**The Shared Task on Multimodal Translation and**

**Multilingual Image Description**

**Tasks:**

1. **Task 1: Multimodal translation:**
   1. 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.
2. **Task 2: Multilingual image description:**
   1. 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.

**Dataset Description:**

1. Multi30K dataset is triple aligned image dataset, with English descriptions translated into both German and French.
2. Source language descriptions are no longer observed for test images.
3. An extension of the 215 existing Multi30K dataset, and a “teaser” evaluation dataset COCO with images carefully chosen to contain ambiguities in the source language. 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.

**Participants:**

1. **Task 1: Multimodal translation:**
   1. **AFRL-OHIOSTATE (Air Force Research Laboratory & Ohio State University):**

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 a set of 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.

* 1. **CUNI (Univerzita Karlova v Praze)**

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. 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.

* 1. **DCU-ADAPT (Dublin City University)**

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. They are model IMGW, in which image features are used as words in the source-language encoder; model IMGE, where image features are used to initialise the hidden states of the forward and backward encoder RNNs; and model IMGD, where the image features are used as additional signals to initialise the decoder hidden state

* 1. **LIUMCVC (C Laboratoire d’Informatique de l’Universite du Maine & Universitat Autonoma de Barcelona Computer Vision Center)**

LIUMCVC experiment with two approaches: a multimodal attentive NMT with separate attention over source text and convolutional image features, and an NMT where global visual features are multiplicatively interacted with word embeddings. More specifically, each target word embedding is multiplied with global visual features in an elementwise fashion in order to visually contextualize word representations.

* 1. **NICT (National Institute of Information and Communications Technology & Nara Institute of Science and Technology)**

These are constrained submissions for both language pairs. First, a hierarchical phrase-based (HPB) translation system s built using Moses with standard features. Then, an attentional encoder-decoder network is trained and used as an additional feature to rerank the n-best output of the HPB system.

* 1. **OREGONSTATE (Oregon State University)**

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. Additionally, the decoder uses alignment to source words by a global attention mechanism.

* 1. **UvA-TiCC ( Universiteit van Amsterdam & Tilburg University)**

The submitted systems are Imagination models 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.

* 1. **SHEF (University of Sheffield)**

The SHEF systems utilize the predicted posterior probability distribution over the image object classes as image features. The model follows a standard encoder-decoder NMT approach using softdot attention as described in. 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.

1. **Task 2: Multilingual image description:**
   1. **CMU (Carnegie Melon University)**

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.

* 1. **CUNI (Univerzita Karlova v Praze)**

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.

**Submission Baseline:**

**Baseline — Task 1:**

The baseline system for the multimodal translation task is a text-only neural machine translation system built with the Nematus toolkit.

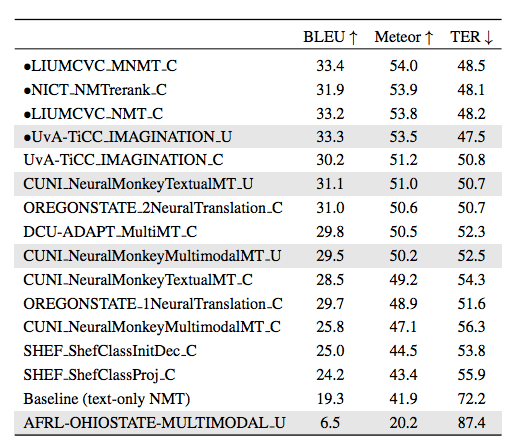
**Baseline — Task 2:**

The baseline for the multilingual image description task is an attention-based image description system trained over only the German image descriptions.

**Metric Results:**

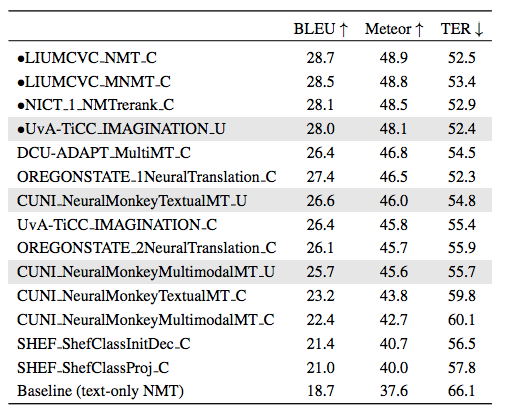
**Task 1: English -> German:**

1. **Multi30K 2017 test data**

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.

1. **Ambiguous COCO**

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.

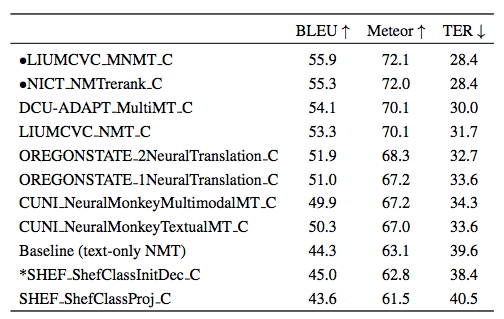
The systems are mostly in the same order as on the Multi30K 2017 test data, with the same four systems performing best. However, two systems (DCU-ADAPT MultiMT C and OREGONSTATE 1NeuralTranslation C) are ranked higher on this test set than on the in-domain Flickr dataset, indicating that they are relatively more robust and possibly better at resolving the ambiguities found in the Ambiguous COCO dataset.

**Task 1: English -> French:**

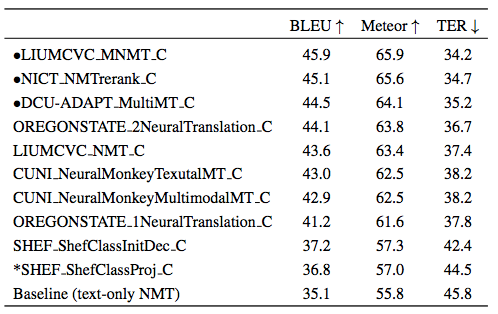
1. **Multi30K 2017 test data**

the evaluation metrics are in better agreement about the ranking of the submissions. English→French is an easier task than English→German systems, as reflected in the higher metric scores.

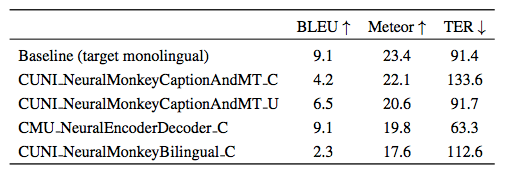
Eight out of the ten submissions outperformed the English→French baseline system. Two of the best submissions for English→German remain the best for English→French (LIUMCVC MNMT C and NICT NMTrerank C), the text-only system (LIUMCVC NMT C) decreased in performance, and no UvA-TiCC IMAGINATION U system was submitted for French.

****

1. **Ambiguous COCO**

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 OREGONSTATE 2NeuralTranslation C again perform relatively better on this dataset.

**Task 2: English -> German:**

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.

**Human Judgement Results:**

When comparing automatic and human evaluations, we can observe that they globally agree with each other with German showing better agreement than French. We point out two interesting disagreements: First, in the English→French language pair, CUNI NeuralMonkeyMultimodalMT C and DCUADAPT 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.

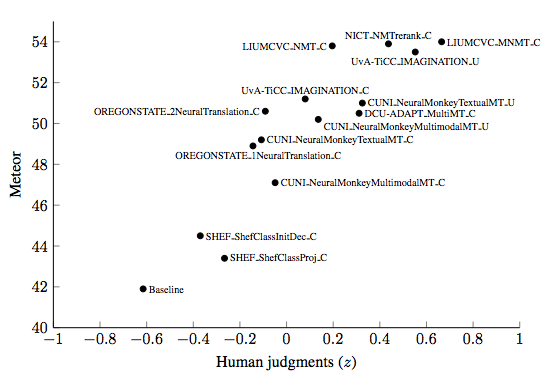
**Discussion:**

**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.

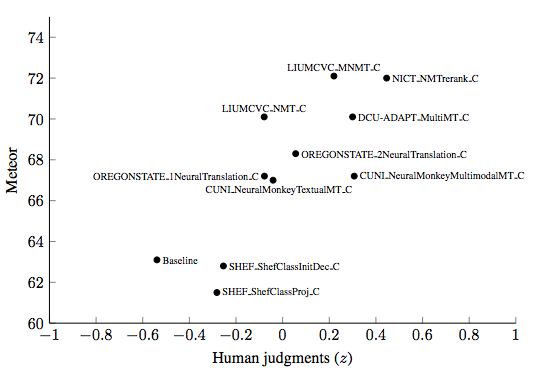
The human evaluation results are perhaps more promising: nearly all the highest ranked systems (with the exception of NICT) are multimodal.

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.

System performance on the English→German Multi30K 2017 test data as measured by human evaluation against Meteor scores:

System performance on the English→French Multi30K 2017 test data as measured by human

evaluation against Meteor scores:



**Ambiguous COCO dataset**

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.

**Conclusions:**

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.