Findings of the Third Shared Task on Multimodal Machine Translation

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, Stella Frank

Findings of the Third Shared Task on Multimodal Machine Translation

Loïc Barrault[1], **Fethi Bougares**[1], **Lucia Specia**[2],
Chiraag Lala[2], **Desmond Elliott**[3] and **Stella Frank**[4]
[1]LIUM, University of Le Mans
[2]Department of Computer Science, University of Sheffield
[3]Department of Computer Science, University of Copenhagen
[4]Centre for Language Evolution, University of Edinburgh
loic.barrault@univ-lemans.fr

Abstract

We present the results from the third shared task on multimodal machine translation. In this task a source sentence in English is supplemented by an image and participating systems are required to generate a translation for such a sentence into German, French or Czech. The image can be used in addition to (or instead of) the source sentence. This year the task was extended with a third target language (Czech) and a new test set. In addition, a variant of this task was introduced with its own test set where the source sentence is given in multiple languages: English, French and German, and participating systems are required to generate a translation in Czech. Seven teams submitted 45 different systems to the two variants of the task. Compared to last year, the performance of the multimodal submissions improved, but text-only systems remain competitive.

1 Introduction

The Shared Task on Multimodal Machine Translation tackles the problem of generating a description of an image in a target language using the image itself and its English description. This task can be addressed as either a pure translation task from the source English descriptions (ignoring the corresponding image), or as a multimodal translation task where the translation process is guided by the image in addition to the source description.

Initial results in this area showed the potential for visual context to improve translation quality (Elliott et al., 2015; Hitschler et al., 2016). This was followed by a wide range of work in the first two editions of this shared task at the WMT in 2016 and 2017 (Specia et al., 2016; Elliott et al., 2017).

This year we challenged participants to target the task of multimodal translation, with two variants:

- 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 source-target parallel sentences and their corresponding images.

- Task 1b: **Multisource multimodal translation** takes an image with a description in three source languages that is then translated into a target language. The training data consists of source-target parallel data and their corresponding images, but where the source sentences are presented in three different languages, all parallel.

Task 1 is identical to previous editions of the shared task, however, it now includes an additional Czech target language. Therefore, participants can submit translations to any of the following languages: German, French and Czech. This extension means the Multi30K dataset (Elliott et al., 2016) is now 5-way aligned, with images described in English, which are translated into German, French and Czech.[1] Task 1b is similar to Task 1; the main difference is that multiple source languages can be used (simultaneously) and Czech is the only target language.

We introduce two new evaluation sets that extend the existing Multi30K dataset: a set of 1071 English sentences and their corresponding images and translations for Task 1, and 1,000 translations for the 2017 test set into Czech for Task 1b.

Another new feature of this year's shared task is the introduction of a new evaluation metric: Lexical Translation Accuracy (LTA), which measures

[1]The current version of the dataset can be found here: https://github.com/multi30k/dataset

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 308–327

https://doi.org/10.18653/v1/W18-6402

the accuracy of a system at translating correctly a subset of ambiguous source language words.

Participants could submit both constrained (shared task data only) and unconstrained (any data) systems for both tasks, with a limit of two systems per task variant and language pair per team.

2 Datasets

The Multi30K dataset (Elliott et al., 2016) is the primary resource 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.

Over the two last years, we have extended the Multi30K dataset with 2,071 new images and two additional languages for the translation task: French and Czech. Table 1 presents an overview of the new evaluation datasets. Figure 1 shows an example of an image with an aligned English-German-French-Czech description.

This year we also released a new version of the evaluation datasets featuring a subset of sentences that contain ambiguous source language words, which may have different senses in the target language. We expect that these ambiguous words could benefit from additional visual context.

In addition to releasing the parallel text, we also distributed two types of visual features extracted from a pre-trained ResNet-50 object recognition model (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.

Multi30K Czech Translations

This year the Multi30K dataset was extended with translations of the image descriptions into Czech. The translations were produced by 15 workers (university and high school students and teachers, all with a good command of English) at the cost of EUR 3,500. The translators used the same platform that was used to collect the French translations for the Multi30K dataset. The Czech translators had access to the source segment in English and the image only (no automatic translation into Czech was presented). The translated segments were automatically checked for mismatching punctuation, spelling errors (using `aspell`), inadequately short and long sentences, and non-standard characters. The segments containing errors were manually checked and fixed if needed. In total, 5,255 translated segments (16%) were corrected. After the manual correction, 1% of the segments were sampled and manually annotated for translation quality. This annotation task was performed by three annotators (and every segment was annotated by two different people to measure annotation agreement). We found that 94% of the segments did not contain any spelling errors, 96% of the segments fully preserved the meaning, and 75% of translations were annotated as fluent Czech. The remaining 25% contained some stylistic problems (usually inappropriate lexical choice and/or word order adopted from the English source segment). However, the annotation agreement for stylistic problems was substantially lower compared to other categories due to the subjectivity of deciding on the best style for a translation.

En: A boy dives into a pool near a water slide.
De: Ein Junge taucht in der Nähe einer Wasserrutsche in ein Schwimmbecken.
Fr: Un garçon plonge dans une piscine près d'un toboggan.
Cs: Chlapec skáče do bazénu poblíž skluzavky.

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

Test 2018 dataset

As our new evaluation data for Task 1, we collected German, French and Czech translations for the test set used in the 2017 edition of the Multilingual Image Description Generation task, which only contained English descriptions. This test set contains images from five of the six Flickr groups used to create the original Flickr30K dataset[2]. We

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

	Training set	Development set	Test set 2018 - Task 1	Test set 2018 - Task 1b
Instances	29,000	1,014	1071	1,000

Table 1: Overview of the Multi30K training, development and 2018 test datasets. The figures correspond to tuples with an image and parallel sentences in four languages: English, German, French and Czech.

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

Table 2: Distribution of images in the Test 2018 dataset by Flickr group.

sampled additional images from two thematically related groups (Everything Outdoor and Flickr Social Club) because Outdoor Activities only returned 10 new CC-licensed images and Flickr-Social no longer exists. The translations were collected using the same procedure as before for each of the languages: professional translations for German and internally crowdsourced translations for French and Czech (see (Elliott et al., 2017)), as described above. The new evaluation data for Task 1b consists of Czech translations, which we collected following the procedure described above. 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, 1,000 were used for Task 1 and 1,071 for Task 1b.

Dataset for LTA

In this year's task we also evaluate systems using Lexical Translation Accuracy (LTA) (Lala and Specia, 2018). LTA measures how accurately a system translates a subset of ambiguous words found in the Multi30K corpus. To measure this accuracy, we extract a subset of triplets form the Multi30K dataset in the form (i, aw, clt) where i is the index representing an instance in the test set, aw is an ambiguous word in English found in that instance i, and clt is the set of correct lexical translations of aw in the target language that conform to the context i. A word is said to be ambiguous in the source language if it has multiple translations (as given in the Multi30K corpus) with different meanings.

We prepared the evaluation dataset following the procedure described in Lala and Specia (2018), with some additional steps. First, the parallel text in the Multi30K training and the validation sets are decompounded with SECOS (Riedl and Biemann, 2016) (for German only) and lemmatised[3]. Second, we perform automatic word alignment using fast_align (Dyer et al., 2013) to identify the English words that are aligned to two or more different words in the target language. This step results in a dictionary of $\{key : val\}$ pairs, where key is a potentially ambiguous English word, and val is the set of words in the target language that align to key. This dictionary is then filtered by humans, students of translation studies who are fluent in both the source and target languages, to remove incorrect/noisy alignments and unambiguous instances, resulting in a cleaned dictionary containing $\{aw : lt\}$ pairs, where aw is an ambiguous English word, and lt is the set of lexical translations of aw in the corpus. For English-Czech, we were unable to perform this 'human filtering' step, and so we use the unfiltered, noisy dictionary. Table 3 shows summary statistics about number of ambiguous words and the total number of their instances in the training and validation sets.

Given a dictionary, we identify instances i in the test sets[4] which contain an ambiguous word aw from the dictionary, resulting in triplets of the form (i, aw, lt). At this stage we again involve human

[3]For English, German and French, we use the tool from `http://staffwww.dcs.shef.ac.uk/people/A.Aker/activityNLPProjects.html`. For Czech, we pre-processed the data using MorphoDiTa (Straková et al., 2014) from `http://ufal.mff.cuni.cz/morphodita`

[4]The test data and the submissions undergo the same pre-processing steps as the training and the validation sets.

Language Pair	Ambiguous Words	Instances
EN-DE	745	53,868
EN-FR	661	44,779
EN-CS	3217	187,495

Table 3: Statistics of the ambiguous words extracted from the training and validation sets after human filtering (dictionary filtering). For EN-CS, the numbers are larger because we could not perform the dictionary filtering step.

annotators (students of translation studies) to select, from the set of lexical translations lt, only those translations, denoted as clt, which conform to the source context i - both image and its English description. For example, in the test instance shown in Figure 2, *hat* is an ambiguous word aw and {*kappe, mütze, hüten, kopf, kopfbedeckung, kopfbedeckungen, hut, helm, hüte, helmen, mützen*} is the set of its lexical translations lt. The human annotator looked at both the image and its description and then selected the following subset {*kappe, mütze, mützen*} as the correct lexical translations clt that conform to the context of the test instance in Figure 2. We also asked annotators to expand the *clt* set with other synonyms outside the *lt* set that satisfy the context if they can. The number of ambiguous words and instances for each language pair in the resulting dataset for the test instances is given in Table 4. For English-Czech, while the first human filtering step (dictionary filtering) was not performed, the second human filtering step (test set filtering) was done. We note that this cleaning done by the Czech-English annotators was very selective, most likely due to the noisier nature of the initial annotations from the unfiltered dictionary.

Given a human filtered dictionary, the LTA evaluation is straight forward: for each MT system submission, we check if any word in clt is found in the translation of the submission's i^{th} instance.The preprocessing steps may result in mismatches due to sub-optimal handling of morphological variants, but we do not expect this to be a rare event because the dictionaries, gold standard text, and system submissions are pre-processed using the same tools.

3 Participants

This year we attracted submissions from seven groups. Table 5 presents an overview of the groups

En: a cute boy with his **hat** looking out of a window.
De: ein süß jung mit mütze blicken aus einem fenster.
aw: **hat**
lt: {kappe, mütze, hüten, kopf, kopfbedeckung, kopfbedeckungen, hut, helm, hüte, helmen, mützen}
clt: {kappe, mütze, mützen}

Figure 2: A test instance with ambiguous word *aw* and lexical translation options *lt*. Human annotator corrects/selects those options *clt* which conform to the source sentence *En* and corresponding image.

Language Pair	Ambiguous Words	Test instances
EN-DE	38	358
EN-FR	70	438
EN-CS	29	140
EN-CS(1B)	28	52

Table 4: Statistics of dataset used for the LTA evaluation after human filtering.

and their submission identifiers.

AFRL-OHIO-STATE (Task 1)

The AFRL-OHIO-STATE team builds on their previous year Visual Machine Translation (VMT) submission by combining it with text-only translation models. Two types of models were submitted: AFRL-OHIO-STATE_1_2IMPROVE_U is a system combination of the VMT system and an instantiation of a Marian NMT model (Junczys-Dowmunt et al., 2018), and AFRL-OHIO-STATE_1_4COMBO_U is a systems combination of the VMT system along with instantiations of Marian, OpenNMT, and Moses (Koehn et al., 2007).

CUNI (Task 1)

The CUNI submissions use two architectures based on the self-attentive Transformer model (Vaswani et al., 2017). For German and Czech, a language model is used to extract pseudo-in-

ID	Participating team
AFRL-OHIOSTATE	Air Force Research Laboratory & Ohio State University (Gwinnup et al., 2018)
CUNI	Univerzita Karlova v Praze (Helcl et al., 2018)
LIUMCVC	Laboratoire d'Informatique de l'Université du Maine & Universitat Autonoma de Barcelona Computer Vision Center (Caglayan et al., 2018)
MeMAD	Aalto University, Helsinki University & EURECOM (Grönroos et al., 2018)
OSU-BAIDU	Oregon State University & Baidu Research (Zheng et al., 2018)
SHEF	University of Sheffield (Lala et al., 2018)
UMONS	Université de Mons (Delbrouck and Dupont, 2018)

Table 5: Participants in the WMT18 multimodal machine translation shared task.

domain data from all available parallel corpora and mix it with the original Multi30k data and the EU Bookshop corpus. At inference time, both submitted models use only the text input. The first model was trained using the parallel data only. The second model is a reimplementation of the Imagination model (Elliott and Kádár, 2017) adapted to the Transformer architecture. During training, the model uses the encoder states to predict the image representation. This allows using additional English-only captions from the MSCOCO dataset (Lin et al., 2014).

LIUMCVC (Task 1)

LIUMCVC proposes a refined version of their multimodal attention model (Caglayan et al., 2016), where source-side information from the textual encoder (i.e. last hidden state of the bidirectional gated recurrent units (GRU)) is now used to filter the convolutional feature maps before the actual decoder-side multimodal attention is computed. The authors also experiment with the impact of L_2 normalisation and input image size for convolutional feature extraction process and found that multimodal attention without L_2 normalisation performs significantly worse than baseline NMT.

MeMAD (Task 1)

The MeMAD team adapts the Transformer neural machine translation architecture to a multimodal setting. They use global image features extracted from Detectron (Girshick et al., 2018), a pre-trained object detection and localisation neural network, and two additional training corpora: MS-COCO (Lin et al., 2014) (an English multimodal dataset, which they extend with synthetic multilingual data) and OpenSubtitles (Lison and Tiedemann, 2016) (a multilingual, text-only dataset). Their experiments show that the effect of the visual features in the system is small; the largest differences in quality amongst the systems tested is attributed to the quality of the underlying text-only neural MT system.

OSU-BAIDU (Tasks 1 and 1b)

For Task 1, the OREGONSTATE system ensembles models including some neural machine translation models which only consider text information and multimodal machine translation models which also consider image information. Both types of models use global attention mechanism to align source to target words. For the multimodal model, 1024 dimensional vectors are extracted as image information from a ResNet-101 convolutional neural network and these are used to initialize the decoder. The models are trained using scheduled sampling (Bengio et al., 2015) and reinforcement learning (Rennie et al., 2017) to further improve performance.

For Task 1b, for each language in the multisource inputs, single-source models are trained using the same architecture as in Task 1. The resulting models are ensembled with different combinations. The final submissions only ensemble models trained from English-to-Czech pair, which outperforms other combinations on the development set.

SHEF (Tasks 1 and 1b)

For Task 1, SHEF adopts a two-step pipeline approach. In the first (base) step – submitted as a baseline system – they use an ensemble of standard attentive text-only neural machine translation models built using the NMTPY toolkit (Caglayan et al., 2017) to produce 10-best high quality trans-

lation candidates. In the second (re-ranking) step, the 10-best candidates are re-ranked using word sense disambiguation (WSD) approaches: (i) most frequency sense (MFS), (ii) lexical translation (LT) and, (iii) multimodal lexical translation (MLT). Models (i) and (ii) are baselines, whilst MLT is a novel multimodal cross-lingual WSD model. The main idea is to have the cross-lingual WSD model select the translation candidate which correctly disambiguates ambiguous words in the source sentence and the intuition is that the image could help in the disambiguation process. The re-ranking cross-lingual WSD models are based on neural sequence learning models for WSD (Raganato et al., 2017; Yuan et al., 2016) trained on the Multimodal Lexical Translation Dataset (Lala and Specia, 2018). More specifically, they train LSTMs as taggers to disambiguate/translate every word in the source sentence.

For Task 1b, the SHEF team explores three approaches. The first approach takes the concatenation of the 10-best translation candidates of German-Czech, French-Czech and English-Czech neural MT systems and then re-ranks them using the same multimodal cross-lingual WSD model as in Task 1. The second approach explores consensus between the different 10-best lists. The best hypothesis is selected according to the number of times it appears in the different n-bests. The highest ranked hypothesis with the majority votes is selected. The third approach uses data augmentation: extra source (Czech) data is generated by building systems that translate from German into English and French into English. An English-Czech neural machine translation system is then built and the 10-best list is generated. For re-ranking, classifiers are trained to predict binary scores derived from Meteor for each hypothesis in the 10-best list using word embeddings and image features.

UMONS (Task 1)

The UMONS submission uses as baseline a conditional GRU decoder. The architecture is enhanced with another GRU that receives as input the global visual features provided by the task (i.e. 2048-dimensional ResNet pool5 features) as well as the hidden state of the second GRU. Each GRU disposes of 256 computational units. All non-linear transformations in the decoder (apart from the textual attention module) use gated hyperbolic tangent activations. Both visual and textual representation are separately projected onto a vocabulary-sized space. At every timestep, the decoder ends up with two modality-dependent probability distributions over the target tokens, eventually merged with an element-wise addition.

Baseline (Tasks 1 and 1b) The baseline system for both tasks is a text-only neural machine translation system built with the NMTPY (Caglayan et al., 2017) following a standard attentive approach (Bahdanau et al., 2015) with a conditional GRU decoder. The baseline was trained using the Adam optimizer, with a learning rate of $5e^{-5}$ and a batch size of 64. The input embedding dimensionality was set to 128 and the remainder of the hyperparameters were kept as default. Bite-pair encoding with 10,000 merge operations was used for all language pairs. For Task 1b, only the English-Czech portion of the training corpus is used.

4 Automatic 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.[5] 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 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

Table 6 shows the results on the Test 2018 dataset with a German target language. The first observation is that the best-performing system, MeMAD_1_FLICKR_DE_MeMAD-OpenNMT-mmod_U, is substantially better than other systems, although it uses unconstrained data. The MeMAD team did not submit a constrained or monomodal submission, so we cannot conclude whether this improvement comes from the use of multimodal data or from the additional parallel data. However, as mentioned in Section 3, the

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

authors themselves state that the gains mainly come from the additional parallel text data in the monomodal system. The vast majority of systems beat the strong text-only Baseline by a considerable margin. For other teams submitting monomodal and multimodal versions of their systems (e.g. CUNI and LIUMCVC), there does not seem to be a marked difference in automatic metric scores.

We can also observe that the ambiguous word evaluations (LTA) does not lead to the same system ranking as the automatic metrics. While this could stem mainly from the fact that the LTA evaluation is only performed on a small subset of the test cases, we consider that these two automatic evaluations are complementary. General translation quality is measured with the standard metrics (BLEU, METEOR and TER), while the LTA evaluations captures the ability of the system to model complex words which, in many cases, could require the use of the image input to disambiguate them.

4.2 Task 1: English → French

Table 7 shows the results for the Test 2018 dataset with French as target language. Once again, the MeMAD_1_FLICKR_FR_MeMAD-OpenNMT-mmod_U system performs significantly better than the other systems.[6] For teams submitting monomodal and multimodal versions of their systems (e.g. CUNI and LIUMCVC), there does not seem to be a marked difference in automatic metric scores. Another interesting observation is that in this case the clearly superior performance of the MeMAD_1_FLICKR_FR_MeMAD-OpenNMT-mmod_U system also shows in the LTA evaluation.

All submissions significantly outperformed the English→French baseline system. For this language pair, the evaluation metrics are in better agreement about the ranking of the submissions, however, the LTA metric is once again less correlated.

4.3 Task 1: English → Czech

The Czech language is a new addition to the 2018 evaluation campaign. Table 8 shows the results for the Test 2018 dataset with Czech as target language. A smaller number of teams have submitted systems for this language pair. This is a more complex language pair as demonstrated by the lower automatic scores obtained by the systems. The best results are obtained by the CUNI_1_FLICKR_CS_NeuralMonkeyImagination_U system, under the unconstrained conditions. The constrained systems all perform similarly to each other, and all except CUNI_1_FLICKR_CS_NeuralMonkeyTextual_U are significantly better than the baseline system. Interestingly, for the OSU-BD submissions, LTA seems to disagree significantly with the other metrics. More analysis is necessary to understand why this is the case.

[6]We note that their original submission had tokenisation issues, which were fixed by the task organisers.

4.4 Task 1b: Multisource English, German, French → Czech

Multisource multimodal translation is a new task this year. This task invites participants to use multiple source language inputs, as well as the image, in order to generate Czech translations. Only a few systems have been submitted compared to the other tasks. The results for the Test 2018 dataset are presented in Table 9. We observe that all teams outperformed the text-only baseline, even though in some cases the difference is not significant. No teams used unconstrained data in their submissions.

Again, the LTA results do not follow those of the automatic metrics, particularly for the two top submissions: LTA scores differ by a large margin, while all other metric scores are the same or very similar. This could however result from the very small number of samples available for LTA evaluation for this task: only 52 test instances. Differences in the translation of a very few number of instances can therefore result in considerably differences in LTA scores.

5 Human Judgment Results

In addition to the automatic metrics evaluation, we conducted human evaluation to assess the translation quality of the submissions. This evaluation was undertaken for the Task 1 German, French and Czech outputs as well as for the Task 1b Czech outputs for the Test 2018 dataset. This section describes how we collected the human assessments and computed the results. We are grateful to all of the assessors for their contributions.

5.1 Methodology

The system outputs indicated as the primary submission were manually evaluated by bilingual Direct Assessment (DA) (Graham et al., 2015) using

EN → DE	BLEU ↑	Meteor ↑	TER ↓	LTA ↑
•MeMAD_1_FLICKR_DE_MeMAD-OpenNMT-mmod_U (P)	38.5	56.6	44.5	47.49
CUNI_1_FLICKR_DE_NeuralMonkeyTextual_U	32.5	52.3	50.8	46.37
CUNI_1_FLICKR_DE_NeuralMonkeyImagination_U (P)	32.2	51.7	51.7	47.21
UMONS_1_FLICKR_DE_DeepGru_C (P)	31.1	51.6	53.4	48.04
LIUMCVC_1_FLICKR_DE_NMTEnsemble_C (P)	31.1	51.5	52.6	46.65
LIUMCVC_1_FLICKR_DE_MNMTEnsemble_C (P)	31.4	51.4	52.1	45.81
OSU-BD_1_FLICKR_DE_RLNMT_C (P)	32.3	50.9	49.9	45.25
OSU-BD_1_FLICKR_DE_RLMIX_C	32.0	50.7	49.6	46.09
SHEF_1_DE_LT_C	30.4	50.7	53.0	48.04
SHEF_1_DE_MLT_C (P)	30.4	50.7	53.0	48.32
SHEF1_1_DE_ENMT_C	30.8	50.7	52.4	44.41
SHEF1_1_DE_MFS_C (P)	30.3	50.7	53.1	48.32
LIUMCVC_1_FLICKR_DE_MNMTSingle_C	28.8	49.9	55.6	45.25
LIUMCVC_1_FLICKR_DE_NMTSingle_C	29.5	49.9	54.3	47.77
Baseline	27.6	47.4	55.2	45.25
AFRL-OHIO-STATE_1_FLICKR_DE_4COMBO_U (P)	24.3	45.4	58.6	46.09
AFRL-OHIO-STATE_1_FLICKR_DE_2IMPROVE_U	10.0	25.4	79.0	25.42

Table 6: Official automatic results for the MMT18 Task 1 on the English → German Test 2018 dataset (ordered by Meteor). Grey background indicate use of resources that fall outside the constraints provided for the shared task. (P) indicate a primary system designated for human evaluation.

EN → FR	BLEU ↑	Meteor ↑	TER ↓	LTA ↑
•MeMAD_1_FLICKR_FR_MeMAD-OpenNMT-mmod_U (P)	44.1	64.3	36.9	73.08
CUNI_1_FLICKR_FR_NeuralMonkeyTextual_U	40.6	61.0	40.7	68.44
CUNI_1_FLICKR_FR_NeuralMonkeyImagination_U (P)	40.4	60.7	40.7	69.29
UMONS_1_FLICKR_FR_DeepGru_C (P)	39.2	60.0	41.8	68.82
LIUMCVC_1_FLICKR_FR_MNMTEnsemble_C (P)	39.5	59.9	41.7	68.53
LIUMCVC_1_FLICKR_FR_NMTEnsemble_C (P)	39.1	59.8	41.9	68.44
SHEF_1_FR_LT_C	38.8	59.8	41.5	69.57
SHEF_1_FR_MLT_C (P)	38.9	59.8	41.5	69.86
SHEF1_1_FR_ENMT_C	38.9	59.8	41.2	67.87
SHEF1_1_FR_MFS_C (P)	38.8	59.7	41.6	67.58
OSU-BD_1_FLICKR_FR_RLNMT_C (P)	39.0	59.5	41.2	68.91
OSU-BD_1_FLICKR_FR_RLMIX_C	38.6	59.3	41.5	67.68
LIUMCVC_1_FLICKR_FR_MNMTSingle_C	37.9	58.5	43.4	67.77
LIUMCVC_1_FLICKR_FR_NMTSingle_C	37.6	58.4	43.2	67.11
Baseline	36.3	56.9	54.3	66.26

Table 7: Official automatic results for the MMT18 Task 1 on the English → French Test 2018 dataset (ordered by Meteor). Grey background indicate use of resources that fall outside the constraints provided for the shared task. (P) indicate a primary system designated for human evaluation.

EN → CS	BLEU ↑	Meteor ↑	TER ↓	LTA ↑
•CUNI_1_FLICKR_CS_NeuralMonkeyImagination_U (P)	31.8	30.6	48.2	70.00
OSU-BD_1_FLICKR_CS_RLMIX_C	30.1	29.7	51.2	54.29
OSU-BD_1_FLICKR_CS_RLNMT_C (P)	30.2	29.5	50.7	60.71
SHEF1_1_CS_ENMT_C	29.0	29.4	51.1	71.43
SHEF1_1_CS_MFS_C (P)	27.8	29.2	52.4	73.57
SHEF_1_CS_LT_C	28.3	29.1	51.7	72.14
SHEF_1_CS_MLT_C (P)	28.2	29.1	51.7	71.43
Baseline	26.5	27.7	54.4	62.14
*CUNI_1_FLICKR_CS_NeuralMonkeyTextual_U	26.8	27.1	55.2	52.14

Table 8: Official automatic results for the MMT18 Task 1 on the English → Czech Test 2018 dataset (ordered by Meteor). Grey background indicate use of resources that fall outside the constraints provided for the shared task. (P) indicate a primary system designated for human evaluation. Submissions marked with * are not significantly different from the Baseline.

EN,DE,FR → CS	BLEU ↑	Meteor ↑	TER ↓	LTA ↑
OSU-BD_1b_CS_RLMIX_C	26.4	28.2	52.7	55.77
OSU-BD_1b_CS_RLNMT_C (P)	26.4	28.0	52.1	61.54
SHEF_1b_CS_CON_C	24.7	27.6	52.1	61.54
*SHEF_1b_CS_MLTC_C (P)	24.5	27.5	52.5	61.54
SHEF1_1b_CS_ARNN_C (P)	25.2	27.5	53.9	51.92
*SHEF1_1b_CS_ARF_C	24.1	27.1	54.6	51.92
Baseline	23.6	26.8	54.1	53.85

Table 9: Official automatic results for the MMT18 Task 1b on the English,German,French → Czech Test 2018 dataset (ordered by Meteor). Submissions marked with * are not significantly different from the Baseline.

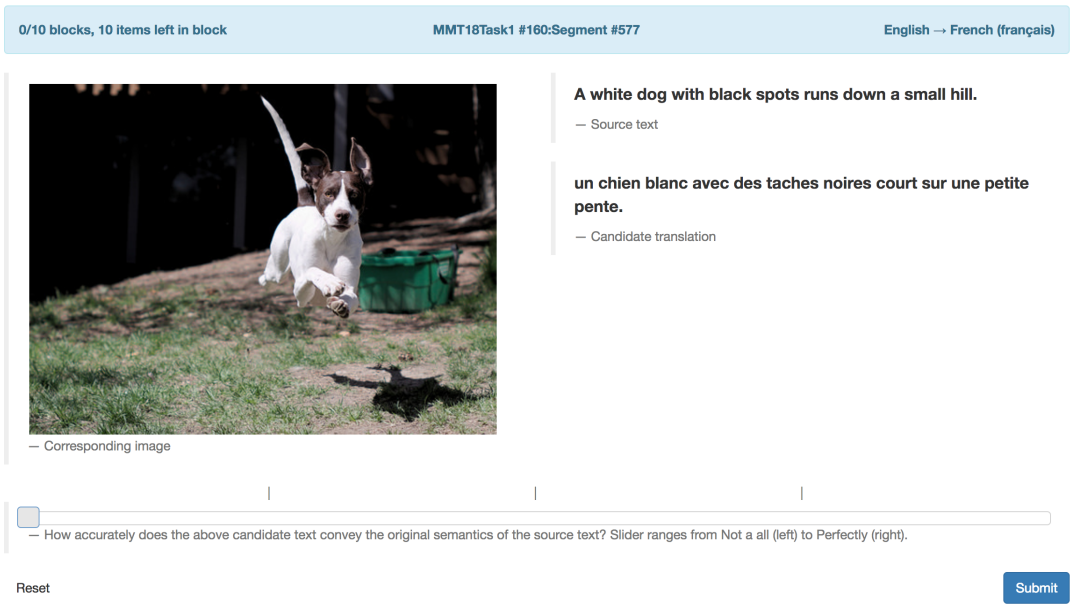

Figure 3: Example of the human direct assessment evaluation interface.

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, French or Czech. For the Multisource Task (1b), only the English source is presented. For the evaluation task, the image was shown along with the source sentence and the candidate translation. Evaluators were ask 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 to avoid influencing the assessment. Instead, as a control experiment to estimate the quality of the reference sentences (and test the quality of the annotations), we included the references as hypotheses for human evaluation. Figure 3 shows an example of the direct assessment interface used in the evaluation. The score of each translation candidate ranges from 0 (the meaning of the source is not preserved in the target language sentence) to 100 (the meaning of the source is "perfectly" preserved). The overall score of a given system (z) corresponds to the mean standardised score of its translations.

5.2 Results

For Task 1 English-German translation, we collected 3,422 DAs, resulting in a minimum of 300 and a maximum of 324 direct assessments per system submission, respectively. We collected 2,938 DAs for the English-French translations. This results in a minimum of 280 and a maximum of 307 direct assessments per system submission, respectively. We collected 8,096 DAs for the Task 1 English-Czech translation, representing a minimum of 1,330 and a maximum of 1,370 direct assessments per system submission. For Task 1b English,German,French→Czech translation, we collected 6,827 direct assessments. The least evaluated system received 1,345 assessments, while the most evaluated system received 1,386 direct assessments.

Tables 10, 11, 12 and 13 show the results of the human evaluation for the English to German, English to French and English to Czech Multimodal Translation task (Test 2018 dataset) as well as the Multisource Translation task. The systems are ordered by standardised mean DA scores and clustered according to the Wilcoxon signed-rank test at p-level $p \leq 0.05$. Systems within a cluster are considered tied. The supplementary Wilcoxon signed-rank scores can be found in Tables 14, 15 and 16 in Appendix A.

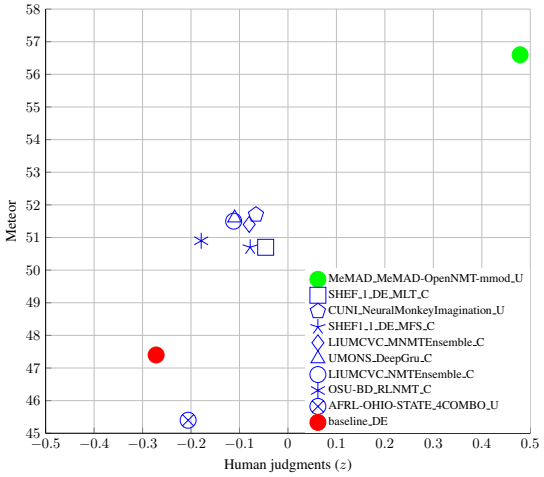

Figure 4: System performance on the English→German Test 2018 dataset as measured by human evaluation against Meteor scores.

The comparison between automatic and human evaluation are presented in Figures 4, 5, 6 and 7. We can observe that METEOR scores are well correlated with the human evaluation.

6 Discussion

As mentioned in Section 5, we included the reference sentences in the DA evaluation as if they were candidate translations generated by a system. The first observation is that for all language pairs and all tasks, the references (see gold_* in Tables 10, 11, 12 and 13) are significantly better than all automatic systems with average raw scores above 90%. This does not only validates the references but also the DA evaluation process.

For the first time in the MMT evaluation campaign series, using additional (unconstrained) data resulted in some significant improvement both in terms of automatic score and human evaluation. The biggest improvements come from the unconstrained MeMAD system (for the English-German and English-French), which achieves large improvements in Meteor score compared to the second best system. This is also the case in terms of human evaluation. For English-German, for example, the average raw DA score (87.2, see second column of Table 10) is only 4.5% away from the result of the reference evaluation (91.7). The MeMAD team use a transformer NMT architec-

English→German			
#	Ave %	Ave z	System
1	91.7	0.69	gold_DE_1
2	87.2	0.479	MeMAD_MeMAD-OpenNMT-mmod_U
3	73.5	-0.046	SHEF_1_DE_MLT_C
	73.8	-0.066	CUNI_NeuralMonkeyImagination_U
	72.6	-0.078	SHEF1_1_DE_MFS_C
	71.6	-0.08	LIUMCVC_MNMTEnsemble_C
	72.1	-0.11	UMONS_DeepGru_C
	72.5	-0.112	LIUMCVC_NMTEnsemble_C
	71.1	-0.179	OSU-BD_RLNMT_C
	68.6	-0.206	AFRL-OHIO-STATE_4COMBO_U
	67.4	-0.272	baseline_DE

Table 10: Results of the human evaluation of the WMT18 English-German Multimodal Translation task (Test 2018 dataset). 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 14). Systems using unconstrained data are identified with a gray background.

English→French			
#	Ave %	Ave z	System
1	90.3	0.487	gold_FR_1
2	86.8	0.349	MeMAD_MeMAD-OpenNMT-mmod_U
3	78.5	0.047	CUNI_NeuralMonkeyImagination_U
	77.3	-0.005	UMONS_DeepGru_C
	74.9	-0.05	LIUMCVC_NMTEnsemble_C
	74.9	-0.075	SHEF1_1_FR_MFS_C
	74.5	-0.088	SHEF_1_FR_MLT_C
	73.0	-0.11	LIUMCVC_MNMTEnsemble_C
	74.4	-0.12	OSU-BD_RLNMT_C
	66.0	-0.376	baseline_FR

Table 11: Results of the human evaluation of the WMT18 English-French Multimodal Translation task (Test 2018 dataset). 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 15). Systems using unconstrained data are identified with a gray background.

English→Czech			
#	Ave %	Ave z	System
1	93.2	0.866	gold_CS_1
2	70.2	0.097	CUNI_NeuralMonkeyImagination_U.txt
	62.4	-0.162	SHEF_1_CS_MLT_C
	60.6	-0.225	SHEF1_1_CS_MFS_C
	59.1	-0.248	OSU-BD_RLNMT_C
3	57.8	-0.337	baseline_CS

Table 12: Results of the human evaluation of the WMT18 English-Czech Multimodal Translation task (Test 2018 dataset). 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 16). Systems using unconstrained data are identified with a gray background.

English,French,German→Czech			
#	Ave %	Ave z	System
	93.6	0.803	gold_CS_1b
	63.3	-0.149	SHEF_1b_CS_MLTC_C
	61.8	-0.178	SHEF1_1b_CS_ARNN_C
	62.1	-0.206	OSU-BD_1b_CS_RLNMT_C
	59.4	-0.284	baseline_CS_task1b

Table 13: Results of the human evaluation of the WMT18 English,French,German-Czech Multisource Multimodal Translation task (Test 2018 dataset). 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 17).

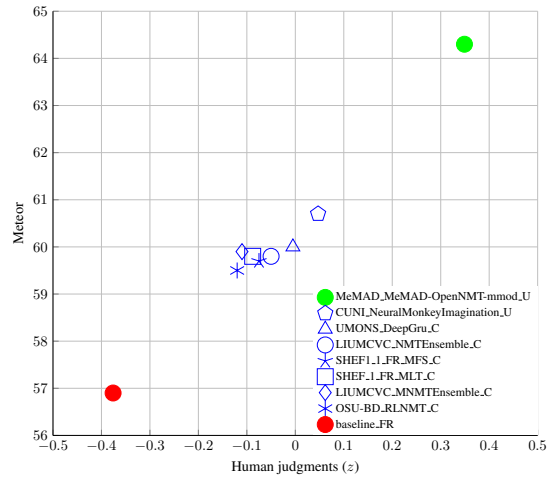

Figure 5: System performance on the English→French Test 2018 dataset as measured by human evaluation against Meteor scores.

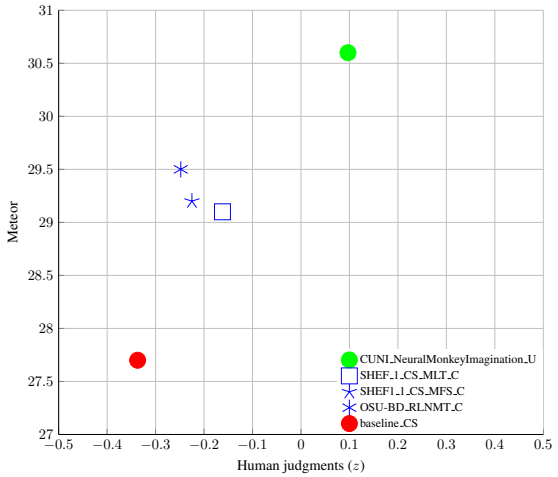

Figure 6: System performance on the English→Czech Test 2018 dataset as measured by human evaluation against Meteor scores.

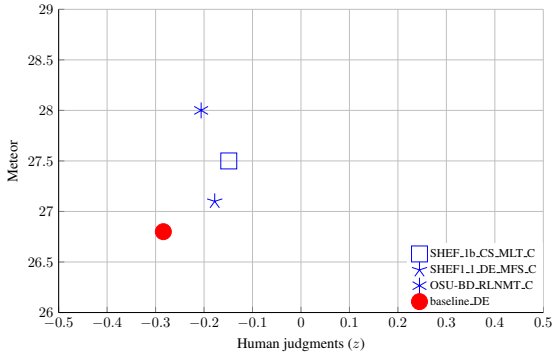

Figure 7: System performance on the English,German,French→Czech Test 2018 dataset as measured by human evaluation against Meteor scores.

ture (as opposed to recurrent neural networks) combined with global image feature that are different from the ResNet features made available by the task organisers. However, according to the authors it seems that most of the improvements come from the additional parallel data.

Many teams proposed a combination of several systems. This is the case for AFRL-OHIO-STATE, LIUMCVC, OSU-BAIDU and SHEF teams. LIUMCVC also submitted a non-ensembled version of each system. Their conclusion is that ensembling multiple systems benefit monomodal and multimodal systems.

Lexical Translation Accuracy LTA was a new evaluation for this campaign. Unlike other automatic metrics, LTA only evaluates a specific aspect of translation quality, namely lexical disambiguation. One of the motivations for multimodality in machine translation is that the visual features could help to disambiguate ambiguous words (Elliott et al., 2015; Hitschler et al., 2016). Our aims in introducing the LTA metric was to directly evaluate the disambiguation performance of participating systems.

The LTA columns in Tables 6, 7, 8, and 9 show some interesting trends. First, for teams submitting text-only and multimodal variants of models, the multimodal versions seem to perform better at LTA compared to their text-only counterparts (e.g. CUNI's systems). This trend is not visible using the Meteor, BLEU, or TER metrics. Second, the SHEF systems that were built precisely to perform cross-lingual LTA-style WSD perform well on this metric but they are not always the best-performing system on this metric.

Multisource multimodal translation Only two teams participated in this task. The automatic results are presented in Table 9, the human evaluation results are presented in Table 13 and the comparison between automatic and human evaluation results are shown in Figure 6. Although many direct assessments have been collected for this task, it was not possible to separate the systems into different clusters. We can see that there is still a large margin between the performance of the systems and the human gold reference, but this was also the case for the English-Czech language pair in Task 1.

7 Conclusions

We presented the results of the third shared task on multimodal translation. The shared task attracted submissions from seven groups, who submitted a total of 45 systems across the two proposed tasks. The Multimodal Translation task attracted the majority of the submissions, with fewer groups attempting multisource multimodal translation.

The main findings of the shared task are:

(i) Additional data can greatly improve the results as demonstrated by the winning unconstrained systems.

(ii) Almost all systems achieved better results compared to the baseline text-only translation system. Various text and visual integration schemes have been proposed, leading to only slight changes in the automatic and human evaluation results.

(iii) Automatic metrics and human evaluation provided similar results. However, it is difficult to evaluate the impact of the multimodality. In the future, submission of monomodal equivalent of the systems will be encouraged in order to better emphasize the effect of using the visual inputs.

We are considering to change the data in favor of a more ambiguous task where all modalities should be used in order to generate the output. A possibility would be to re-use the list of ambiguous words extracted for LTA computation and select the image/sentence pairs containing one or more of those words.

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), by the MultiMT project (EU H2020 ERC Starting Grant No. 678017 – Lucia Specia and Chiraag Lala), and by an Amazon Research Award (Desmond Elliott). We thank the Charles University team for collecting and describing the Czech data. The Czech data collection was supported by the Czech Science Foundation, grant number P103/12/G084.

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*.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, pages 1171–1179. MIT Press.

Ozan Caglayan, Adrien Bardet, Fethi Bougares, Loïc Barrault, Kai Wang, Marc Masana, Luis Herranz, and Joost van de Weijer. 2018. LIUM-CVC submissions for WMT18 multimodal translation task. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, Brussels, Belgium. Association for Computational Linguistics.

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

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

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.

Jean-Benoit Delbrouck and Stéphane Dupont. 2018. Umons submission for wmt18 multimodal translation task. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

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

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2013)*, pages 644–649.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 215–233, Copenhagen, Denmark.

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.

Desmond Elliott and Ákos Kádár. 2017. Imagination improves Multimodal Translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 130–141, Taipei, Taiwan.

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

Ross Girshick, Ilija Radosavovic, Georgia Gkioxari, Piotr Dollár, and Kaiming He. 2018. Detectron. `https://github.com/facebookresearch/detectron`.

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.

Stig-Arne Grönroos, Benoit Huet, Mikko Kurimo, Jorma Laaksonen, Bernard Merialdo, Phu Pham, Mats Sjöberg, Umut Sulubacak, Jörg Tiedemann, Raphael Troncy, and Raúl Vázquez. 2018. The MeMAD submission to the WMT18 multimodal translation task. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

Jeremy Gwinnup, Joshua Sandvick, Michael Hutt, Grant Erdmann, John Duselis, and James Davis. 2018. The afrl-ohio state wmt18 multimodal system: Combining visual with traditional. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

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, Jindřich Libovický, and Dušan Variš. 2018. CUNI system for the WMT18 multimodal translation tasks. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

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.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

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.

Chiraag Lala, Pranava Madhyastha, Carolina Scarton, and Lucia Specia. 2018. Sheffield submissions for wmt18 multimodal translation shared task. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

Chiraag Lala and Lucia Specia. 2018. Multimodal Lexical Translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.

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.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*.

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.

Alessandro Raganato, Claudio Delli Bovi, and Roberto Navigli. 2017. Neural sequence learning models for word sense disambiguation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1156–1167.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1179–1195, Honolulu, Hawaii.

Martin Riedl and Chris Biemann. 2016. Unsupervised compound splitting with distributional semantics rivals supervised methods. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 617–622.

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.

Jana Straková, Milan Straka, and Jan Hajič. 2014. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18, Baltimore, Maryland. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

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.

Dayu Yuan, Julian Richardson, Ryan Doherty, Colin Evans, and Eric Altendorf. 2016. Semi-supervised word sense disambiguation with neural models. *arXiv preprint arXiv:1603.07012*.

Renjie Zheng, Yilin Yang, Mingbo Ma, and Liang Huang. 2018. Ensemble sequence level training for multimodal mt: Osu-baidu wmt18 multimodal machine translation system report. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

A Significance tests

Tables 14, 15 and 16 show the Wilcoxon signed-rank test used to create the clustering of the systems.

English→German											
	gold_DE_1	MeMAD_MeMAD-OpenNMT-mmod_U	SHEF_1_DE_MLT_C	CUNI_NeuralMonkeyImagination_U	SHEF1_1_DE_MFS_C	LIUMCVC_MNMTEnsemble_C	UMONS_DeepGru_C	LIUMCVC_NMTEnsemble_C	OSU-BD_RLNMT_C	AFRL-OHIO-STATE_4COMBO_U	baseline_DE
gold_DE_1	-	9.3e-10	1.2e-30	2.1e-28	2.0e-30	8.1e-28	1.8e-29	5.9e-31	2.5e-36	3.7e-37	4.2e-38
MeMAD_MeMAD-OpenNMT-mmod_U	-	-	4.4e-14	1.5e-12	2.7e-14	1.0e-12	7.0e-14	5.5e-15	1.2e-19	3.2e-21	3.5e-22
SHEF_1_DE_MLT_C	-	-	-	-	-	-	-	-	5.0e-02	1.2e-02	3.0e-03
CUNI_NeuralMonkeyImagination_U	-	-	-	-	-	-	-	-	4.8e-02	1.2e-02	3.1e-03
SHEF1_1_DE_MFS_C	-	-	-	-	-	-	-	-	-	2.9e-02	8.3e-03
LIUMCVC_MNMTEnsemble_C	-	-	-	-	-	-	-	-	-	3.2e-02	8.8e-03
UMONS_DeepGru_C	-	-	-	-	-	-	-	-	-	-	1.5e-02
LIUMCVC_NMTEnsemble_C	-	-	-	-	-	-	-	-	-	-	2.0e-02
OSU-BD_RLNMT_C	-	-	-	-	-	-	-	-	-	-	-
AFRL-OHIO-STATE_4COMBO_U	-	-	-	-	-	-	-	-	-	-	-
baseline_DE	-	-	-	-	-	-	-	-	-	-	-

Table 14: English → German Wilcoxon signed-rank test at p-level p $\leq$ 0.05. '-' means that the value is higher than 0.05.

English→French	gold_FR_1	MeMAD_MeMAD-OpenNMT-mmod_U	CUNI_NeuralMonkeyImagination_U	UMONS_DeepGru_C	LIUMCVC_NMTEnsemble_C	SHEF1_1_FR_MFS_C	SHEF_1_FR_MLT_C	LIUMCVC_MNMTEnsemble_C	OSU-BD_RLNMT_C	baseline_FR
gold_FR_1	-	2.7e-02	1.3e-09	1.8e-10	1.5e-10	5.1e-12	6.2e-13	4.4e-11	6.6e-14	3.3e-20
MeMAD_MeMAD-OpenNMT-mmod_U	-	-	3.0e-05	6.6e-06	3.0e-06	3.3e-07	8.1e-08	9.0e-07	1.4e-08	3.9e-14
CUNI_NeuralMonkeyImagination_U	-	-	-	-	-	-	-	-	4.9e-02	5.1e-05
UMONS_DeepGru_C	-	-	-	-	-	-	-	-	-	2.1e-04
LIUMCVC_NMTEnsemble_C	-	-	-	-	-	-	-	-	-	6.6e-04
SHEF1_1_FR_MFS_C	-	-	-	-	-	-	-	-	-	1.9e-03
SHEF_1_FR_MLT_C	-	-	-	-	-	-	-	-	-	3.5e-03
LIUMCVC_MNMTEnsemble_C	-	-	-	-	-	-	-	-	-	3.0e-03
OSU-BD_RLNMT_C	-	-	-	-	-	-	-	-	-	1.0e-02
baseline_FR	-	-	-	-	-	-	-	-	-	-

Table 15: English → French Wilcoxon signed-rank test at p-level $p \leq 0.05$. '-' means that the value is higher than 0.05.

English→Czech	gold_CS_1	CUNI_NeuralMonkeyImagination_U	SHEF_1_CS_MLT_C	SHEF1_1_CS_MFS_C	OSU-BD_RLNMT_C	baseline_CS_encs
gold_CS_1	-	6.9e-100	5.9e-150	3.6e-166	8.3e-158	1.3e-170
CUNI_NeuralMonkeyImagination_U	-	-	1.5e-10	1.4e-15	2.1e-16	1.5e-22
SHEF_1_CS_MLT_C	-	-	-	-	2.2e-02	6.7e-05
SHEF1_1_CS_MFS_C	-	-	-	-	-	8.2e-03
OSU-BD_RLNMT_C	-	-	-	-	-	2.8e-02
baseline_CS_encs	-	-	-	-	-	-

Table 16: English → Czech Wilcoxon signed-rank test at p-level $p \leq 0.05$. '-' means that the value is higher than 0.05.

English,French,German→Czech					
	gold_CS_1b	SHEF_1b_CS_MLTC_C	SHEF1_1b_CS_ARNN_C	OSU-BD_1b_CS_RLNMT_C	baseline_CS_.encs.task1b
gold_CS_task1b	-	4.4e-127	1.3e-115	3.8e-116	4.1e-132
SHEF_1b_CS_MLTC_C	-	-	-	-	4.3e-03
SHEF1_1b_CS_ARNN_C	-	-	-	-	1.2e-02
OSU-BD_1b_CS_RLNMT_C	-	-	-	-	-
baseline_CS_task1b	-	-	-	-	-

Table 17: English,French,German → Czech Wilcoxon signed-rank test at p-level $p \leq 0.05$. '-' means that the value is higher than 0.05.