A Shared Task on Multimodal Machine Translation and Crosslingual Image Description

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

Stella Frank, Khalil Sima'an and **Desmond Elliott**
ILLC, University of Amsterdam, The Netherlands
{s.c.frank, k.simaan, d.elliott}@uva.nl

Abstract

This paper introduces and summarises the findings of a new shared task at the intersection of Natural Language Processing and Computer Vision: the generation of image descriptions in a target language, given an image and/or one or more descriptions in a different (source) language. This challenge was organised along with the Conference on Machine Translation (WMT16), and called for system submissions for two task variants: (i) a translation task, in which a source language image description needs to be translated to a target language, (optionally) with additional cues from the corresponding image, and (ii) a description generation task, in which a target language description needs to be generated for an image, (optionally) with additional cues from source language descriptions of the same image. In this first edition of the shared task, 16 systems were submitted for the translation task and seven for the image description task, from a total of 10 teams.

1 Introduction

In recent years, significant research has been done to address problems that require joint modelling of language and vision. Examples of popular applications involving both Natural Language Processing (NLP) and Computer Vision (CV) include image description generation and video captioning (Bernardi et al., 2016), image retrieval based on textual and visual cues (Feng and Lapata, 2010), visual question answering (Yang et al., 2015), among many others (see (Ramisa et al., 2016) for more examples). With very few exceptions (Grubinger et al., 2006; Funaki and Nakayama, 2015; Gao et al., 2015), these applications are inherently monolingual and existing work explore mostly English data. In an attempt to push this interdisciplinary field to incorporate a multilingual component, we propose the first shared task on two new applications: Multimodal Machine Translation and Crosslingual Image Description. Generally speaking, this shared task targets the generation of image descriptions in a target language, given an image and one or more descriptions in a different (source) language. More specifically, the task can be addressed from two perspectives:

1. Task 1: a **Multimodal Machine Translation** task, which takes a source language description and translates it into the target language, where this process can be supported by information from the image; see Figure 1, and
2. Task 2: a **Crosslingual Image Description** task, which takes an image and generates a description for it in the target language, where this process can be supported by the source language description; see Figure 2.

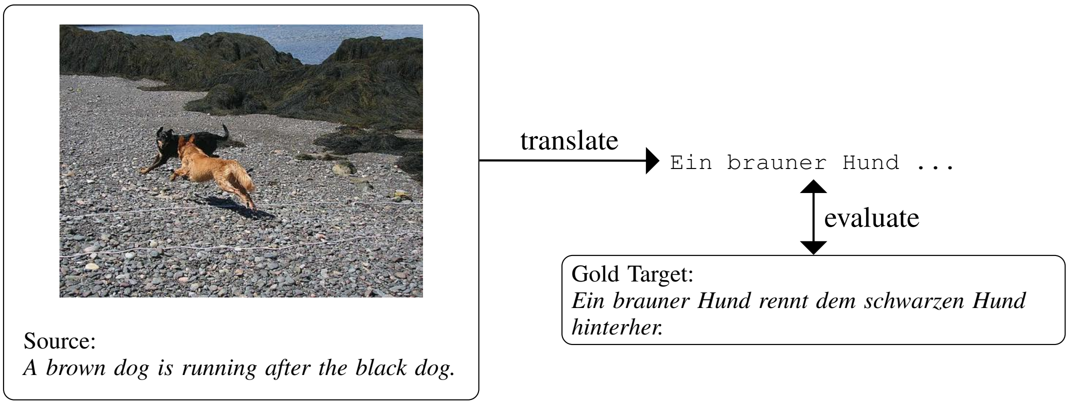
This shared task has the following main goals:

- To push existing work on multimodal language processing towards multilingual multimodal language processing.
- To investigate the effectiveness of information from images in machine translation.
- To investigate the effectiveness of crosslingual textual information in image description generation.

The challenge was organised in the framework of the well-established WMT series of shared tasks.[1] Participants were called to submit systems focusing on either or both of these task variants. The tasks differ in the training data and in

[1] http://www.statmt.org/wmt16/

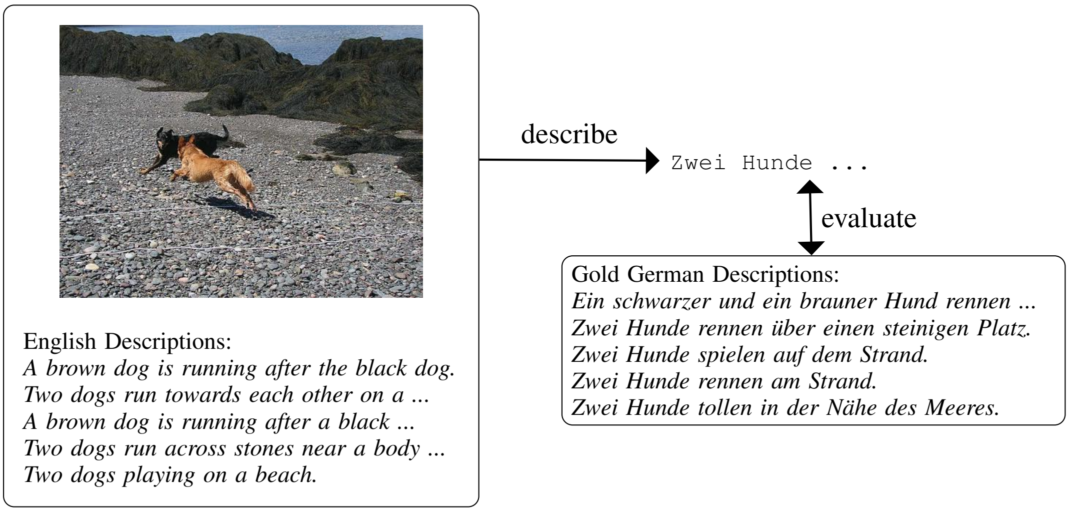
Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 543–553,
Berlin, Germany, August 11-12, 2016. ©2016 Association for Computational Linguistics



Gold Target:
Ein brauner Hund rennt dem schwarzen Hund hinterher.

Source:
A brown dog is running after the black dog.

Figure 1: Multimodal Machine Translation (Task 1). English and translated German image descriptions are grounded to an image.



Gold German Descriptions:
Ein schwarzer und ein brauner Hund rennen ...
Zwei Hunde rennen über einen steinigen Platz.
Zwei Hunde spielen auf dem Strand.
Zwei Hunde rennen am Strand.
Zwei Hunde tollen in der Nähe des Meeres.

English Descriptions:
A brown dog is running after the black dog.
Two dogs run towards each other on a ...
A brown dog is running after a black ...
Two dogs run across stones near a body ...
Two dogs playing on a beach.

Figure 2: Multilingual Image Description (Task 2). The data consist of *independently* produced image descriptions in English and German.

	Sentences	Types	Tokens	Avg. length
Task 1: Translations				
English	31,014	11,420	357,172	11.9
German		19,397	333,833	11.1
Task 2: Descriptions				
English	155,070	22,815	1,841,159	12.3
German		46,138	1,434,998	9.6

Table 1: Corpus-level statistics about the translation and the description data over 31,014 images.

the way the target language descriptions are evaluated: against one translation of the corresponding source description (translation variant) or against five descriptions of the same image in the target language, created independently from the corresponding source description (image description variant). The data used for both tasks is an extended version of the Flickr30K dataset. Participants were also allowed to use external data and resources for unconstrained submissions.

Participants were encouraged to make use of both the sentences and the images as part of their submissions but they were not required to do so. The baseline systems for the translation task were a text-only Moses phrase-based statistical machine translation (SMT) model (Koehn et al., 2007) and the GroundedTranslation multilingual image description model (Elliott et al., 2015) (in particular, the MLM→LM variant). The baseline system for the description generation task was also the GroundedTranslation model.

In this paper we describe the data, image features and participants of the shared task (Sections 2 and 3), present its main findings (Section 4), and discuss interesting issues and directions for future research (Section 5).

2 Datasets and image features

We created a new dataset for the shared task by extending the Flickr30K dataset (Young et al., 2014) into another language. The **Multi30K** dataset (Elliott et al., 2016) contains two types of multilingual data: a corpus of English sentences *translated* into German (used for Task 1), and a corpus of *independently collected* English and German sentences (used for Task 2). For the translation corpus, one sentence (of five) was chosen for professional translation such that the final dataset is a combination of short, medium, and long length sentences. The second corpus consists of crowdsourced descriptions gathered from Crowdflower,[2] where each worker generated an independent description of the image. We used a translation of the original instructions used to gather the English sentences, in order to ensure as much similarity across the German and English descriptions as possible. Table 1 presents an overview of the data available for each task.

The images are publicly available[3] but to encourage participation we released two types of features extracted from the images. The use of such features was not mandatory, and participants could also extract image features from the original images in the Flickr30K dataset using their own algorithms. We released features extracted from the VGG-19 Convolutional Neural Network (CNN), as described in (Simonyan and Zisserman, 2015), from the FC_7 (relu7) and $\text{CONV}_{5,4}$ layers. We extracted these image features using Caffe RC2[4] with the `matlab_features_reference` code from NeuralTalk.[5]

3 Participants

Ten teams submitted a total of 23 systems for the two tasks. The teams are listed in Table 2. In what follows, we summarise the participating systems.

CMU (Task 1) The approach incorporates global and regional visual features with textual features from English (source) and German (target) to jointly train a Recurrent Neural Network (RNN). Visual features extracted from a region-based convolution neural network (RCNN) are designed to be appended in the head/tail of the textual feature or dissipated in parallel long short term memory (LSTM) threads to assist the LSTM reader in computing a representation. For rescoring, an additional bilingual dictionary is used to select the best sentence from candidates generated by five different models. The submission is thus unconstrained, with the German-English Dictionary from GLOSBE[6] used as additional resource.

CUNI (Tasks 1 and 2) The method is a system combination which implements the attentive neural Machine Translation (MT) (Bahdanau et al., 2014). The input of the decoder is a linear combination of the image features obtained from the penultimate layer of the VGG16 convolutional network (Simonyan and Zisserman, 2015) and two recurrent encoders coding the source sentence and its translation obtained from the Moses system. The Moses system uses the with additional language models based on coarse bitoken classes (Stewart et al., 2014).

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

[3] http://illinois.edu/fb/sec/229675

[4] http://github.com/BVLC/caffe/releases/tag/rc2

[5] http://github.com/karpathy/neuraltalk/tree/master/matlab_features_reference

[6] https://glosbe.com/en/de/

ID	Participating team
CMU+NTU	Carnegie Melon University (Huang et al., 2016)
CUNI	Univerzita Karlova v Praze (Libovický et al., 2016)
DCU	Dublin City University (Hokamp and Calixto, 2016)
DCU-UVA	Dublin City University & Universiteit van Amsterdam (Calixto et al., 2016)
HUCL	Universität Heidelberg (Hitschler et al., 2016)
IBM-IITM-Montreal-NYU	IBM Research India, IIT Madras, Université de Montréal & New York University
LIUM	Laboratoire d'Informatique de l'Université du Maine (Caglayan et al., 2016)
SHEF	University of Sheffield (Shah et al., 2016)
UPC	Universitat Politècnica de Catalunya (Rodríguez Guasch and Costa-jussà, 2016)
UPC$_b$	Universitat Politècnica de Catalunya

Table 2: Participants in the WMT16 multimodal machine translation shared task.

DCU (Task 1) Both submissions from DCU are neural MT systems with an attention mechanism on the source-side representation (Bahdanau et al., 2014). The first submission is text-only, and the second submission includes the FC$_7$ image features in the target-side decoder initial state. The FC$_7$ features are passed through a 3-layer fully-connected feedforward network with Tanh non-linearities, and then summed with the final state of the source-side representation. This summed representation is passed through another feed-forward layer, and becomes the initial state for the decoder recurrent transition. The main novelty of our system is that we use a minimum-risk training objective to directly optimise the model for Meteor, instead of the word-level cross entropy loss function which is currently standard for NMT systems. This idea comes from (Shen et al., 2016), although our implementation is somewhat different than the idea outlined in that work. To optimise for expected Meteor, we take up to 100 samples from our model, compute an expectation over these samples, and use Stochastic Gradient Descent to directly optimise the model on this expected score.

DCU-UVA (Task 1) The approach integrates separate attention mechanisms over the source language and the CONV$_{5,4}$ visual features in a single decoder. The source language was represented using a bidirectional RNN with Gated Recurrent Units (GRU); the images were represented as 196x512 matrix from the pre-trained VGG-19 convolutional network. A separate, time-dependent context vector was constructed for the source sentence and the visual features, which were merged into a single multimodal context vector. This time-dependent multimodal context vector was input into the target language decoder, along with the previous hidden state and the previously emitted word. Throughout, 300D word embeddings, 1000D hidden states, and 1000D context vectors were used; the source and target languages were estimated over the entire vocabularies.

HUCL (Task 1) The submitted system for the constrained task extends a standard SMT pipeline by a re-ranking component that makes use of multimodal information. The cdec decoder (Dyer et al., 2010) was used to produce hypothesis lists, which were re-scored by comparison with similar image captions from the training corpus using the pivoting approach described in Hitschler et al. (2016), with some minor differences: Because all data for the shared task was parallel, a constrained model was built by employing a source side matching approach inspired by standard translation memories, instead of retrieving matching captions in the target language by pivoting on larger image-caption data as described by Hitschler et al. (2016), which would have resulted in an unconstrained model. That is, the submission resorted to textual similarity (as measured by the TF-IDF score (Spärck Jones, 1972)) on the source language side as well as visual similarity (as measured by the Euclidean distance between the feature values of the FC$_7$ layer of the

VGG16 deep convolutional model (Simonyan and Zisserman, 2015), supplied by the task organisers) for retrieval of matches. The retrieval model architecture was identical to that in Hitschler et al. (2016). Instead of TF-IDF, a modified version of BLEU (Papineni et al., 2002) was used in order to re-score hypotheses based on the target-language text of retrieved captions. Fixed settings were used for some parameters ($d = 90$, $b = 0.01$ and $k_m = 20$), while k_r and λ were optimised on the validation set (parameters as defined in (Hitschler et al., 2016)).

IBM-IITM-Montreal-NYU (Tasks 1 and 2)[7] The approach for Task 1 is similar to that of (Elliott et al., 2015) with two differences. First, instead of using a RNN based encoder for the source (English) sentence, a simple bag of words encoder is used. In other words, the representation of the source sentence is simply a sum of the representations of the words in it. These word representations are randomly initialised and then learned during training. Second, unlike (Elliott et al., 2015), the image and source sentence representation are fed at every timestep to the target RNN decoder. The approach for Task 2 is same as that for Task 1, except that now instead of having a single source sentence representation, the representations of all the five source sentences are concatenated. This is then further concatenated with the image representation and the result is fed at every timestep to the target decoder. The FC_7 features for images as provided by the task organisers are used and tuned during training. The source and target RNNs contain 512 hidden neurons and the word embeddings are also of size 512. The models for both the tasks are trained for 10 epochs. For the unconstrained setup, the MSCOCO dataset, which contains English captions for images, was explored. These English captions were translated into German using IBM's translation services and then these pseudo Image-English-German tuples were used as additional training data, together with the training data provided by the task organisers. These are referred to as pseudo tuples since the German captions were machine translated and not human generated.

LIUM (Tasks 1 and 2) All submissions from LIUM are constrained. LIUM_1_MosesNMTRnnLMSent2Vec_C and LIUM_1_MosesNMTRnnLMSent2VecVGGFC7_C are phrase-based systems based on Moses (14 standard features plus operation sequence models. They include re-scoring with several models and more particularly with a continuous space language model (CSLM) and a neural MT system (see TextNMT system). The CSLMs can use image feature maps as auxiliary data, in order to provide some context to the probabilities. The LIUM_1_TextNMT_C and LIUM_2_TextNMT_C systems are monomodal (text-only) fully neural MT systems similar to the one proposed by DL4MT school.[8] They are made of a bidirectionnal recurrent encoder followed by a conditional Gated Recurrent Unit decoder which embeds an attention mechanism. The difference between the two systems is the training and development data, as provided by the organisers. Finally, the LIUMCVC_1_MultimodalNMT_C and LIUMCVC_2_MultimodalNMT_C are an extension of the previous systems, where an additional input is given: the convolutional feature maps extracted with a very deep ResNet (up to 152 layers) from the images (He et al., 2015). The attention mechanism is shared across the two modalities (with softmax activations remaining distinct). The architecture of the decoder is the same as before. The difference between the two systems is again the training and development data.

SHEF (Task 1) Both submissions from the Sheffield team are constrained, each focusing on one language direction: SHEF_1_en-de-Moses-rerank_C cover the official task direction (English-German), while SHEF_1_de-en-Moses-rerank_C covers the opposite direction (German-English). Our proposed systems are standard phrase-based statistical MT systems based on the Moses decoder, trained on the provided data. We investigate how image features can be used to re-rank the n-best output of the SMT model, with the aim of improving performance by grounding the translations on images. Image features from a CNN are used to re-rank the n-best list along with standard Moses features. We also propose an alternative scheme for the German-to-English direction, where terms in the English image descriptions are matched with 1,000 WordNet synsets, and the probability of these synsets occurring in the image estimated using CNN predictions on the images.

[7] Systems submitted by Amrita Saha, Mitesh M. Khapra, Janarthanan Rajendran, Sarath Chandar, Kyunghyun Cho

[8] `http://dl4mt.computing.dcu.ie/`

The aggregated probabilities are then used to re-rank the n-best list, with the intuition that the best translations should contain these entities. Our submissions to re-rank the n-best translations with image vectors are able to marginally outperform the strong, text-only baseline Moses system for both directions.

UPC (Task 1) Bidirectional Recurrent Neural Networks (BiRNNs) have shown outstanding results on sequence-to-sequence learning tasks. This architecture becomes especially interesting for multimodal machine translation task, since BiRNNs can deal with images and text. On most translation systems the same word embedding is fed to both BiRNN units. In our submission, we enhance a baseline sequence-to-sequence system (Elliott et al., 2015) by using double embeddings. These embeddings are trained on the forward and backward directions of the input sequence. The system was trained, validated and tested using the task's dataset only.

UPC$_b$ (Task 2)[9] The two submissions from UPC$_b$ use the same method with different training data, one is constrained (UPC_2_MNMT_C), while the other is unconstrained (UPC_2_MNMT2_U). Captions are generated from two different directions. One caption is generated through translating the captions in the source language directly using the method proposed in (Bahdanau et al., 2014). The other one is generated based on the image feature using method proposed in (Vinyals et al., 2015). After that, an SVM-based model decides which one is better according to the sentence's score from a language model and the score from the model that generated the sentence. The only difference between the two submissions is that the unconstrained one used Task 1 dataset in the training of text translator.

Baseline - GroundedTranslation (Tasks 1 & 2) This method follows (Elliott et al., 2015):[10] A source language multimodal RNN model is initialised with a visual feature vector (i.e., a multimodal model for the source language). The final hidden state is then used to initialise a target language model, which generates the target language description. The source language multimodal RNN language model was trained until the loss stopped falling on the validation data. The target model was initialised with the final hidden state transferred from the source model and trained until the loss stopped falling on the validation data. The source model and target models were parameterised with 300D word embeddings and 1000D GRU hidden states; the source model was initialised with the 4096D FC$_7$ visual feature vector; for Task 1, the target model was initialised with a 1000D source model feature vector; for Task 2 the feature vectors corresponding to each source language description were summed into a 1000D feature vector. For both tasks, we found the optimal combination of target model language generation timesteps and beam width size using grid search.

Baseline - Moses (Task 1) This baseline system uses text-only information. It is a standard phrase-based SMT system built using the Moses toolkit (Koehn et al., 2007). The models were trained using the extended version of Flickr30K parallel dataset provided for the task only ($29,000$ sentence pairs), and tuned with the official validation dataset ($1,014$ segment pairs). Default settings and features in Moses were used, with a 4-gram language model trained on the target side of the parallel data.

4 Results

Tables 3 and 4 present the official results for the Multimodal Machine Translation and Crosslingual Image Description tasks. We evaluated the submissions based on Meteor (Denkowski and Lavie, 2014) (primary), BLEU (Papineni et al., 2002) and TER (Snover et al., 2006) using `MultEval` (Clark et al., 2011)[11] with default parameters.

4.1 Task 1

Table 4 shows the final results for the Multimodal Machine Translation task on the official test set, where systems are ranked by their Meteor scores. Meteor, BLEU and TER were computed based on the single reference (human translation) provided for the test set. For Meteor, we replaced the default version by the latest version of the metric (Meteor Version 1.5). Both reference and system submissions were first normalised for punctuation.

[9] Systems submitted by Zhiwen Tang and Marta Ruiz Costa-jussà; code available: https://github.com/Z-TANG/re-scorer.

[10] https://github.com/elliottd/GroundedTranslation

[11] https://github.com/jhclark/multeval

System submissions that preserved casing or had been tokenised were further processed for lowercasing and detokenisation.[12] For all of these preprocessing steps, we used Moses scripts.[13]

It is interesting to note that while the three evaluation metrics do not fully agree on the ranking of participating systems, their overall Pearson's correlation (English-German direction) is very high: 0.98 between Meteor and BLEU, and -0.97 between Meteor and TER.

The three winning submissions from the LIUM and SHEF teams are heavily based on the output of a standard phrase-based SMT system (Moses) built using only the shared task data. This is a remarkable result, given the size of the dataset: 29,000 parallel segments. They all use additional features to re-rank the k-best output of a text-only phrase-based system, including visual features, although these seem to play a minor role and lead to only marginally better results.

Submissions based on the output of a Moses translation model – like the main baseline (1_en-de-Moses_C) – have very similar Meteor scores. In fact, SHEF_1_en-de-Moses-rerank_C and CMU+NTU_1_MNMT+RERANK_U are not considered significantly different from this baseline Shah et al. (2016) provide some analysis on the differences between SHEF_1_en-de-Moses-rerank_C and 1_en-de-Moses_C. They show that the output of these systems differ in 260 out of the 1,000 segments. However, despite differences in the actual translations, the Meteor scores for many of these cases may be the same/close.

Disappointingly, truly multimodal systems, which in most cases use neural MT approaches (e.g. CUNI_1_MMS2S-1_C, DCU_1_min-risk-multimodal_C) do not fare as well as the text-only SMT systems (or those followed by multimodal-based translation rescoring), except when additional resources are used for rescoring translations (CMU_1_MNMT+RERANK_U).

Only two submissions made use of additional data (unconstrained submissions, _U) and in both cases it proved helpful in comparison with the constrained submissions by the same teams.

4.2 Task 2: Crosslingual Image Description

Table 4 presents the final results for the Crosslingual Image Description task. Meteor is the primary evaluation measure because it has been shown to have a much stronger correlation with human judgements than BLEU or TER for this task (Elliott and Keller, 2014). The data for this task was lowercased and had punctuation removed where necessary.

The strongest performing *constrained* submission (LIUM_2_TextNMT_C) does not use any visual features. Including multimodal features (i.e., LIUM_2_MultimodalNMT_C) results in a 2.8 Meteor drop in performance for that model type. The baseline system 2_GroundedTranslation_C outperformed all but these two systems. In general, there is a wide range of performances, and an intriguing discrepancy between Meteor and BLEU rankings. This discrepancy was much larger than the one observed in Task 1, where the overall ranking trend for all metrics is similar. We believe the difference between metrics in Task 2 is due to the different ways in which these metrics use multiple references (which are only available for Task 2). While Meteor (and TER) will match the single closest reference (the entire sentence) to the system output, BLEU allows n-grams from different references to be used for its n-gram matching.

Only two groups submitted *unconstrained* runs, marked in grey and with _U in Table 4. The IBM-IITM-Montreal-NYU_2_NeuralTranslation_U submission resulted in a small improvement over the IBM-IITM-Montreal-NYU_2_NeuralTranslation_C submission, but the UPC_2_MNMT_U resulted in a small decrease compared to the analogous constrained submission UPC_2_MNMT_C.

5 Discussion

Although the Multimodal Machine Translation and Crosslingual Description tasks are based on the same collection of images, there are a number of important differences in the textual data, outlined below, which lead to different patterns of results for both tasks.

The nature of the sentences The sentences in Task 1 are professional translations, whereas the sentences in Task 2 are independent descriptions. The differences between translations and descriptions may affect the performance of image de-

[12] We note that MultEval does not perform any normalisation of the segments. Scores with tokenised texts would be consistently higher.

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

System ID	**Meteor** ↑	BLEU ↑	TER ↓
English-German			
●LIUM_1_MosesNMTRnnLMSent2Vec_C	53.2	34.2	48.7
●LIUM_1_MosesNMTRnnLMSent2VecVGGFC7_C	53.2	34.1	48.7
●*SHEF_1_en-de-Moses-rerank_C	52.6	32.8	49.8
1_en-de-Moses_C	52.5	32.5	50.2
*CMU_1_MNMT+RERANK_U	51.9	33.6	52.4
HUCL_1_RROLAPMBen2de_C	51.5	32.2	51.1
CMU_1_MNMT_C	50.8	35.1	49.2
DCU_1_min-risk-baseline_C	49.7	31.8	52.6
LIUM_1_TextNMT_C	49.2	32.5	51.6
DCU_1_min-risk-multimodal_C	48.4	32.5	49.8
CUNI_1_MMS2S-1_C	46.5	29.7	53.5
DCU-UVA_1_doubleattn_C	46.4	27.4	59.7
LIUMCVC_1_MultimodalNMT_C	45.0	27.8	57.3
DCU-UVA_1_imgattninit_C	44.1	26.5	60.1
IBM-IITM-Montreal-NYU_1_NeuralTranslation_U	39.1	21.8	61.9
UPC_1_SIMPLE-BIRNN-DEMB_C	37.7	22.1	60.4
IBM-IITM-Montreal-NYU_1_NeuralTranslation_C	31.1	16.0	69.4
1_GroundedTranslation_C	24.7	9.4	77.2
German-English			
●*SHEF_1_de-en-Moses-rerank_C	36.5	39.8	41.0
●1_de-en-Moses_C	36.2	38.1	40.8
HUCL_1_RROLAPMBde2en_C	35.1	37.0	42.4

Table 3: Official results for the WMT16 Multimodal Machine Translation task. The baseline results are underlined. Systems with grey background indicate use of resources that fall outside the constraints provided for the shared task. The winning submissions are indicated by a ●. These are the top-scoring submission 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 a * indicate those that are not significantly different from the main baseline (1_Moses_C) according to the same test.

scription models relative to the translation models. This can be seen by comparing the results of teams that submitted the same systems (but separately trained) to both tasks: LIUM, IBM-IITM-Montreal-NYU, and the GroundedTranslation baseline. The LIUM and IBM-IITM-Montreal-NYU submissions seem to benefit from training over translation data instead of the description data, as suggested by the higher Meteor scores achieved in Task 1 (1 reference) vs. Task 2 (5 references); the GroundedTranslation submissions exhibit the opposite effect (this may be explained by the fact that this submission is an image description model and not a translation model). We hypothesize that the differences in performance may originate from the possibility that (a) the description data is merely a *comparable* corpus instead of a *parallel* corpus leading to noisier pairing of source-target pairs, and/or (b) in the description task the training data is less compatible with the test data than in the translation task. This demands further exploration.

The number of training examples Submissions for Task 1 are trained over 29,000 parallel instances (one sentence pair per image), whereas submissions for Task 2 are trained over 145,000 (five independent sentences per language per image). The number of training examples for each task further complicates the analysis of the difference in performance between the two tasks, as the larger-data scenario in Task 2 does not lead to a straightforward improvement in performance. The type and the quality of the parallel translation data – despite its small size – makes it relatively easy to train high-performing translation models, as we can see by comparing the absolute Meteor scores

System ID	**Meteor** ↑	BLEU ↑	TER ↓	Visual Features?
English-German				
● LIUM_2_TextNMT_C	35.1	23.8	62.1	—
LIUM_2_MultimodalNMT_C	32.3	19.2	70.0	ResNet
2_GroundedTranslation_C	31.2	15.8	76.4	FC_7
IBM-IITM-Montreal-NYU_2_NeuralTranslation_U	29.5	9.7	89.0	FC_7
IBM-IITM-Montreal-NYU_2_NeuralTranslation_C	29.1	17.8	60.0	FC_7
CUNI_2_MMS2S-2_C	13.1	1.2	73.3	FC_7
UPC_b_2_MNMT_C	12.1	1.5	63.1	FC_7
UPC_b_2_MNMT_U	11.7	1.0	82.2	FC_7

Table 4: Official results for the WMT16 Crosslingual Image Description task. The baseline results are underlined. Systems with grey background indicate use of resources that fall outside the constraints provided for the shared task. The winning submission, indicated by a ●, is significantly different from all other submissions based on Meteor scores. Submissions marked with a * are not significantly different compared to the baseline (2_GroundedTranslation_C).

in Tables 3 and 4. In fact, it is quite remarkable that both statistical and neural MT approaches performed so well with only 29,000 sentence pairs for training, particularly for English→German translation. In different text domains (e.g. Europarl, News), this language pair and direction is well known as a challenging case. The two languages are structurally distant and the target language – German – is morphologically richer than English, which poses a problem in machine translation particularly when not enough training instances are available with examples of the various morphological variants of target words. The fact that the performance for Task 1 was so high seems to indicate that the data for this task is much simpler and probably significantly more repetitive than data used in other shared tasks, for example, the News translation task at WMT (Bojar et al., 2015).

The amount of evaluation data Task 1 submissions are evaluated against one reference translation and Task 2 submissions are evaluated against five independent sentences. The larger number of references for Task 2 should make it easier for submissions to achieve high Meteor scores but this is not borne out in the results. One reason for this could be that each independently collected description had a free choice in what to describe and how to describe it (Elliott and Keller, 2014). This has led to collected descriptions that are not translations of their English counterparts. We could collect five professionally translated references for each image to study this issue. We would expect the absolute Meteor scores for Task 1 to increase with more references (Dreyer and Marcu, 2012); however, we should also bear in mind that the image descriptions are quite simple and there is likely to be very high similarity among translations.

Further research is needed to determine whether having more parallel translation data or more references for evaluation will lead to better performance for both tasks. However, this data would be very expensive to collect. Collecting more independent descriptions would be significantly cheaper.

Use of visual information The use of visual information had very different effects in the two tasks. While for Task 1 this information only proved marginally useful in indirect ways (i.e. rescoring k-best translations), visual information featured prominently in submissions for Task 2: six submissions used the FC_7 features, one submission used features extracted from the ResNet-50 network, and one submission used no visual features. The submission with ResNet-50 features outperformed all submissions with FC_7 features, which is not surprising given the difference in object categorisation performance between the models (4.49% top-5 error on the ILSVRC validation data (Russakovsky et al., 2014) compared to 7.1% error). However, the submission without visual features achieved the best performance for Task 2.

In light of our aim of furthering multimodal research with multilingual multimodal data, this is a somewhat disappointing result. However, we believe that it only reinforces the call to develop more robust models that can integrate visual and linguistic features into a single model. Building more realistic and challenging datasets is also an interesting direction for future research.

Acknowledgments

SF was supported by European Union's Horizon 2020 research and innovation programme under grant agreement nr. 645452. DE and KS are supported by the NWO Vici grant nr. 277-89-002.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

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. *CoRR*, abs/1601.03896.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal.

Ozan Caglayan, Walid Aransa, Yaxing Wang, Marc Masana, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, and Joost van de Weijer. 2016. Does multimodality help human and machine for translation and image captioning? In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany.

Iacer Calixto, Desmond Elliott, and Stella Frank. 2016. Dcu-uva multimodal mt system report. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, pages 176–181, Portland, Oregon.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland.

Markus Dreyer and Daniel Marcu. 2012. Hyter: Meaning-equivalent semantics for translation evaluation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 162–171, Montréal, Canada, June.

Chris Dyer, Adam Lopez, Juri Ganitkevitch, Johnathan Weese, Ferhan Ture, Phil Blunsom, Hendra Setiawan, Vladimir Eidelman, and Philip Resnik. 2010. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 7–12, Uppsala, Sweden.

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*, pages 452–457, Baltimore, Maryland.

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. Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German Image Descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, Berlin, Germany.

Yansong Feng and Mirella Lapata. 2010. Topic models for image annotation and text illustration. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 831–839, Los Angeles, California.

Ruka Funaki and Hideki Nakayama. 2015. Image-mediated learning for zero-shot cross-lingual document retrieval. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 585–590, Lisbon, Portugal.

Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. 2015. Are you talking to a machine? dataset and methods for multilingual image question answering. *Advances in Neural Information Processing Systems*, pages 2287–2295.

Michael Grubinger, Paul D. Clough, Henning Muller, and Thomas Desealers. 2006. The IAPR TC-12 benchmark: A new evaluation resource for visual information systems. In *Proceedings of the Language Resources and Evaluation Conference*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

Julian Hitschler, Shigehiko Schamoni, and Stefan Riezler. 2016. Multimodal Pivots for Image Caption Translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany.

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

Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. 2016. Attention-based multimodal neural machine translation. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany.

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 *Proceedings of the 45th Annual meeting of Association for Computational Linguistics*, pages 177–180, Prague, Czech Republic.

Jindřich Libovický, Jindřich Helcl, Marek Tlustý, Ondřej Bojar, and Pavel Pecina. 2016. Cuni system for wmt16 automatic post-editing and multimodal translation tasks. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany.

Kishore Papineni, Salim Roukos, Todd Ard, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania.

Arnau Ramisa, Fei Yan, Francesc Moreno-Noguer, and Krystian Mikolajczyk. 2016. Breakingnews: Article annotation by image and text processing. *CoRR*, abs/1603.07141.

Sergio Rodríguez Guasch and Marta R. Costa-jussà. 2016. Wmt 2016 multimodal translation system description based on bidirectional recurrent neural networks with double-embeddings. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. 2014. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575.

Kashif Shah, Josiah Wang, and Lucia Specia. 2016. Shef-multimodal: Grounding machine translation on images. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum Risk Training for Neural Machine Translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations*.

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

Karen Spärck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21.

Darlene Stewart, Roland Kuhn, Eric Joanis, and George Foster. 2014. Coarse split and lump bilingual language models for richer source information in SMT. In *Proceedings of the Eleventh Conference of the Association for Machine Translation in the Americas*, pages 28–41, Vancouver, Canada.

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

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola. 2015. Stacked attention networks for image question answering. *CoRR*, abs/1511.02274.

Peter Young, Alice Lai, Micha 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.