. Intell. Res.*, 55:409–442.

Ozan Caglayan, Loïc Barrault, and Fethi Bougares. 2016. Multimodal attention for neural machine translation. *CoRR*, abs/1609.03976.

Ozan Caglayan, Walid Aransa, Adrien Bardet, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, Marc Masana, Luis Herranz, and Joost van de Weijer. 2017a. LIUM-CVC Submissions for WMT17 Multimodal Translation Task. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 432–439.

Ozan Caglayan, Mercedes García-Martínez, Adrien Bardet, Walid. Aransa, Fethi. Bougares, and Loïc Barrault. 2017b. NMTPY: A Flexible Toolkit for Advanced Neural Machine Translation Systems. *CoRR*, 1706.00457.

Iacer Calixto, Koel Dutta Chowdhury, and Qun Liu. 2017a. DCU System Report on the WMT 2017 Multi-modal Machine Translation Task. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 440–444.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017b. Incorporating Global Visual Features into Attention-Based Neural Machine Translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325.

Jonathan H Clark, Chris Dyer, Alon Lavie, and Noah A Smith. 2011. Better hypothesis testing for statistical

machine translation: Controlling for optimizer instability. In *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 176–181.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *EACL 2014 Workshop on Statistical Machine Translation*.

John Duselis, Michael Hutt, Jeremy Gwinnup, James Davis, and Joshua Sandvick. 2017. The AFRL-OSU WMT17 Multimodal Translation System: An Image Processing Approach. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 445–449.

Desmond Elliott and Ákos Kádár. 2017. Imagination improves Multimodal Translation. *CoRR*, abs/1705.04350.

Desmond Elliott and Frank Keller. 2014. Comparing automatic evaluation measures for image description. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: Short Papers*, pages 452–457.

Desmond Elliott, Stella Frank, and Eva Hasler. 2015. Multi-language image description with neural sequence models. *CoRR*, abs/1510.04709.

Desmond Elliott, Stella Frank, Khalil Simaan, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. In *5th Workshop on Vision and Language*, pages 70–74.

Gertrud Faaß and Kerstin Eckart. 2013. SdeWaC–a corpus of parsable sentences from the web. In *Language processing and knowledge in the Web*, pages 61–68. Springer.

Christian Federmann. 2012. Appraise: An open-source toolkit for manual evaluation of machine translation output. *The Prague Bulletin of Mathematical Linguistics*, 98:25–35, September.

Spandana Gella, Mirella Lapata, and Frank Keller. 2016. Unsupervised visual sense disambiguation for verbs using multimodal embeddings. In *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 182–192, San Diego, California.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2015. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Jindřich Helcl and Jindřich Libovický. 2017. CUNI System for the WMT17 Multimodal Translation Task. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 450–457.

Julian Hitschler, Shigehiko Schamoni, and Stefan Riezler. 2016. Multimodal Pivots for Image Caption Translation. In *54th Annual Meeting of the Association for Computational Linguistics*, pages 2399–2409.

Alan Jaffe. 2017. Generating Image Descriptions using Multilingual Data. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 458–464.

Douwe Kiela. 2016. MMFeat: A toolkit for extracting multi-modal features. In *Proceedings of ACL-2016 System Demonstrations*, pages 55–60.

Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. 2017. Re-evaluating automatic metrics for image captioning. In *Proceedings of EACL 2017*, pages 199–209.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *45th Annual meeting of Association for Computational Linguistics*, pages 177–180.

D.T. Le, R. Bernardi, and J.R.R. Uijlings. 2014. TUHOI: Trento Universal Human Object Interaction Dataset. In *Vision and Language Workshop at the 26th International Conference on Computational Linguistics*.

Xirong Li, Weiyu Lan, Jianfeng Dong, and Hailong Liu. 2016. Adding chinese captions to images. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pages 271–275.

Jindřich Libovický and Jindřich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. *CoRR*, abs/1704.06567.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Mingbo Ma, Dapeng Li, Kai Zhao, and Liang Huang. 2017. OSU Multimodal Machine Translation System Report. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 465–469.

Pranava Swaroop Madhyastha, Josiah Wang, and Lucia Specia. 2017. Sheffield MultiMT: Using Object Posterior Predictions for Multimodal Machine Translation. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 470–476.

Takashi Miyazaki and Nobuyuki Shimizu. 2016. Cross-lingual image caption generation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1780–1790.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *41st Annual Meeting on Association for Computational Linguistics*, pages 160–167, Sapporo, Japan.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *40th Annual Meeting on Association for Computational Linguistics*, pages 311–318.

Janarthanan Rajendran, Mitesh M Khapra, Sarath Chandar, and Balaraman Ravindran. 2016. Bridge correlational neural networks for multilingual multimodal representation learning. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 171–181.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725, Berlin, Germany.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, et al. 2017. Nematus: a toolkit for neural machine translation. In *Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–68.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*.

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

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Eight International Conference on Language Resources and Evaluation (LREC'12)*.

Mesut Erhan Unal, Begum Citamak, Semih Yagcioglu, Aykut Erdem, Erkut Erdem, Nazli Ikizler Cinbis, and Ruket Cakici. 2016. Tasviret: Görüntülerden otomatik türkçe açıklama oluşturma İçin bir denektaçı veri kümesi (TasvirEt: A benchmark dataset for automatic Turkish description generation from images). In *IEEE Sinyal İşleme ve İletişim Uygulamaları Kurultayı (SIU 2016)*.

Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2017. Cross-linguistic differences and similarities in image descriptions. In *Proceedings of the 10th International Conference on Natural Language Generation*.

Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. 2017. Stair captions: Constructing a large-scale japanese image caption dataset. *CoRR*, abs/1705.00823.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. 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.

Jingyi Zhang, Masao Utiyama, Eiichro Sumita, Graham Neubig, and Satoshi Nakamura. 2017. NICT-NAIST System for WMT17 Multimodal Translation Task. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 477–482, Copenhagen, Denmark.

A Significance tests

Tables 11 and 12 show the Wilcoxon signed-rank test used to create the clustering of the systems.

English→German

	LIUMCVC_MNMT_C	UvA-TiCC_IMAGINATION_U	NICT_NMTrerank_C	CUNI_NeuralMonkeyTextualMT_U	DCU-ADAPT_MultiMT_C	LIUMCVC_NMT_C	CUNI_NeuralMonkeyMultimodalMT_U	UvA-TiCC_IMAGINATION_C	CUNI_NeuralMonkeyMultimodalMT_C	OREGONSTATE_2NMT_C	CUNI_NeuralMonkeyTextualMT_C	OREGONSTATE_1NMT_C	SHEF_ShefClassProj_C	SHEF_ShefClassInitDec_C	BASELINE	AFRL-OHIOSTATE_MULTIMODAL_C
LIUMCVC_MNMT_C		2.8e-2	3.1e-4	4.8e-5	1.1e-6	4.5e-8	6.6e-10	6.2e-13	7.9e-18	1.0e-18	2.9e-17	3.2e-18	5.2e-26	1.2e-30	5.5e-41	3.1e-43
UvA-TiCC_IMAGINATION_U			4.9e-2	1.8e-2	1.2e-3	6.8e-5	3.5e-6	2.0e-8	3.2e-12	2.1e-13	4.7e-12	5.6e-13	3.2e-19	9.1e-24	3.9e-33	1.2e-34
NICT_NMTrerank_C				-	-	1.3e-2	1.6e-3	8.1e-5	9.6e-8	1.0e-8	4.6e-8	9.8e-9	1.3e-13	2.2e-17	2.2e-26	4.0e-28
CUNI_NMT_U					-	4.3e-2	1.3e-2	6.8e-4	2.4e-6	4.9e-7	1.3e-6	3.1e-7	1.6e-11	6.8e-15	3.8e-23	4.7e-25
DCU-ADAPT_MultiMT_C						-	4.0e-2	4.9e-3	3.0e-5	3.3e-6	1.3e-5	2.2e-6	1.4e-10	5.3e-14	9.8e-23	2.4e-24
LIUMCVC_NMT_C							-	-	5.7e-3	1.4e-3	2.2e-3	7.5e-4	2.2e-06	8.5e-09	2.0e-15	4.0e-17
CUNI_MNMT_U								-	1.7e-2	5.7e-3	5.6e-3	2.8e-3	4.5e-06	2.7e-08	1.2e-15	9.7e-18
UvA-TiCC_IMAGINATION_C									-	2.8e-2	3.3e-2	1.3e-2	6.2e-05	3.3e-07	3.5e-14	5.0e-16
CUNI_NMMT_C										-	-	-	6.5e-03	1.5e-04	1.5e-10	1.7e-12
OREGONSTATE_2NMT_C											-	-	1.9e-02	1.1e-03	2.0e-09	3.6e-11
CUNI_NMT_C												-	4.7e-02	3.3e-03	8.3e-08	2.1e-09
OREGONSTATE_1NMT_C													-	1.2e-02	9.1e-07	2.6e-08
SHEF_ShefClassProj_C														-	4.7e-05	1.3e-06
SHEF_ShefClassInitDec_C															2.0e-03	1.2e-04
BASELINE																-
AFRL-OHIOSTATE_MULTIMODAL_C																

Table 11: English → German Wilcoxon signed-rank test at p-level p $\leq$ 0.05. '-' means that the value is higher than 0.05.

English→French	NICT_NMTrerank_C	CUNI_NeuralMonkeyMultimodalMT_C	DCU-ADAPT_MultiMT_C	LIUMCVC_MNMT_C	OREGONSTATE_2NeuralTranslation_C	CUNI_NeuralMonkeyTextualMT_C	OREGONSTATE_1NeuralTranslation_C	LIUMCVC_NMT_C	SHEF_ShefClassInitDec_C	SHEF_ShefClassProj_C	BASELINE
NICT_NMTrerank_C		-	-	1.0e-03	2.5e-05	5.5e-08	1.5e-09	1.8e-07	2.2e-16	6.8e-15	5.8e-26
CUNI_NeuralMonkeyMultimodalMT_C			-	8.8e-03	2.9e-04	5.2e-07	5.0e-08	2.9e-06	2.4e-14	1.1e-13	2.3e-24
DCU-ADAPT_MultiMT_C				2.2e-02	1.3e-03	6.7e-06	7.9e-07	2.3e-05	2.2e-12	1.3e-11	3.9e-21
LIUMCVC_MNMT_C					-	1.7e-02	2.7e-03	1.2e-02	5.3e-07	2.0e-06	5.9e-14
OREGONSTATE_2NeuralTranslation_C						-	-	-	1.7e-04	1.8e-04	7.3e-11
CUNI_NeuralMonkeyTextualMT_C							-	-	7.7e-03	7.7e-03	5.0e-08
OREGONSTATE_1NeuralTranslation_C								-	2.4e-02	1.8e-02	2.3e-07
LIUMCVC_NMT_C									1.8e-02	1.6e-02	5.6e-07
SHEF_ShefClassInitDec_C										-	3.2e-04
SHEF_ShefClassProj_C											1.8e-03
BASELINE											

Table 12: English → French Wilcoxon signed-rank test at p-level $p \leq 0.05$. '-' means that the value is higher than 0.05.