A Dataset and Reranking Method for Multimodal MT of User-Generated Image Captions

Shigehiko Schamoni schamoni@cl.uni-heidelberg.de
Julian Hitschler hitschler@cl.uni-heidelberg.de
Department of Computational Linguistics, Heidelberg University, 69120 Heidelberg, Germany
Stefan Riezler riezler@cl.uni-heidelberg.de
Department of Computational Linguistics and IWR, Heidelberg University, 69120 Heidelberg, Germany

Abstract

We present a dataset and method for improving the translation of noisy image captions that were created by users of Wikimedia Commons. The dataset is multilingual but non-parallel, and is several orders of magnitude larger than existing parallel data for multimodal machine translation. Our retrieval-based method pivots on similar images and uses the associated captions in the target language to rerank translation outputs. This method only requires small amounts of parallel captions to find the optimal ensemble of retrieval features based on textual and visual similarity. Furthermore, our method is compatible with any machine translation system, and allows to quickly integrate new data without the need of re-training the translation system. Tests on three different datasets showed that size and diversity of the data is crucial for the performance of our method. On the introduced dataset we observe consistent improvements of up to 5 BLEU points and 3 points in Character F-score over strong neural MT baselines for three different language pairs.

1 Introduction

Image caption translation is the task of translating a caption associated with an image into another language. What differentiates this task from purely text-based machine translation is the incorporation of image information into the translation process. Images associated with text usually add a new modality of information. Such information helps to ground the meaning of the corresponding text and is thus especially useful in a translation task. Interest in this task has surged since the first instantiation of the shared task on multimodal machine translation where a dataset of 30,000 German translations of crowdsourced English captions was presented (Specia et al., 2016). However, this dataset has limitations: The captions were created by human annotators that were guided to produce "conceptual" descriptions that identify the objects depicted in the image (Hodosh et al., 2013). This leads to relatively short captions amounting to a comparatively easy translation task with little room for improvement by incorporating visual information. This is confirmed by recent results showing that improvements over a text-only MT baseline are inconsistent and hard to achieve (Lala et al., 2017; Elliott et al., 2017).

While caption translation in previous work has been conducted solely on clean, manually labeled captions based on MS COCO (Lin et al., 2014), Flickr30k (Rashtchian et al., 2010), or its multilingual variant Multi30k (Elliott et al., 2016), the goal of our work is to lift multimodal caption translation to a more realistic setup. For this purpose, we extracted a dataset of 4M "cap-

tions in the wild" as they appear in the user-generated Wikimedia Commons database. This new dataset is very different to previous image-caption data, as it contains highly diverse types of user-generated texts associated with images. The English captions in this dataset are around 34 tokens long, compared to 11 and 14 for MS COCO and Multi30k, respectively. Caption translation of Wikimedia Commons data thus contrasts to previous image-caption translation tasks. However, we find the new dataset to provide a lot of room for improvement by incorporating visual information into the translation process. The dataset is described in Section 4.

Since the dataset only contains very small subsets of parallel captions (which we use for tuning and testing), the proper way to integrate visual information is to leverage monolingual image-caption pairs. Hitschler et al. (2016) presented an approach based on a crosslingual reranking framework where monolingual captions in the target language are used to rerank translation hypotheses given a source caption and the corresponding image. In order to retrieve captions for reranking, they pivot on target language image-captions pairs in two ways: A list of monolingual captions is obtained by a joint textual and visual similarity model by comparison between the hypotheses and the captions in the target language. To calculate the visual similarity component of their joint model, they use rich image feature representations from a convolutional neural network (Simonyan and Zisserman, 2015). Our approach is an extension of Hitschler et al. (2016), who rely on manually tuned hyperparameters, to a pairwise ranking approach to learn an optimal ensemble of different rerankers. We also implement separate textual and visual similarity components to incorporate them as distinct features into our reranking model. Furthermore, we investigate a stronger text-only baseline that is based on neural MT (NMT). Our translation and reranking methods are described in Section 3.

We present an evaluation on caption data from Wikimedia Commons. We find gains of 5 BLEU points (Papineni et al., 2002) and 3 points in Character F-score (Popović, 2015) by reranking over strong NMT baselines across three different language pairs. In order to discern the contribution of our new learning method, we compare our approach to the only other monolingual reranking approach that we are aware of, namely Hitschler et al. (2016). On the MS COCO data, we observe gains by neural MT over phrase-based MT, and small but consistent gains by reranking. We also evaluate our approach on the Multi30k (Elliott et al., 2016) dataset that was used for the WMT17 Multimodal Shared Task 2 (Elliott et al., 2017). Due to the limited size of data available for retrieval we found no significant improvements over the NMT baseline here. Our experiments indicate a strong dependency of our approach's performance on the type and size of retrieval data. The experiments are described in Section 5.

2 Related Work

The dataset presented in this paper is to our knowledge the first publicly available resource of user-generated image captions at the size of 4M image-caption pairs. The dataset that is closest to ours is the SBU captioned photo dataset (Ordonez et al., 2011) that contains 1M images and captions. However, this dataset was filtered to include specific terms and to limit description lengths, resulting in an average sentence length of around 13 tokens. See Ferraro et al. (2015) for an overview over image-caption datasets.

Multimodal caption translation on parallel caption data (see the approaches described in Specia et al. (2016)) incorporate visual information directly into the sequence-to-sequence caption translation model or into a reranking component, or into both (see for example the attention-based LSTM approach of Huang et al. (2016)), or they use back-translation to generate synthetic parallel data (see for example Calixto et al. (2017)). However, obtaining parallel captioning data or retraining NMT models on large synthetic datasets is either financially or computationally expensive. We thus opt for a way that does not require large amounts of parallel captions to improve translation quality.

	Hitschler et al. (2016)	Our work
Dataset	MS COCO: clean, limited vocabulary	Wikimedia Commons: captions "in the wild"
Retrieval	multimodal joint model	orthogonal models for image & text
Reranker	interpolation of scores	trained model-based reranker
MT-System	Statistical (`cdec`)	Neural (`nematus`)
Languages	de-en	de-en, fr-en, ru-en

Table 1: Comparison of Hitschler et al. (2016) to our work.

Our work can be seen as an extension of the idea of Wäschle and Riezler (2015) to multimodal data. Their approach is based on cross-lingual retrieval techniques to find sentences in a large target-language document collection, which are then used to rerank candidate translations. Our approach uses textual relevance and visual similarity (see Section 3.2) to obtain lists of multimodal pivot documents from a monolingual image-caption corpus similar to the idea described in Hitschler et al. (2016). In contrast to their approach, we do not rely on grid search for hyperparameters but instead use a machine learning approach to determine optimal weights of different rerankers to get the best scoring ensemble. In order to discern the contribution of our learning method, we compare it to the monolingual reranking approach of Hitschler et al. (2016) on the MS COCO v2014 dataset that was used in their work.

3 Method

3.1 Overview

Following the idea of Hitschler et al. (2016), we use retrieval models to find similar images and image captions in a target language image dataset to rerank target language caption translations e of a source caption f. This approach does not rely on large amounts of parallel data, but only requires monolingual target image-captions pairs. Modularizing the retrieval and translation component makes our method applicable to any existing translation system. Additionally, we can make use of new retrieval data instantly without expensive retraining of the translation model. This is a valuable property for active production systems where larger amounts of images and associated texts are available, e.g. in online shops.

The main difference between our model and Hitschler et al. (2016) is the way we implement the multimodal retrieval component and the reranker. We do not combine the visual and textual similarity models in a joint model, but let our model chose the best ensemble of rerankers that operate on top-k lists generated on visual and textual similarity. Our motivation behind separating textual and visual components is that textual and visual retrieval provide orthogonal information which can be best combined in an ensemble of reranking components operating on different feature sets. Thus, we do not manually tune any hyperparameters and instead apply a supervised training approach following a learning-to-rank strategy. See Table 1 for an overview comparison.

3.2 Multimodal Retrieval Model

The middle part of Figure 1 illustrates the textual and visual retrieval components for selecting pivot documents (i.e. image-caption pairs) from a target document collection m. For each source image-caption pair (i, f_i), our model uses the image i and a decoder-generated target hypothesis list N_{f_i} to yield two lists of pivot documents, namely a list M_i based on visual similarity, and a list L_{f_i} based on textual similarity. These lists are input to the parameterized retrieval score function defined in Section 3.3.

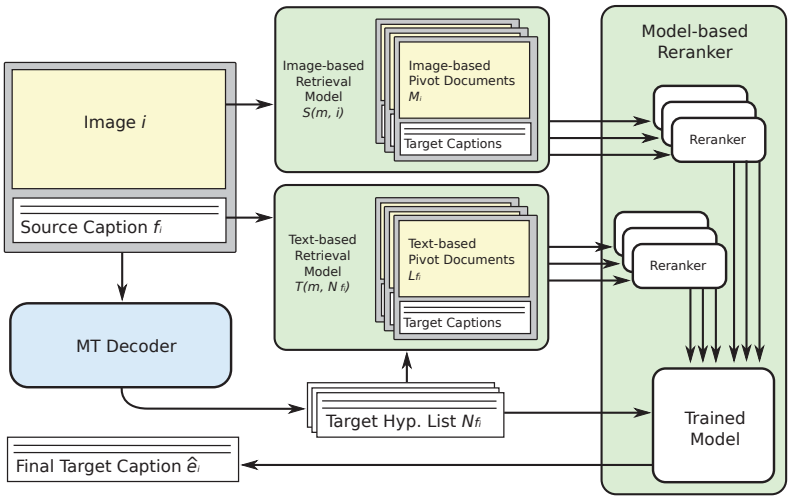

Figure 1: Given a source image-caption pair (i, f_i), we apply image-based and text-based retrieval models to obtain separate lists of pivot documents (M_i and L_{f_i}). These lists are input to multiple reranker components, which are combined in a model-based reranker. Final output is the highest scoring target caption $\hat{e}_i$ selected from the decoder's hypothesis list N_{f_i}.

Pivot documents based on textual similarity Using the translation hypotheses N_{f_i} as query against the monolingual target document collection, we select the top-k most similar pivot documents using the standard TFIDF metric from information retrieval. Given a target document collection m and the translation hypotheses N_{f_i}, the text-based retrieval model $T(m, N_{f_i})$ returns a list of image-caption pairs L_{f_i} ordered by an unsmoothed TFIDF score (Spärck Jones, 1972).

Pivot documents based on visual similarity This list consists of the top-k nearest neighboring image-caption pairs to the source image in visual space. Our distance metric $s(i, j)$ between two images i and j is the cosine of the 4,096-dimensional feature representations $\mathbf{v}_i$ and $\mathbf{v}_j$ of images i and j taken from the penultimate layer of the VGG16 model by Simonyan and Zisserman (2015), which was pre-trained on ImageNet (Russakovsky et al., 2014): $s(i, j) = \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{\|\mathbf{v}_i\|_2 \|\mathbf{v}_j\|_2}$. Given a target document collection m and a source image i, the image-based retrieval model $S(m, i)$ returns a list of image-caption pairs M_i sorted by visual similarity.

3.3 Parameterized Retrieval Score

We formulate a parameterized retrieval score function as follows. Based on a relevance score function $g_m(x, y)$ that returns the relevance of caption y to translation hypothesis x, define a retrieval score $RS_{r,m}(h, t_k)$ that calculates the average relevance of a hypothesis h to a sequence t_k of top-k retrieved captions up to a cutoff level r:

$$RS_{r,m}(h, t_k) = \frac{1}{N_{g_m}} \sum_{n=1}^{r} g_m(h, t_{k,n}). \quad (1)$$

Here, $t_{k,n}$ denotes the n^{th} element of the sequence, and N_{g_m} is a normalization parameter for a given choice of relevance function g_m.

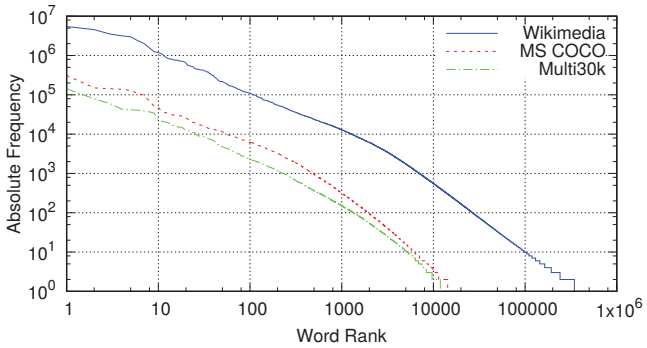

Figure 2: Absolute frequencies of lowercased words against their rank in Multi30k, MS COCO and Wikimedia Commons data, illustrating the differences in corpus size and vocabulary of the three corpora.

The retrieval score defined in Equation (1) can make use of different top-k caption lists, L_{f_i} or M_i, based on textual or visual similarity, respectively. The relevance function $g_m(x, y)$ can be instantiated to any retrieval score or similarity function that suits our needs. In our experiments, we applied the standard TFIDF metric from information retrieval and smoothed sentence-based BLEU (S-BLEU) (Chen and Cherry, 2014). The normalization parameter N_{g_m} depends on the relevance function and is in our case either the number of top-k captions for S-BLEU or the number of words for TFIDF.

Based on this formulation we combine up to 36 ranking functions defined on different top-k sequences, relevance functions, and cutoff levels. This setup can be easily extended by additional relevance score functions.

3.4 Learning to Rank Captions

Our final ranking score function $RE_\theta, \theta \in \mathbb{R}^{37}$ is defined as linear combination of up to 36 retrieval scores plus 1 translation model score as additional feature as follows:

$$RE_\theta(h) = \sum_i \alpha_{m,r,t_k} \cdot RS_{r,m}(h, t_k), \text{ where } \alpha_{m,r,t_k} = \theta_i. \quad (2)$$

We optimize this model by pairwise ranking to determine the importance of each reranker as follows: given a list of hypothesis ordered by a metric reflecting the translation quality with respect to a reference, our system should rank a higher scoring hypothesis above a lower scoring one. We used Character F-score as the metric to optimize throughout our experiments.

This is formalized by the following hinge-loss objective, where h^+ is a higher and h^- is a lower scoring hypothesis, H is the set of all such pairings in the training set, and $RE_\theta()$ denotes the current ensemble of rerankers defined by θ:

$$\underset{\theta}{\text{argmin}} \sum_{(h^+,h^-)\in H} \max(RE_\theta(h^-) - RE_\theta(h^+), 0). \quad (3)$$

For each source sentence, we pair all translation hypotheses which differ in Character F-score with respect to the reference translation. Thus, we can extract up to $\frac{n(n-1)}{2}$ pairs for each sentence given a hypothesis list of n elements. At $n = 100$ our models need not more than 500 to 1,000 parallel image-captions pairs in practice. Training was done using the `Vowpal Wabbit` toolkit.[1]

[1] `http://hunch.net/~vw/`

	Multi30k	MS COCO v2014	Wikimedia Commons
Images with captions	30,014	82,783	7,073,243
with English captions	30,014	82,783	4,149,659
Captions per image	5	5	1.14 on avg.
Caption language(s)	English, German	English	English, German, French, Russian, ...
Type of caption	descriptive	descriptive	indeterminable
Avg. tokens per English caption	13.51	11.32	34.35
Unique types	18,078	24,117	738,479

Table 2: Comparison between Multi30k, MS COCO and Wikimedia Commons data.

The rightmost box in Figure 1 illustrates how our model-based reranker uses information from image-based and text-based pivot documents M_i and L_{f_i} in an ensemble of rerankers to select the highest scoring caption translation $\hat{e}_i$ from a decoder-generated hypothesis list N_{f_i}.

4 Data: Manually Annotated Captions versus Captions "In the Wild"

Image caption translation has been mostly applied to relatively clean data, where monolingual captions are generated by annotators (Flickr30k, Multi30k, MS COCO) and translations are afterwards added by translators (Elliott et al., 2016; Hitschler et al., 2016). Our work investigates the questions whether image caption translation can make use of much noisier and more natural datasets, such as captions found in Wikimedia Commons. Thus, we conduct experiments on two fundamentally different types of captions, i.e. Multi30k and MS COCO on one side, and Wikimedia Commons on the other side. The two types of corpora are clearly distinguishable in Figure 2, which shows the absolute frequencies against the rank of lowercased words in all three corpora: Multi30k and MS COCO are relatively close, while our Wikimedia Commons dataset contains much more tokens and types. Furthermore, the latter has considerably more images and significantly longer captions. Table 2 lists the main characteristics of the three datasets.

Descriptive Captions Most work on image caption translation is done on artificially created captions that were generated by annotators based on clear instructions. Such captions are usually descriptive[2] and omit named entities and other information not present in the image. Table 2 shows that the average sentence length of our descriptive captions lies around 11 and 13 tokens. Figure 3 (left) gives an example of the descriptive captions found in the MS COCO dataset. Captions found in the Multi30k dataset are very similar by their nature.

Captions in the Wild Largely available image-caption data in the web does not fall into the category described in the previous paragraph. For the most part, existing image-caption pairs are not strongly descriptive, but mention product names, locations, and other aspects that are not obviously encoded in the image. Table 2 shows that the average sentence length of captions in Wikimedia Commons is about 34 tokens. See Figure 3 (right) for an example of a typical caption from Wikimedia Commons.

5 Experiments

5.1 Image-Caption Data

We constructed a retrieval dataset based on English image-caption pairs from a recent dump of Wikimedia Commons created by the `wikimgrab.pl` utility.[3] We filtered out images with

[2] Hodosh et al. (2013) call such captions *conceptual descriptions*.

[3] `https://commons.wikimedia.org/wiki/User:AzaToth/wikimgrab.pl`

- A bunch of boats parked at a busy and full harbor.
- A group of boats floating on top of a river near a city.
- A group of boats on water next to pier.
- Boats lined up in rows in water at the dock
- Motor boats parked near a dock in a marina.

- View of a disused pier in North Woolwich, London Borough of Newham, London.

London, North-Woolwich, Thames 27.jpg by *Kleon3* is licensed under CC BY-SA 4.0.

Figure 3: Image-caption examples from MS COCO (left) and Wikimedia Commons (right).

extreme aspect ratios (>3:1) and kept only images in JPEG- and PNG-format, as other formats are likely to contain data not useful for our approach (e.g. SVG-encoded logos or PDF-scanned document pages). Furthermore, we randomly selected 2,000 images with parallel captions for each of three language pairs, German-English (de-en), French-English (fr-en), and Russian-English (ru-en). One half of the parallel data is used as development set for training the reranker, the other half is used for testing.

During retrieval, we only use images and captions that were not included in the development or test set data, totaling in 3,816,940 images with mostly a single corresponding caption. In case of duplicate images, we discarded retrieval matches with an identical image to the query image or where the annotated target caption was identical to the gold standard target caption.

We applied the standard utilities from the `cdec`[4]-toolkit to tokenize and convert the data, namely `tokenize-anything.sh`, `lowercase.pl`. Parallel captions were additionally filtered by `filter-length.pl`. Table 3 gives an overview of the visual and textual data sources we used in our experiments. Examples for image-caption pairs from our dev and test sets together with a script to retrieve the full data set is available for download.[5] The data is released in accordance to the respective licenses.

5.2 Translation Systems

We trained our baseline translation system with Nematus (Sennrich et al., 2017), a state of the art toolkit for neural machine translation.[6] We tokenized and converted all training data to lower case using the same `cdec` utilities as were used for pre-processing of the retrieval data. In addition, we performed 20,000 steps of byte pair encoding (Sennrich et al., 2016) on the input and output vocabularies, giving our systems an open output vocabulary in principle. We used default parameters for learning, measured the cross-entropy of a held-out validation set after processing every 10,000 training samples and stopped training accordingly. For training and validation data, we filtered out sentences longer than 70 words (before byte pair encoding).

[4] `https://github.com/redpony/cdec`

[5] `http://www.cl.uni-heidelberg.de/wikicaps`

[6] `https://github.com/rsennrich/nematus`, git commit hash (unique revision identifier): `54be147dc363603d69643c35b700ae5d9de2ad93`

	Images	Captions	Languages
development	$1,000$	$1,000$	de-en, fr-en, ru-en
test	$1,000$	$1,000$	de-en, fr-en, ru-en
retrieval	$3,816,940$	$3,825,132$	en

Table 3: Number of images and captions in the dataset extracted from Wikimedia Commons.

For reranking and retrieval, we generated n-best lists of length $n = 100$ using beam search. The same beam size of $n = 100$ was used for our baseline systems. We used the following training data for the three language pairs:

French-English We trained our French-English translation system on data made available for the WMT 2015 translation shared task.[7] We used the Europarl, News Commentary and Common Crawl data for training.

Russian-English Our Russian-English translation system was trained with the Europarl and News Commentary data from the WMT 2016 shared task on news translation.[8]

German-English In order to enable direct comparison with Hitschler et al. (2016), we used the same training data as was used for their statistical machine translation system (Europarl, News Commentary and Common Crawl Data as provided for WMT 2015). For the experiments on COCO, we domain-adapted our system on the same data as the in-domain system of Hitschler et al. (2016), the corpus of parallel image captions provided for the WMT 2016 shared task on multimodal machine translation.[9] This was achieved by continuing training on the in-domain data once training was complete on the out-of-domain training data.

5.3 Ranking Components

The rankers we combine operate on features that make use of retrieval- and translation-based metrics such as TFIDF and S-BLEU. These features are extracted from textual- or visual-retrieval-based pivot documents. To evaluate the contribution of different combinations of rerankers in an ablation experiment, we use the following letters to identify the types of rerankers in a combination:

- **`T`** : textual-retrieval-based TFIDF
- **`B`** : textual-retrieval-based S-BLEU
- **`V`** : visual-retrieval-based TFIDF
- **`W`** : visual-retrieval-based S-BLEU

For example, a system that operates solely on textual-retrieval-based TFIDF and S-BLEU (T,B) is labeled TB, while a system that uses textual- as well as visual-retrieval-based S-BLEU (B,W) and visual-retrieval-based TFIDF (V) is labeled BWV. A system that is trained on all available rerankers is labeled TVBW.

The length of the hypothesis list for the reranker was selected on dev using Character F-score as our primary metric for tuning. See Tables 8, 9, 10, and 11 for details.

We also modified the length of pivot document list across rankers and implemented different cutoff levels for the similarity as defined in Equation 1 at 3, 5, 10, 20, 50, and 100. In total we obtain 36 different rerankers. Note that it is straightforward to add new rerankers based on different similarity score functions or cutoff levels.

[7] `http://www.statmt.org/wmt15/translation-task.html`
[8] `http://www.statmt.org/wmt16/translation-task.html`
[9] `http://www.statmt.org/wmt16/multimodal-task.html`

	baseline	T	B	TB	V	W	TV	BW	BWV	TVBW
BLEU	21.66	21.44	21.55	21.54	21.34	21.33	21.19	21.84	21.90	21.76
Character-F	49.35	49.16	49.29	49.19	49.33	49.11	49.28	49.22	49.18	49.19

Table 4: BLEU and Character F-scores on German-English Multi30k test data from the WMT17 Multimodal Task 2 (Elliott et al., 2017). Due to the limited data available for retrieval our approach did now show improvements over the `nematus` baseline.

	System	baseline	T	B	TB	V	W	TV	BW	BWV	TVBW
BLEU	`cdec` in-dom. (Hitschler et al., 2016)	29.6	–	–	–	–	–	–	–	–	–
BLEU	TSR-CNN (Hitschler et al., 2016)	30.6	–	–	–	–	–	–	–	–	–
BLEU	New reranker	29.6	26.96	28.47	30.83	27.54	30.59	‡31.04	†30.79	30.67	30.76
BLEU	Our system	33.78	33.88	33.27	32.97	33.92	32.46	34.03	34.30	34.21	34.40
ChF		62.74	62.93	62.37	62.29	63.01	61.70	62.90	62.91	62.86	63.00

Table 5: BLEU and Character F-scores (ChF) on de-en MS COCO test data from Hitschler et al. (2016) for their `cdec` in-domain and TSR-CNN systems, the new reranker applied to the `cdec` hypothesis lists, and our new system on different combinations of rerankers. Significant improvements over the baseline system are indicated by preceding † ($p < 0.03$) and ‡ ($p < 0.003$) as reported by MultEval's randomization test (Clark et al., 2011).

5.4 Results

Our experiments revealed a strong connection between certain properties of retrieval data and the performance of our approach. On very clean, manually constructed data of limited size and complexity like the Multi30k dataset, our retrieval-based method fails to extract additional useful information. As the data available for retrieval grows, we observe small gains like in the experiments on MS COCO. The biggest improvements, however, can be found on the much larger and inherently diverse Wikimedia Commons dataset we described before. We discuss the results of an ablation experiment for various combinations of rerankers on the different datasets in detail in the following paragraphs.

Multi30k The results listed in Table 4 show that the retrieval-based approach does not lead to gains in BLEU or Character F-score on the Multi30k dataset. We see two main reasons for this: Firstly, the dataset available for retrieval is orders of magnitude smaller that the Wikimedia Commons dataset. Secondly, the descriptive captions are of low complexity, e.g. built of a small vocabulary and simple sentence structure. Thus, the additional information contributed to the reranking step by pivot images and documents is very limited. The necessity to restrict reranking to short hypothesis lists (top-5 only, see Table 9) underlines this problem, as the ensemble of rerankers is not able to select better translations further down in the list.

MS COCO As shown in Table 5, on manually annotated caption data, our neural machine translation baseline outperformed the best system of Hitschler et al. (2016) by more than 3 BLEU percentage points, demonstrating the advantages of neural over statistical machine translation. We were able to achieve nominal gains over this very strong baseline using our multimodal reranking setup. We found small but consistent gains in terms of BLEU if textual and visual information is combined (TV, BW, BWV, TVBW). The best system improved over the neural machine translation baseline by 0.62 BLEU points and 0.27 Character F-score points.

We also applied our model-based reranker to the original `cdec`'s hypothesis lists provided by Hitschler et al. (2016). The orthogonal nature of textual and visual components is particularly

	System	baseline	T	B	TB	V	W	TV	BW	BWV	TVBW
BLEU	de-en	35.29	†36.45	†39.20	34.00	35.35	†37.09	†36.53	†**40.90**	†40.76	†39.39
	fr-en	27.81	29.79	†34.70	†**35.27**	27.85	25.09	†29.82	†33.74	†33.62	†29.76
	ru-en	12.85	12.89	†15.08	†**15.74**	12.93	12.04	12.91	†15.51	†15.45	†14.61
ChF	de-en	66.57	67.42	69.01	65.79	66.68	68.50	67.51	**69.01**	68.95	68.27
	fr-en	67.99	69.28	70.81	**71.28**	68.27	65.59	69.50	69.66	69.75	68.24
	ru-en	45.33	46.56	46.63	**47.07**	45.84	44.77	46.40	46.94	47.03	46.41

Table 6: BLEU and Character F-scores (ChF) on Wikimedia test data. Best scoring systems for each language pair were selected on the dev set and are printed in bold. For BLEU, the preceding † indicates a significant improvement ($p < 0.005$) over the baseline system as reported by MultEval's randomization test (Clark et al., 2011).

apparent in the combination of T and V rerankers, where TV showed the biggest improvement of 1.49 BLEU points over the baseline system. All combinations of textual and visual components showed consistent improvements over the best joint system (TSR-CNN) on this dataset.

Wikimedia Commons On Wikimedia Commons, improvements over the neural machine translation baseline were much larger as our retrieval-based reranking approach was very effective in improving translation quality: As shown in Table 6, the improvement were as high as 7.46 BLEU points and 3.29 Character F-score points (French-English). However, text-based retrieval presented a strong baseline on this dataset, which was not always outperformed by additional multimodal retrieval components. Best performance was achieved by a multimodal system only on the German-English data; there was no consistent improvement from incorporating image retrieval data across all three language pairs. It should be noted that irrespective of multi-modality, the best retrieval-based systems always comfortably outperformed the neural machine translation baseline on this dataset. Here, the ensemble of rerankers can make use of hypothesis lists up to the maximum length (see Table 11), as it is able to successfully identify good translations by exploiting information from pivot images and documents.

5.5 Examples

Table 7 shows two examples of source image-caption pairs, the reference translation, and target translations produced by the `nematus` baseline, a pure text-retrieval-based reranking system (TB), and three combinations of text- and image-based reranking components (BW, BWV, TVBW). In both examples, the text-retrieval-based reranking system was not able to select a different translation than the plain MT system. In the left example, additional visual information supports the translation mentioning the "garden", which is a prominent part of the image. In the right example, visual information again helps to select the more complete translation containing the phrase "leaning tower". Notably, the same translation was favored in all four combinations of text- and image-based reranking components (i.e. TV, BW, BWV, TVBW) in the left example, and in the three combinations listed in the table in the right example.

6 Conclusions

We presented a dataset and method for improving caption translation of noisy user-generated data without the need of large parallel captions. Our dataset contains 4M image-caption pairs extracted from Wikimedia Commons, with an average sentence length of 34 tokens and a vocabulary size that is orders of magnitude larger than in previously used caption translation data.

The key idea of our method is to retrieve matches on monolingual target captions, based on textual and visual similarity, and re-score translation hypotheses according to similarity with target matches. This allows us to modularize the translation and the retrieval components of

Image		
Source	Schlossgarten Oldenburg	Kathedrale von Pisa & schiefer Turm von Pisa , Pisa , Italien
`nematus`	oldenburg palace	pisa cathedral of pisa , pisa , italy
Text only `TB`	oldenburg palace	pisa cathedral of pisa , pisa , italy
Text+visual `BW`, `BWV`, `TVBW`	castle garden of oldenburg	pisa cathedral , leaning tower of pisa , pisa , italy .
Reference	oldenburg castle garden	pisa cathedral & leaning tower of pisa , pisa , italy

Table 7: Examples for improved caption translation "in the wild" by multimodal feedback. *Oldenburg Schlossgarten 6.JPG* (left) by *Corradox* and *Pisa Cathedral & Leaning Tower of Pisa.jpg* (right) by *TheVelocity* are licensed under CC BY-SA 4.0.

our system. In practice, this means that if new retrieval data becomes available, the translation system does not need to be re-trained, enabling fast adaptation of the system to new data. This is economically interesting in situations where data is constantly changing and frequent retraining of a system is prohibitive, like in an e-commerce environment.

Our results show the potential benefit of retrieval-based multimodal machine translation in the challenging setting of caption translation on data from Wikimedia Commons. The learning-to-rank setup for optimizing the ensemble of rerankers is able to exploit the orthogonal information from textual and visual retrieval of target images (and captions) and achieves large improvements over strong neural machine translation baselines. We also achieved gains by reranking on manually annotated data from MS COCO. However, the results on the Multi30k dataset emphasize that a large retrieval database is crucial for the performance of the reranking approach and that it especially benefits from more complex and diverse data.

In future work we would like to investigate possibilities to integrate monolingual image-caption as feedback signal in a reinforcement learning setup to neural caption translation (see for example He et al. (2016)). We would also like to further evaluate and enhance our method on other realistic datasets encountered "in the wild" like web-crawled content, product descriptions, and reviews.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments and suggestions. This research was supported in part by DFG grant RI-2221/2-1 "Grounding Statistical Machine Translation in Perception and Action", and by an Amazon Academic Research Award (AARA) "Multimodal Pivots for Low Resource Machine Translation in E-Commerce Localization".

	top-n	T	B	TB	V	W	TV	BW	BWV	TVBW
Character-F	100	56.77	59.31	59.57	58.64	55.79	59.59	59.00	58.77	58.96
	50	56.96	60.50	60.28	58.21	57.25	60.08	59.94	59.78	60.03
	20	57.57	61.10	60.96	58.69	58.11	60.16	60.51	60.39	60.43
	10	58.00	**61.62**	61.39	**59.85**	59.42	60.39	60.63	60.36	60.30
	5	58.54	61.50	**61.60**	59.04	**60.06**	**60.57**	**61.07**	**60.92**	**61.28**

Table 8: Influence of hypothesis lists length n on Character F-score for the reranking experiment on MS COCO dev data where we applied the new reranker on the original `cdec`'s hypothesis lists. Numbers in bold indicate highest score above the baseline (59.64) within a column.

	top-n	T	B	TB	V	W	TV	BW	BWV	TVBW
Character-F	100	49.01	48.26	47.98	48.78	48.02	48.74	48.41	48.29	48.26
	50	49.02	48.75	48.54	48.83	48.29	48.78	48.65	48.50	48.52
	20	49.06	48.86	48.72	49.10	48.62	49.04	48.92	48.85	48.89
	10	49.10	49.14	49.10	49.24	48.78	49.14	49.17	49.02	49.12
	5	49.16	49.29	49.19	49.33	49.11	49.28	49.22	49.18	49.19

Table 9: Hypothesis lists length n and Character F-score on Multi30k dev data. No combination of rerankers was able to surpass the baseline (49.35) on this dataset. Note that increasing the hypothesis lists length always leads to degradation, because the system is unable to identify better translations in the list.

	top-n	T	B	TB	V	W	TV	BW	BWV	TVBW
Character-F	100	64.34	62.24	62.25	64.97	62.42	64.77	62.95	63.15	62.91
	50	64.36	62.89	62.61	65.01	62.78	64.84	63.31	63.66	63.57
	20	64.59	63.80	63.85	**65.14**	63.42	64.93	63.92	63.98	63.97
	10	64.68	64.36	64.32	65.05	64.00	64.93	64.12	64.30	64.22
	5	**64.79**	64.18	64.29	65.11	64.09	**64.94**	64.10	64.25	64.17

Table 10: Hypothesis lists length n and Character F-score on MS COCO dev data. Numbers in bold indicate highest score above the baseline (64.76) within a column.

		top-n	T	B	TB	V	W	TV	BW	BWV	TVBW
Character-F	de-en	100	**68.54**	69.74	66.15	67.73	**69.70**	**68.58**	**70.42**	**70.33**	69.43
		50	68.53	**69.81**	66.52	67.72	69.67	68.58	70.33	70.30	**69.53**
		20	68.50	69.58	66.70	67.73	69.43	68.56	69.86	69.80	69.18
		10	68.40	69.43	66.96	**67.79**	69.12	68.45	69.51	69.47	68.85
		5	68.28	68.97	66.91	67.74	68.72	68.33	69.19	69.19	68.67
Character-F	fr-en	100	68.10	69.41	**69.96**	**67.73**	58.18	68.26	67.91	68.05	66.23
		50	**68.33**	**69.59**	69.87	67.70	62.17	**68.46**	**68.21**	**68.21**	66.96
		20	68.24	69.01	69.25	67.68	63.37	68.43	67.85	67.90	67.12
		10	68.28	68.74	69.04	67.59	63.96	68.46	67.77	67.74	67.22
		5	68.08	68.50	68.67	67.52	64.81	68.23	67.75	67.69	67.25
Character-F	ru-en	100	46.72	46.16	47.24	46.59	40.41	46.82	45.88	46.15	45.87
		50	47.31	46.97	**47.90**	46.60	43.81	47.19	47.24	47.42	47.06
		20	47.28	47.07	47.49	46.58	44.74	47.22	**47.49**	**47.57**	47.05
		10	**47.35**	47.22	47.74	**46.62**	45.41	**47.36**	47.28	47.49	47.48
		5	47.23	**47.26**	47.89	46.59	45.78	47.20	47.26	47.47	**47.50**

Table 11: Influence of hypothesis lists length n on Character F-score on Wikimedia dev data. Numbers in bold indicate highest score above the baseline within a column group. The baseline scores are 67.65, 67.35, and 46.16 for the de-en, fr-en, and ru-en systems, respectively.

References

Calixto, I., Stein, D., Matusov, E., Lohar, P., Castilho, S., and Way, A. (2017). Using images to improve machine-translating e-commerce product listings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Valencia, Spain.

Chen, B. and Cherry, C. (2014). A systematic comparison of smoothing techniques for sentence-level BLEU. In *Proceedings of the Ninth Workshop on Statistical Machine Translation (WMT)*, Baltimore, Maryland, USA.

Clark, J., Dyer, C., Lavie, A., and Smith, N. (2011). Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the Association for Computational Lingustics (ACL)*, Portland, Oregon, USA.

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 Conference on Machine Translation (WMT)*, Copenhagen, Denmark.

Elliott, D., Frank, S., Sima'an, K., and Specia, L. (2016). Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th ACL Workshop on Vision and Language*, Berlin, Germany.

Ferraro, F., Mostafazadeh, N., Huang, T.-H. K., Vanderwende, L., Devlin, J., Galley, M., and Mitchell, M. (2015). A survey of current datasets for vision and language research. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Lisbon, Portugal.

He, D., Xia, Y., Qin, T., Wang, L., Yu, N., Lie, T.-Y., and Ma, W.-Y. (2016). Dual learning for machine translation. In *Advances in Neural Information Processing Systems (NIPS)*, Barcelona, Spain.

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 (ACL)*, 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.

Huang, P.-Y., Liu, F., Shiang, S.-R., Oh, J., and Dyer, C. (2016). Attention-based multimodal neural machine translation. In *Proceedings of the First Conference on Machine Translation (WMT)*, Berlin, Germany.

Lala, C., Madhyastha, P., Wang, J., and Specia, L. (2017). Unraveling the contribution of image captioning and neural machine translation for multimodal machine translation. *The Prague Bulletin of Mathematical Linguistics (PBML)*, (108):197–208.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollar, P. (2014). Microsoft COCO: Common objects in context. arXiv:1405.0312 [cs.CV].

Ordonez, V., Kulkarni, G., and Berg, T. L. (2011). Im2Text: Describing images using 1 million captioned photographs. In *Neural Information Processing Systems (NIPS)*, 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 (ACL)*, Philadelphia, Pennsylvania, USA.

Popović, M. (2015). chrF: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation (WMT)*, Lisbon, Portugal.

Rashtchian, C., Young, P., Hodosh, M., and Hockenmaier, J. (2010). Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Los Angeles, CA.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. S., Berg, A. C., and Li, F. (2014). Imagenet large scale visual recognition challenge. *Computing Research Repository*, abs/1409.0575.

Sennrich, R., Firat, O., Cho, K., Birch, A., Haddow, B., Hitschler, J., Junczys-Dowmunt, M., Läubli, S., Miceli Barone, A. V., Mokry, J., and Nadejde, M. (2017). Nematus: a toolkit for neural machine translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Valencia, Spain.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, Berlin, Germany.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*, San Diego, CA.

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

Specia, L., Frank, S., Sima'an, K., and Elliott, D. (2016). A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation (WMT)*, Berlin, Germany.

Wäschle, K. and Riezler, S. (2015). Integrating a large, monolingual corpus as translation memory into statistical machine translation. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation (EAMT)*, Antalya, Turkey.