
Article:
Frank, S., Elliott, D. and Specia, L. orcid.org/0000-0002-5495-3128 (2018) Assessing multilingual multimodal image description: Studies of native speaker preferences and translator choices. Natural Language Engineering, 24 (3). pp. 393-413. ISSN 1351-3249

https://doi.org/10.1017/S1351324918000074

Natural Language Engineering **1** (1): 1–19. Printed in the United Kingdom

Assessing multilingual multimodal image description: Studies of native speaker preferences and translator choices

Stella Frank
Centre for Language Evolution
University of Edinburgh, UK

Desmond Elliott
Institute of Language, Cognition and Computation
University of Edinburgh, UK

Lucia Specia
Department of Computer Science
University of Sheffield, UK

(*Received Deadline: November 30th*)

Abstract

Two studies on multilingual multimodal image description provide empirical evidence towards two hypotheses at the core of the task: (i) whether target language speakers prefer descriptions generated directly in their native language, as compared to descriptions translated from a different language; (ii) the role of the image in human translation of descriptions. These results provide guidance for future work in multimodal natural language processing by firstly showing that on the whole, translations are not distinguished from native language descriptions, and secondly delineating and quantifying the information gained from the image during the human translation task.

1 Introduction

Multimodal natural language processing (NLP) combines linguistic and non-linguistic modalities with the goal of grounding language in non-linguistic context, such as the visual context provided by an image. Modelling language in a grounded environment is important because it reflects how humans acquire, understand, and use language, namely, contextualised within a multimodal environment. Multimodal NLP research covers a broad range of topics, including image–sentence retrieval based on learning shared multimodal spaces (Hodosh et al., 2013), natural language generation from images and video (Bernardi et al., 2016), question answering given multimodal visual context (Antol et al., 2015), modelling the linguistic

attributes of images (Silberer and Lapata, 2014), and grounding the meaning of words in visual context (Lazaridou et al., 2015). Multimodality can also improve the performance of models for more traditional NLP problems, such as prepositional phrase attachment (Berzak et al., 2015) and co-reference resolution (Ramanathan et al., 2014).

The prototypical multimodal NLP task is image description[1] generation, which will be the focus of this paper. However, we are interested in image description from a *multilingual* perspective, specifically in a translation or transfer setting. This is an example of *multilingual multimodal* NLP, which broadly covers everything that involves images or other multimodal resources linked to text in multiple languages.

We assume a situation in which there is a resource-rich 'source' language (English) along with a 'target' language with fewer resources, but a need for image descriptions. There are (at least) two possible approaches to solving this need: Firstly, we could collect new multimodal data in the target language to train a monolingual target language image description system; secondly, we could translate English descriptions (either existing or machine generated) into the target language. For the first approach, we may want to also use the available English multimodal data, making the multimodal system *multilingual*, or rather crosslingual; for the second approach, we may want to take the image on which the description is based into account while collecting the translation, making the translation process *multimodal*.

These two approaches lead to different types of generated descriptions, which will serve different purposes. For example, when generating alt-text for stock photos online, it will be more important to generate descriptions that are appropriate for the user and context, without closely following the original language descriptions; this reflects the crosslingual multimodal scenario, in which the source language plays only a supporting role. On the other hand, when translating a manual with illustrations, staying faithful to the original text is crucial. In this case, the image can provide essential disambiguating information to the translator, leading to better translations. Prior work in multimodal NLP has shown the benefit of including multimodal inputs in a variety of visually-centered linguistic domains, such as user-generated captions on social media sites (Ordonez et al., 2011), product descriptions on e-commerce sites, and captioned images from newswire (Ramisa et al., 2017; Hollink et al., 2016) and historical newspaper corpora (Elliott and Kleppe, 2016).

In the crosslingual scenario, a more flexible relationship between texts in different languages allows for language or culture specific discrepancies between the texts. Different cultures may interpret the same image differently, which will be reflected in how they describe the image. Automatically generated descriptions will need to accommodate these differences where they are important for understand-

[1] We use the term image *description* in contrast to *captions* deliberately: we define descriptions as sentences that are solely and literally about an image, whereas captions are sentences associated with, but not necessarily descriptions of, an image. Descriptions datasets are usually gathered intentionally (as with the dataset used in this paper), e.g. using crowdsourcing, whereas captions are harvested from naturally appearing sources. Contrast the descriptions in Figure 1 with captions seen in newspapers or on social media, which usually include background information not depicted in the image.

ing. Differences often arise from shared cultural knowledge that may be unknown or less salient in a different language. For example, "tailgating" (elaborate picnicking around the back of a vehicle, usually associated with a sports event) is a popular activity in the U.S.A. that is obscure to German and Dutch speakers (van Miltenburg et al., 2017). A description of an image depicting tailgating thus needs to be phrased differently depending on the audience, which varies with language.

The two scenarios outlined above have been codified as a multimodal translation task and a description generation task, respectively, as part of the Multimodal Machine Translation shared task held at the Conference for Machine Translation in 2016 and 2017 (Specia et al., 2016; Elliott et al., 2017). The two tasks use different training data: the multimodal translation task is based on images with parallel translations of descriptions, while in the crosslingual image description task, the training data consists of images with independently authored descriptions in multiple languages. Evaluation also proceeds differently, since multimodal translation is evaluated as translation, based on the faithfulness of the target language description to the source language description, while crosslingual image description is evaluated based on the similarity of the generated description to multiple reference target-language descriptions, collected independently.

In this paper we re-evaluate and test the assumptions, outlined above, behind the multimodal translation and description generation tasks. First, we assess whether the division into two separate tasks, one based on flexible description generation in the crosslingual scenario and one focussed on literal translations, is actually necessary for the image description setting: do they result in measurably different descriptions? In particular, do target language speakers prefer descriptions created in their own language over translations from a different language? Note that the human-generated target language descriptions constitute an upper bound, in terms of quality: automatically generated descriptions based on source- and target-language training data, in the form of either translations or independent descriptions, are expected to perform less well. If, for example, German speakers do not differentiate between German descriptions and translations into German, this has important implications for multilingual multimodal NLP in the crosslingual, non-translation, setting.

Second, we examine the role of visual information in multimodal translation. Again, we take an approach based on human performance, but here we study how human translators use images during translation. Professional translators first translate image descriptions without seeing the image, then do post-editing to transform the 'image-blind' translation into an 'image-aware' translation. This enables us to quantify the difference that the image makes to translation, as well as to develop a classification of frequent error types arising in text-only, image-blind translation.

The remainder of this paper is structured as follows: we first outline the current state of multilingual multimodal NLP and describe available datasets and evaluation, in Section 2, with a focus on the above-mentioned shared task on multimodal translation. In Section 3, we present the human evaluation study comparing descriptions and translations. The multimodal translation study, comparing text-only and image-aware translation, is in Section 4. We conclude with a discussion of the

(a) Multilinguality by translation.

(b) Multilinguality by description.

Fig. 1: Multilingual annotations resulting from (a) a deliberate translation process from an "original" language into a new language, or (b) independently collecting annotations for the image in a new language.

implications of our findings and recommendations for future work on multimodal NLP, with an emphasis on resource design and evaluation methods.

2 Background

The availability of resources plays a critical role in the development and evaluation of computational models for multilingual multimodal image description. However, resources that are both multilingual and multimodal do not occurr naturally, unlike (unimodal) parallel texts, which can be found in parliamentary proceedings, or (monolingual) newswire captioned images. We review existing multilingual multimodal resources collected through crowdsourcing and professional translation. We also discuss evaluation methods for state-of-the-art multimodal translation models, which have to date mainly involved automatic metrics.

2.1 Multilingual Multimodal Resources

We define a multimodal resource as a collection of multimedia artefacts paired with textual annotations. Multimedia artefacts include photographs, videos, diagrams, line sketches, sound recordings, and video games, *inter alia*, while the textual annotations can range from single words, e.g. tags or keywords, to sentences, paragraphs, or entire documents. Given these definitions, examples of multimodal resources include datasets of tagged images, e.g. the COREL 5K dataset (Duygulu et al., 2002); images paired with crowdsourced descriptions, e.g. the Flickr30K dataset (Young

	Images	Sentences	Languages
Translation datasets			
Multi30K	31,014	31,014	English, German, French
Flickr8K-CN	8,018	40,090	English, Chinese
DECOCO	1,000	1,000	English, German
Multi30K-2017	1,000	1,000	English, German, French
AmbiguousCOCO	461	461	English, German, French
Description datasets			
STAIR-Captions	164,062	820,310	English, Japanese
Multi30K	31,014	155,070	English, German
YJ Captions 26k	26,500	131,740	English, Japanese
Flickr8K-CN	8,018	40,090	English, Chinese
Tasviret	8,018	24,054	English, Turkish
DutchDescription	2,014	10,070	English, Dutch

Table 1: Summary statistics of multilingual image description datasets.

et al., 2014); and videos paired with crowdsourced descriptions, e.g. the Microsoft Research Video Description corpus (Chen and Dolan, 2011). In this paper, we are primarily interested in *multilingual* multimodal resources, which are datasets that consist of multimedia artefacts with textual annotations in more than one language. The multilinguality of the textual annotations can take two forms: (i) it can arise as a process of translating annotations from one language into another language, or (ii) it can arise from creating textual annotations independently of those in the other language(s), given the multimedia artefact (see Figure 1). We will refer to these processes as *Translation* and *Description* throughout the rest of the paper, and we will study multilinguality that arises from both of these processes.

One of the earliest multilingual multimodal resources is the Microsoft Research Video Description corpus (Chen and Dolan, 2011), which consists of short YouTube videos with crowdsourced descriptions. The descriptions were not limited to English, and thus cover a broad range of languages. However, two-thirds of the descriptions are in English, and we are unaware of any work using the non-English descriptions.

More recently, there has been increased efforts to create multilingual image description datasets. These datasets consist of images paired with literal descriptions in multiple languages, created either by translation or independent description. Such resources currently exist with annotations in German (Elliott et al., 2016; Hitschler et al., 2016; Rajendran et al., 2016), Turkish (Unal et al., 2016), Chinese (Li et al., 2016), Japanese (Miyazaki and Shimizu, 2016; Yoshikawa et al., 2017), Dutch (van Miltenburg et al., 2017), and French (Elliott et al., 2017). Table 1 presents an overview of multilingual image description datasets. We observe that the datasets with multilingual annotations created by translation are an order of magnitude smaller than those created independently of each other. This is, in part, due to the expense of translation compared to crowdsourcing independent de-

	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 2: Corpus-level statistics of Multi30K dataset.

scriptions in each language. For example, the 31,014 translations in the Multi30K Translations data cost €23,000 to collect, whereas the 155,070 descriptions in the Descriptions data cost only $10,000 (Elliott et al., 2016).

The Multi30K dataset is the most commonly used multilingual image description dataset; it consists of images described in English, German, and French (Elliott et al., 2016; Elliott et al., 2017). This resource is derived from the Flickr30K dataset of images originally described in English (Young et al., 2014). The multilingual annotations exist in two forms: a translation corpus of parallel texts, and a corpus of independently collected descriptions.[2] For the translation corpus, one sentence (of five) was chosen for professional translation in a way that ensured that the final dataset was a combination of short, medium, and long sentences. The professional translations were created without the images, resulting in 'image-blind' translation data. We examine the consequences of this method of collecting multilingual annotations in Section 4. The second corpus consists of crowd-sourced descriptions gathered via Crowdflower[3] where each worker produced an independent description of the image. Table 2 presents an overview of the data available for each task.

An alternative approach to creating multilingual multimodal datasets is to create a parallel text using an off-the-shelf machine translation system. This approach also does not use the image to construct the input data. The Flickr8K-CN dataset contains translations of the English sentences into Mandarin using the Google[4] and Baidu[5] online translation systems (Li et al., 2016). There was no attempt to create human-quality Chinese translations of the English source data, e.g. by post-editing (possibly also with the image). The Flickr8K-CN dataset also contains crowdsourced descriptions created independently of the English originals.

2.2 Evaluating Multilingual Multimodal Models

Multimodal machine translation (MMT) has been the subject of two large-scale Shared Task evaluations at the Conference on Machine Translation (Specia et al.,

[2] The French data consists of translations only.
[3] `http://www.crowdflower.com`
[4] `http://translate.google.com`
[5] `http://translate.baidu.com/`

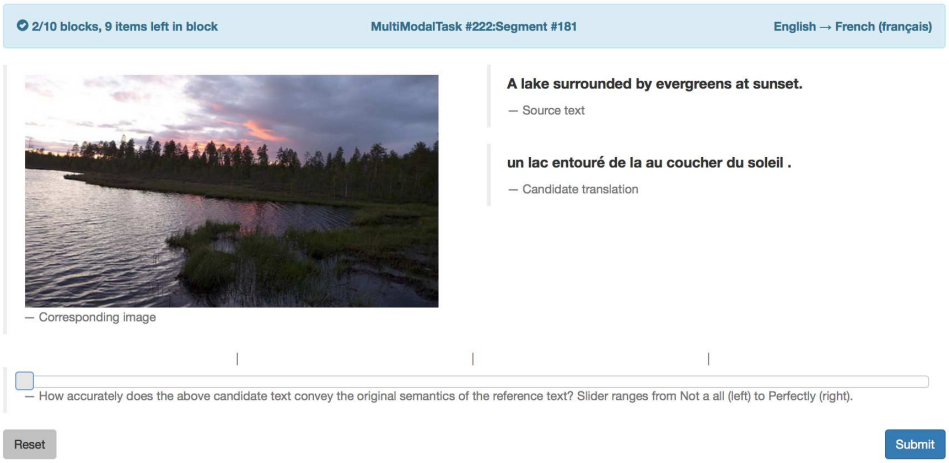

Fig. 2: An example of a direct assessment interface for multimodal translation. Human judges signify their assessment for the candidate translation, given the source text and the corresponding image, using a fine-grained sliding-bar interface.

2016; Elliott et al., 2017), which we refer to as MMT16 and MMT17. These shared tasks have focused on generating descriptions of images in non-English languages, by either translating parallel text or crosslingual description using independently collected sentences. At these shared tasks, and throughout the literature, multimodal translation is usually evaluated using text-based similarity metrics, e.g. the BLEU (Papineni et al., 2002) and Meteor scores (Denkowski and Lavie, 2014). However, these metrics are known to be problematic for machine translation and image description evaluation (Elliott and Keller, 2014; Kilickaya et al., 2017). More recently, multimodal translation has been evaluated using human direct assessment (Graham et al., 2015), in which humans express a judgement about the quality of a translation, given the source language description and image (Elliott et al., 2017). Figure 2 shows an example of the direct assessment interface for English→French MMT.

Human evaluation is extremely important for evaluating MMT models: In the MMT shared task, initial results based on automated metrics suggested that incorporating images into the translation process did not significantly outperform a text-only baseline (Specia et al., 2016). However, the use of human evaluation has confirmed that visual context does improve translation quality compared to text-only baselines (Elliott et al., 2017). In this paper, we study the human perspective of sentences that describe images in a multilingual corpus. In particular, we focus on two related issues: (i) do people have a preference for sentences translated from a different language or sentences written independently by speakers of their language? And (ii) what is the role of the image in translation and what types of disambiguation does it facilitate?

Ein Boston Terrier läuft über saftig-grünes Gras vor einem weißen Zaun.

Wie gut beschreibt dieser Satz das Bild? (required)

	1	2	3	4	5	6	7	
Sehr schlecht	○	○	○	○	○	○	○	Sehr gut

Fig. 3: The Crowdflower interface used to collect ratings of how well a sentence describes the given image. Participants are required to express a rating on a seven-point scale from "very badly" ("sehr schlecht") to "very good" ("sehr gut").

3 Quality Assessment of Native Language Descriptions vs. Translations

In this section we investigate whether, for the purposes of image description, there is any significant difference in quality between descriptions crowdsourced directly in the target language and translations into the target language.

This inquiry can inform decisions about resource creation and data collection for multilingual multimodal NLP. There are practical advantages to using translations from an existing dataset in another language. Professional translations require less quality control, compared to crowdsourcing new descriptions. This is especially valuable if the researchers do not speak the target language. The cost of collecting a single translation can be comparable to multiple crowdsourced descriptions, if aggressive quality control is necessary. Crowdsourcing can also be difficult for smaller languages with few workers on crowdsourcing platforms. Furthermore, if we find that human translation is adequate for image description, machine translation may soon also be of sufficiently high quality to automatically create descriptions in a new language. The Flickr8K-CN dataset was created by machine translating the original English descriptions into Mandarin (Li et al., 2016), but its quality has not yet been rigorously compared to human translations.

However, descriptions sourced directly in the target language may have the advantage of being more culturally appropriate than translation (van Miltenburg et al., 2017). Different languages tend to align with different cultures and different shared bodies of knowledge, which may have different ways of "carving up the world": these differences are reflected in what people focus on when describing the same image. The domain of sports offers clear examples of differences in cultural knowledge between the US (the source of the English descriptions in the Flickr30K dataset used for translation) and Germany, since different sports are culturally important in these countries. For instance, a description of people playing softball is likely to be confusing to a German speaker who may not know that there is a distinction be-

tween baseball and softball (or even that a sport named softball exists); conversely, a German speaker may describe an image depicting a soccer event in more detail than the average American would (and vice-versa for American football). Similar examples of what is considered shared knowledge (and thus appropriate to use in an image description) can be found in many other domains: famous buildings and other locations, traditional food and dress, celebrities, among many others. These are concrete entities and objects that are often depicted in images and a description that does *not* name a culturally recognisable object is jarring. For example, if an image of the Notre Dame in Paris does not mention the cathedral but only the automobiles on the nearby street (as found in the MS COCO dataset), then this would be an inappropriate description from the perspective of a French speaker (Elliott and Kleppe, 2016)).

In this study reported in this section, we asked German crowdworkers to rate how well a given sentence (which was either a translation from English or a description originally written in German) described an image, using a seven-point Likert scale. We then examined whether there is any consistent and significant difference between the ratings given to descriptions versus translations. If cultural differences are widespread in the images, descriptions would presumably be preferred over translations. However, if such differences are rare, there will be no clear difference between translations and descriptions. Moreover, if the translations are higher quality than the crowdsourced descriptions, we may even find a preference for the translations.

3.1 Materials and Procedure

The items for the study were taken from the WMT16 shared task test set, which is based on Multi30K. We chose 100 images at random; each image had an associated translation into German and five German descriptions (see Section 2.1 for more details). The translations were image-aware, that is, the translator was able to see the image while translating. We picked the description that was closest in length (counted in words) to the translation. A preliminary study had shown that length was a strong confounding factor when assessing translation quality. On average, the translations were still longer than the description, due to a handful of extremely long outliers (means: 11.2 vs. 10.1 words; medians: 11 vs. 10 words).

The data was collected in sets of five items, featuring four real items and one control item. Figure 3 presents the crowdsourcing interface used to collect a single item. The controls ("test questions") were descriptions applied to the wrong image, manually inspected to make sure that they did not match. Participants who did not give these items a score of '1' were automatically rejected. The order of the items was randomised across pages, as well as on each page.

We collected 10 ratings per image-sentence pair from German crowdworkers on the Crowdflower platform. Participation was restricted to those who had a Crowdflower language qualification for German and were at least Level 3 Crowdflower

workers.[6] 17 workers did not pass the test questions; we removed a further three participants who gave all items a rating of '1'. In total, 49 workers contributed to the final dataset, which consisted of 1,968 ratings. Each worker contributed an average of 40 ratings (max: 128, min: 4). Workers were paid $0.03 per item and the total cost of collecting the data was $90.36 at a rate of $6.53/hour.

3.2 Results

Overall, the ratings were very skewed towards the higher end of the scale. Two thirds of the items were rated 7, the highest rating, with a mean rating of 6.5 ($SD = 0.95$). This is to be expected, since these are human-generated descriptions: these results can also function as an oracle upper-bound for systems evaluation.

A Wilcoxon-Mann-Whitney test showed that the ratings given to descriptions ($M = 6.57, SD = 0.85$) were significantly higher than those given to translations ($M = 6.37, SD = 1.02$): $p = 2.08\text{e-}06$ ($U = 529970$, one-sided test). However, the effect size is very small, 2.4e-05, as measured by the Hodges-Lehmann Estimator, which captures the median difference in samples of descriptions and translations. This means that, while on average the descriptions will receive higher ratings than the translations, the median difference between the two is negligible, and so for practical purposes the two sentence types will receive equivalent ratings.

We were interested in whether other factors, particularly length, played a role in how participants rated the quality of an item. Note that descriptions were on average shorter, but were rated slightly better than translations. Did this mean that descriptions were overcoming a length disadvantage, or was length not an important factor for description quality?

We fit an ordinal regression model with mixed effects using the **ordinal** package in R. This model[7] had the rating (1–7) as the ordinal outcome variable, while the type of sentence (description or translation) and the length in words were fixed effect predictors with additional random effects for participant and item (intercepts only). The random effects capture the tendency for participants and items to have differing baseline ratings (e.g. a particular crowdworker may consistently rate items higher than other crowdworkers).

The estimated coefficients for the fixed effects are shown in Table 3. After controlling for subject and item effects, only the length coefficient is significantly different from zero, while the type of sentence is not. Likelihood ratio tests gave equivalent results, indicating that length is a significant predictor of rating (controlling for type), while adding type does not improve predictive power (controlling for length).

It is interesting to note here that these results seemingly contradict our earlier results from a Wilcoxon-Mann-Whitney test, in which on average descriptions were rated higher, while translations were on average longer: here we find that longer sentences will be rated higher, *caeteris paribus*, and the type of sentence does not

[6] This corresponds to the most trusted workers on the platform, at the time of writing.
[7] In R notation: `clmm(rating ~ type + length + (1|participant) + (1|item))`

Predictor	Estimate	SE	Wald's Z	p-value
type	0.197	0.299	0.66	0.51
length	0.072	0.029	2.48	0.013

Table 3: Summary of estimated values for the fixed effects in the ordinal regression model. The reference value for the 'type' predictor is description, so the estimate represents the increase in rating when going from description to translation. For 'length' the estimator represents the increase in rating gained when increasing sentence length by one word. The p-values are calculated using the Wald test, with the null-hypothesis that the value of the predictor is zero.

change this preference. This apparent contradiction can be resolved by noting the random effects structure included in the ordinal regression model. The random effects are a significant (both statistically and in magnitude) contributor to the goodness of fit of the full model (compared against a model with the same fixed effects but no random effects, $\chi^2(df = 2) = 821.5, p < 2.2\text{e-}16$). They allow the mixed effects model to control for high between-subject ($SD = 2.5$) and between-item ($SD = 1.3$) variation. Given that the data is crowdsourced, capturing and controlling for subject (crowdworker) variability is essential.

To return to the original question: we conclude that, for the images we tested, there was no consistent difference between the target-language descriptions and the translations from English. Other factors, specifically sentence length, are more important. The implications of this result are two-fold. Firstly, when building new multilingual image description resources, translating existing resources into a new language will most likely result in equally good descriptions as collecting new descriptions. Secondly, automatic multilingual image description generation can possibly rely on (machine) translation as a strategy, training on parallel text, rather than using comparable (but not parallel) sets of descriptions of the same image in multiple languages.

We note some caveats about the generalisability of these results. For one, English and German are relatively linguistically and culturally similar, likely providing a more straightforward translation path than for more distant language pairs. The Flickr images for the most part have a Western perspective shared by both Americans and Germans. For other domains and language pairs where the images are less familiar (or familiar for different reasons) to one or the other language or culture, it will likely remain important to go beyond translation to more flexible re-describing in the target language.

4 Multimodality in Translation of Image Descriptions

Having established that translations can function as replacements for image descriptions, we now turn to the question of how important the images themselves are for the translation process. The multimodal translation task is based on the assumption that translations — particularly translations of image descriptions and

	Sentences	Post-edited (PE)	Distance Original-PE
Validation set	1,014	6.11%	0.173
Test set	1,000	13.8%	0.157

Table 4: Percentage of translations post-edited and TER edit distance between their original and post-edited versions for the MMT16 validation and test sets.

other 'visual' language — will improve if humans or models take the image into account. In this section we aim to confirm this assumption and moreover quantify how adding accompanying images changes translations. To do so, we took text-only, image-blind, translations and then collected post-edits from a translator who can see the images. Post-edits capture the difference between a text-only translation and translation with the image supplied; if we collected separate translations (with and without images) there could be many spurious differences due to human translator decisions that are not necessarily related to presence or absence of the image.

4.1 Image-Aware Corrections

We used the test and validation sets from the MMT16 Shared Task, which had been translated without access to the image. We employed one professional translator to post-edit the original human translations, this time having access to the image along with the source text and the original translation. The post-editor was asked to fix only words that were deemed incorrect in the initial translation and to avoid any changes due to preferences or style.

Table 4 shows the percentage of sentences that were post-edited when a translator was presented with their corresponding image, as well as the average TER edit distance between the original translation and its post-edited version (calculated over post-edited sentences only). The TER (Translation Error Rate) edit distance is an adaptation of the Levenshtein minimum edit distance that includes word reordering as an operation: words or sequences of words can be reordered and this counts as a single edit operation. We computed this edit distance using the TERCOM tool[8].

The reasons behind the differences between the test and validation sets are not entirely clear. These datasets had been translated by the same translator, and they were post-edited by a different translator, who fixed both the validation and test sets. We can only hypothesise that the differences are due to specific features of the two original Flickr30K datasets. Based on the feedback received from the post-editor, the test set was perceived to contain more errors and inaccuracies in the original English descriptions as well as the translations.

In order to confirm whether or not the edits can be attributed to the presence of the images, and to further analyse which additional information the image brings in those cases, we manually checked all edits and categorised them into six categories:

[8] `http://www.cs.umd.edu/~snover/tercom/`

Category	% Validation	% Test
Lexical ambiguity	37.7	27.5
Conjunction ambiguity	1.7	2.2
Gender ambiguity	3.3	7.3
English description inaccurate	36.1	28.0
Original translation too literal	11.5	10.0
Original translation inaccurate	9.7	25.0

Table 5: Distribution of human post-edits in the MMT16 validation and test sets.

1. **Lexical ambiguity**: the edit corrects lexical choices which were the result of ambiguity/vagueness in the source text.
2. **Conjunction ambiguity**: qualifiers in conjoined noun phrases were attatched incorrectly.
3. **Gender ambiguity**: (natural) gender was not marked in English but needed to be marked in German; the edit corrects mistaken gender assignments.
4. **English description inaccurate**: the edit corrects errors due to incorrect or overly vague descriptions.
5. **Original translation too literal**: the edit improves the fluency or style of translation, even though its meaning was not incorrect.
6. **Original translation inaccurate**: other translation errors.

Figures 4 and 5 show examples of the post-edit corrections made for categories 1–5. Categories 1–3 are particularly important as they represent clear cases of ambiguity, where the image was critical to generate the correct translation. (There may be other potential instances of these categories in the dataset where the human translator used their best judgement to "guess" the correct sense and translation, lacking the information that the image would have provided.) Category 4 is an artefact of the Flickr30K dataset, but it also shows how images can help recover from inaccuracies in the original descriptions. Incorrect descriptions are a common problem with user generated content, such as the descriptions in the Flickr30K dataset. Category 5 covers cases where the original translation was correct but could be improved, which in some cases was made possible or facilitated by the presence of the corresponding image. Finally, category 6 covers all other cases where the original translation (without image) was not correct for reasons other than the absence of the image. These cases often happened because the translator was misled by their intuitions based on previous descriptions in the dataset and made incorrect assumptions about what should be the correct translation. This category also includes a few instances of typos and grammar mistakes.

The changes made by human translator when faced with the images corresponding to the English description tended to be very localised. The overall proportion of words edited was very low (2.2% in the test set, 1% in the validation set).

En: A child wearing a red coat and cap is holding a large chunk of snow.

De: Ein Kind in einem roten Mantel und einer Mütze hält einen großen Haufen Schnee.

PE: Ein Kind mit roter Jacke und Mütze hält ein großes Stück Schnee.

(a) **Conjunction ambiguity**

En: Three children in football uniforms of two different teams are playing football on a football field.

De: Drei Kinder in Fußballtrikots zweier verschiedener Mannschaften spielen Fußball auf einem Fußballplatz.

PE: Drei Kinder in Footballtrikots zweier verschiedener Mannschaften spielen Football auf einem Footballplatz.

(b) **Lexical ambiguity**

En: A man in a blue coat grabbing a young boy's shoulder.

De: Ein Mann in einem blauen Mantel hält einen Jungen an der Schulter.

PE: Ein Mann in einer blauen Jacke hält einen Jungen an der Schulter.

(c) **Lexical ambiguity**

Fig. 4: Examples of *conjunction ambiguity* and *lexical ambiguity* post-edits where the image was necessary for correct human translation.

4.2 Multimodal MT Systems Performance on Updated Dataset

The MMT16 shared task on multimodal MT was evaluated against text-only translations in the test set: we now consider whether using the image-aware translations for evaluation would change the results of the submitted systems. In particular, it is possible that the rankings might change, showing that some systems are better at translating descriptions where seeing the image makes a critical difference to the final translation — one might, for example, now expect the multimodal systems to outrank text-only systems.

We compared the overall performance of the participating systems in the MMT16 shared task, on both the original test set and the post-edited test set, using Meteor (Denkowski and Lavie, 2014), the official metric for the MMT16 shared task. We note that the overall small percentage of edits performed by humans is unlikely to make a significant impact in terms of automatic evaluation. The MT system output remained exactly the same, i.e. no re-training or fine-tuning (us-

En: A baseball player in a black shirt just tagged a player in a white shirt.

De: Ein Baseballspieler in einem schwarzen Shirt fängt einen Spieler in einem weißen Shirt.

PE: Eine Baseballspielerin in einem schwarzen Shirt fängt eine Spielerin in einem weißen Shirt.

(a) **Gender ambiguity**

En: The workers are surrounding a hole with a bucket.

De: Die Arbeiter decken ein Loch mit einem Eimer ab.

PE: Die Arbeiter stehen um ein Loch mit einem Eimer herum.

(b) **English description inaccurate**

En: A young man in a blue shirt grinds a rail on a skateboard in an urban area.

De: Ein junger Mann in einem blauen Shirt rutscht in einer städtischen Gegend über ein Geländer.

PE: Ein junger Mann in einem blauen Shirt fährt in einer städtischen Gegend über ein Geländer.

(c) **Original translation too literal**

Fig. 5: Examples of *gender ambiguity*, *inaccurate English description*, and *too literal translations* post-edits where the image was necessary for correct human translation.

ing the post-edited development set) was performed; only the gold-standard data was (marginally) different due to the post-edits.

Table 6 shows the relative difference in system performance when evaluated using the post-edited references as compared to the original ranking (Specia et al., 2016). The differences between performance on the two test sets are nonexistent or marginal and do not lead to any changes in the overall ranking of the systems. According to the original shared task results, there is no significant difference between systems that use visual cues and systems that do not use such cues; this remains the case when using image-aware translations for evaluation.

Overall, the performance of most systems slightly decreased when evaluated with the post-edited references. This probably indicates that systems are mimicking strong biases in the training data, such as the use of male gender in German for any type of unmarked noun in English. When these biases are fixed in the reference test data, the performance of these systems naturally drops. It has recently been

System ID	Δ **Meteor**
•LIUM_1_MosesNMTRnnLMSent2Vec_C	-0.2
•LIUM_1_MosesNMTRnnLMSent2VecVGGFC7_C	-0.2
•*SHEF_1_en-de-Moses-rerank_C	-0.1
1_en-de-Moses_C	-0.1
CMU_1_MNMT+RERANK_U	-0.1
HUCL_1_RROLAPMBen2de_C	-0.2
CMU_1_MNMT_C	-0.1
DCU_1_min-risk-baseline_C	0.0
LIUM_1_TextNMT_C	-0.1
DCU_1_min-risk-multimodal_C	-0.2
CUNI_1_MMS2S-1_C	-0.1
DCU-UVA_1_doubleattn_C	0.1
LIUMCVC_1_MultimodalNMT_C	0.1
DCU-UVA_1_imgattninit_C	-0.1
IBM-IITM-Montreal-NYU_1_NeuralTranslation_U	0.0
UPC_1_SIMPLE-BIRNN-DEMB_C	0.0
IBM-IITM-Montreal-NYU_1_NeuralTranslation_C	0.0
1_GroundedTranslation_C	0.0

Table 6: Difference in Meteor results for the MMT16 English–German task between using the original, image-unaware references, and the image-aware post-edited references. A negative difference indicates that the *original references*, i.e. the image-blind text-only translations, led to higher Meteor scores. The baseline systems are underlined. The winning submissions are indicated by a •. Submissions marked with a * are not significantly different from the text-only baseline (1_Moses_C).

shown that models can amplify these types of gender biases in multi-label object classification and visual semantic role labelling (Zhao et al., 2017). Object classification models constitute the basis for the image models used in many multimodal translation systems. In our case, these biases likely were made even stronger because the training data was translated based on the source descriptions only, rather than on the source descriptions and images.

4.3 Translator Perception of the Importance of Images

The post-editing results showed that the presence or absence of the relevant image affects description translation; here we ask to what extent translators rely on the image while translating.

We compare the two test sets used for evaluation for the MMT17 shared task: firstly, the official MMT17 test set of 1,000 descriptions of Flickr images in the same domain as those in the Multi30K dataset, and secondly, a new set of descriptions created to contain ambiguous verbs which ideally required the image for disambiguation during translation. For this second test set, which we refer to

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

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

Fig. 6: Two senses of the English verb "to pass" in their visual contexts, with the original English and the translations into German and French, taken from the Ambiguous COCO dataset. The verb and its translations are underlined.

as Ambiguous COCO, 461 additional descriptions were selected from the VerSe dataset (Gella et al., 2016), These contain a selection of 56 ambiguous verbs from VerSe appearing in descriptions of MSCOCO images, e.g. *stir*, *pull*, *serve*, with 1–3 instances per sense per verb. The number of instances per verb varies from 3 (e.g. *shake*, *carry*) to 26 (*reach*). We refer the reader to (Elliott et al., 2017) for more details about the dataset.

Both the MMT17 and Ambiguous COCO datasets were translated by the same professional translator in an image-aware setting. In both cases, we asked the translator performing the task to select, after each description was translated, whether or not the image was perceived as "needed" in the translation for whatever reason, e.g. to help disambiguate words or better understand the source description in any way. For example, consider the images of the English verb "to pass" from the Ambiguous COCO dataset shown in Figure 6. In the German translations, the source language verb did not require disambiguation (both German translations use the verb "fährt"), whereas in the French translations, the verb was disambiguated into "dépasse" and "traverse", respectively.

For the WMT17 dataset from Flickr, the image was explicitly judged as needed in 20% of the descriptions, while for the Ambiguous COCO dataset, in 49% of the descriptions. Although a control group was not used to test whether the translations would have been different without the images in this dataset, this large proportion shows that – if nothing else – having access to images makes the translator perceive the translation process as easier, and that English verb ambiguity seems to often transfer into translation ambiguities that can be resolved with the help of the image.

The results in this section have shown that language with a visual context such as image descriptions benefits from image-aware translations, as demonstrated both by translation post-editing and the translator's subjective perception of how much they relied on the images. If human translators, who are professionals at using background knowledge and context to arrive at the correct translations of ambiguous short texts, can improve their translations with the aid of images, automatic translation systems should also be able to benefit.

5 Conclusion

This paper examined two of the assumptions underpinning work on crosslingual image description and multimodal machine translation, namely, that native language descriptions (or generated descriptions that are closer to native language descriptions) are preferable to translations, and that the image is important for the translation of language with a visual context such as image descriptions. We performed a human evaluation experiment to assess the former and a post-editing procedure plus error analysis to assess the latter.

We found that on the whole these assumptions do not entirely hold: a statistical analysis failed to show meaningful differences between the ratings for translations versus native language descriptions, and the post-edit rate from image-blind to image-aware translations was quite low. However, even though these results may seem to imply that simple methods such as text-only translation can often lead to reasonable outcomes, there remain cases where access to the image is essential for translation and where image descriptions should be not simply be translated.

We note that our findings are based on human generated descriptions and translations. Humans use their background knowledge to make sense of short contexts and often correctly guess the right, or at least acceptable, translations of ambiguous source texts. This task is certainly much more complex for computational systems, which may result in multimodality playing a bigger role in machine translation. Human translators are moreover able to adapt descriptions to make them appropriate for a target language and culture. Therefore, a comparison between translations and native language descriptions generated by automatic systems would likely lead to different results.

Our findings are also contingent on the procedure used for collecting the original images and descriptions, which resulted in literal descriptions of fairly straightforward images. Other visual domains (e.g. instruction manuals) may require more attention to the image during translation; other kinds of image-related language data may be harder to translate without the image. Nevertheless, this paper analysed the standard dataset for multimodal machine translation, so our findings are intended to inform future work in the field.

A further caveat concerns the language pair used in this work: German and English are closely related languages and also share significant amounts of cultural knowledge. Future work should also investigate multilingual image description with language pairs that are more distantly related both linguistically and culturally.

The studies presented here can be seen as an examination of how to evaluate

image description: what is an appropriate gold standard to evaluate against? This is a particularly important question to ask when developing a dataset for a new language, since the available resources constrain the direction of future research. The first study indicates that there is no *a priori* reason to discount (image-aware) translation as a source of high quality image descriptions. It also means that image description systems (either based on multimodal MT or crosslinguistic methods) can use translations as a gold standard in evaluation. It will, however, remain important to be aware of potential cultural differences, for example by developing methods for identifying cases in which translation is inappropriate.

The second study is relevant to whether text-only 'image-blind' translations are an appropriate gold standard for evaluating multimodal MT, specifically evaluation using automatic metrics like BLEU or Meteor. The post-editing resulted in only a small number of words being changed, albeit often with significant semantic impact. The minor changes meant that the difference between using text-only and image-aware reference translations led to only minor differences in system evaluations using Meteor. Rather than concluding that text-only translations may be used in evaluation, we take these findings to indicate that automatic metrics should not be used for multimodal MT, since these metrics are not sufficiently sensitive to the information provided by the image (i.e. the difference between pre- and post-edited translations). The difficulties in evaluating multimodal MT are similar to those faced in evaluating discourse-level MT, where small changes (e.g. pronoun choice) often have significant semantic consequences. Given the inappropriateness of word-overlap metrics such as BLEU, the discourse-level MT community has developed sub-tasks focussed on specific translation problems that require discourse awareness, such as pronoun prediction (Hardmeier et al., 2015). Attempts in multimodal MT to create a similar test set of ambiguous instances that require image information, such as the Ambiguous COCO dataset, are a promising future direction.

The increasing use of images and video online, along with the decreasing dominance of English, will make multilingual multimodal NLP important in the future. This paper has furthered research in this direction by delineating the contributions of multimodality in (human) translations and by assessing the different possible sources of image descriptions. However, the main challenges remain: how should we represent such visual cues from images (a job that humans can easily do), and how should such information be used in translation and description models.

References

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. (2015). VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, Santiago, Chile.

Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A., and Plank, B. (2016). Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artifical Intelligence Research*, 55:409–442.

Berzak, Y., Barbu, A., Harari, D., Katz, B., and Ullman, S. (2015). Do you see what I mean? Visual resolution of linguistic ambiguities. In *Proceedings of the 2015 Confer-*

ence on Empirical Methods in Natural Language Processing, pages 1477–1487, Lisbon, Portugal.

Chen, D. L. and Dolan, W. B. (2011). Building a persistent workforce on Mechanical Turk for multilingual data collection. In *Proceedings of The 3rd Human Computation Workshop (HCOMP 2011)*, San Francisco, CA, USA.

Denkowski, M. and Lavie, A. (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.

Duygulu, P., Barnard, K., de Freitas, J. F., and Forsyth, D. A. (2002). Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *European conference on computer vision*, pages 97–112, Copenhagen, Denmark.

Elliott, D., Frank, S., Barrault, L., Bougares, F., and Specia, L. (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.

Elliott, D., Frank, S., Sima'an, K., and Specia, L. (2016). Multi30K: Multilingual English-German Image Descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, Berlin, Germany.

Elliott, D. and Keller, F. (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.

Elliott, D. and Kleppe, M. (2016). 1 million captioned Dutch newspaper images. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

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

Graham, Y., Baldwin, T., Moffat, A., and Zobel, J. (2015). Can machine translation systems be evaluated by the crowd alone? *Natural Language Engineering*, 23(1):3–30.

Hardmeier, C., Nakov, P., Stymne, S., Tiedemann, J., Versley, Y., and Cettolo, M. (2015). Pronoun-focused MT and cross-lingual pronoun prediction: Findings of the 2015 DiscoMT shared task on pronoun translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 1–16, Lisbon, Portugal.

Hitschler, J., Schamoni, S., and Riezler, S. (2016). Multimodal Pivots for Image Caption Translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 2399–2409, Berlin, Germany.

Hodosh, M., Young, P., and Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Hollink, L., Bedjeti, A., van Harmelen, M., and Elliott, D. (2016). A Corpus of Images and Text in Online News. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Kilickaya, M., Erdem, A., Ikizler-Cinbis, N., and Erdem, E. (2017). Re-evaluating automatic metrics for image captioning. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 199–209, Valencia, Spain.

Lazaridou, A., Pham, N. T., and Baroni, M. (2015). Combining language and vision with a multimodal skip-gram model. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 153–163, Denver, Colorado.

Li, X., Lan, W., Dong, J., and Liu, H. (2016). Adding Chinese captions to images. In *Proceedings of the 2016 ACM International Conference on Multimedia Retrieval*, pages 271–275, New York, NY, USA.

Miyazaki, T. and Shimizu, N. (2016). Cross-lingual image caption generation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1780–1790, Berlin, Germany.

Ordonez, V., Kulkarni, G., and Berg, T. L. (2011). Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*, pages 1143–1151, Granada, Spain.

Papineni, K., Roukos, S., Ard, T., and Zhu, W.-J. (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.

Rajendran, J., Khapra, M. M., Chandar, S., and Ravindran, B. (2016). Bridge correlational neural networks for multilingual multimodal representation learning. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 171–181, San Diego, CA, USA.

Ramanathan, V., Joulin, A., Liang, P., and Fei-Fei, L. (2014). Linking people in videos with "their" names using coreference resolution. In *European Conference on Computer Vision*, pages 95–110, Zurich, Switzerland.

Ramisa, A., Yan, F., Moreno-Noguer, F., and Mikolajczyk, K. (2017). BreakingNews: Article annotation by image and text processing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1.

Silberer, C. and Lapata, M. (2014). Learning grounded meaning representations with autoencoders. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 721–732, Baltimore, Maryland.

Specia, L., Frank, S., Sima'an, K., and Elliott, D. (2016). A shared task on multimodal machine translation and crosslingual image description. In *First Conference on Machine Translation, Volume 2: Shared Task Papers*, WMT, pages 540–550, Berlin, Germany.

Unal, M. E., Citamak, B., Yagcioglu, S., Erdem, A., Erdem, E., Cinbis, N. I., and Cakici, R. (2016). Tasviret: Görüntülerden otomatik türkçe açıklama oluşturma İçin bir denektaçı veri kümesi (TasvirEt: A benchmark dataset for automatic Turkish description generation from images). In *IEEE Sinyal İşleme ve İletişim Uygulamaları Kurultayı (SIU 2016)*.

van Miltenburg, E., Elliott, D., and Vossen, P. (2017). Cross-linguistic differences and similarities in image descriptions. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 21–30, Santiago de Compostela, Spain.

Yoshikawa, Y., Shigeto, Y., and Takeuchi, A. (2017). STAIR captions: Constructing a large-scale Japanese image caption dataset. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 417–421, Vancouver, Canada.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (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.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2941–2951, Copenhagen, Denmark.