NICT-NAIST System for WMT17 Multimodal Translation Task

Jingyi Zhang[1,2], **Masao Utiyama**[1], **Eiichro Sumita**[1]
Graham Neubig[2], **Satoshi Nakamura**[2]
[1]National Institute of Information and Communications Technology,
3-5Hikaridai, Keihanna Science City, Kyoto 619-0289, Japan
[2]Graduate School of Information Science, Nara Institute of Science and Technology,
Takayama, Ikoma, Nara 630-0192, Japan
jingyizhang/mutiyama/eiichiro.sumita@nict.go.jp
neubig/s-nakamura@is.naist.jp

Abstract

This paper describes the NICT-NAIST system for the WMT 2017 shared multimodal machine translation task for both language pairs, English-to-German and English-to-French. We built a hierarchical phrase-based (Hiero) translation system and trained an attentional encoder-decoder neural machine translation (NMT) model to rerank the n-best output of the Hiero system, which obtained significant gains over both the Hiero system and NMT decoding alone. We also present a multimodal NMT model that integrates the target language descriptions of images that are similar to the image described by the source sentence as additional inputs of the neural networks to help the translation of the source sentence. We give detailed analysis for the results of the multimodal NMT model. Our system obtained the first place for the English-to-French task according to human evaluation.

1 Introduction

We participated in the WMT 2017 shared multimodal machine translation task 1, which translates a source language description of an image into a target language description. We built systems for both English-to-German and English-to-French language pairs.

Our baseline systems only use text information. We compared three text-only approaches: a hierarchical phrase-based (Hiero) translation system (Chiang, 2005), an attentional encoder-decoder neural machine translation (NMT) system (Bahdanau et al., 2015), and a system using the NMT model to rerank the n-best output of the Hiero system.

We also explored ways to improve the NMT model with image information. Compared to previous multimodal NMT (MNMT) models that integrate visual features directly (Caglayan et al., 2016; Calixto et al., 2016; Huang et al., 2016; Calixto et al., 2017), we first exploit image retrieval methods to obtain images that are similar to the image described by the source sentence, and then integrate the target language descriptions of these similar images into the NMT model to help the translation of the source sentence. This makes it possible to exploit a large corpus with only images and target language descriptions through an image retrieval step. This is similar to Hitschler et al. (2016)'s multimodal pivots method, which uses target descriptions of similar images for reranking MT outputs, while we use these target descriptions as additional inputs for the NMT model.

2 Text-only MT

We compared three text-only approaches for this translation task.

2.1 Hierarchical Phrase-based SMT

The hierarchical phrase-based model (Chiang, 2005) extracts hierarchical phrase-based translation rules from parallel sentence pairs with word alignments. The word alignments can be learned by IBM models. Each translation rule contains several feature scores. The decoder of hierarchical phrase-based model implements a bottom-up CKY+ algorithm. The weights for different features can be tuned on the development set.

2.2 Attentional NMT

The attentional encoder-decoder networks (Bahdanau et al., 2015) include three parts: an encoder that uses a bi-directional recurrent neural network to learn representations for words in the source

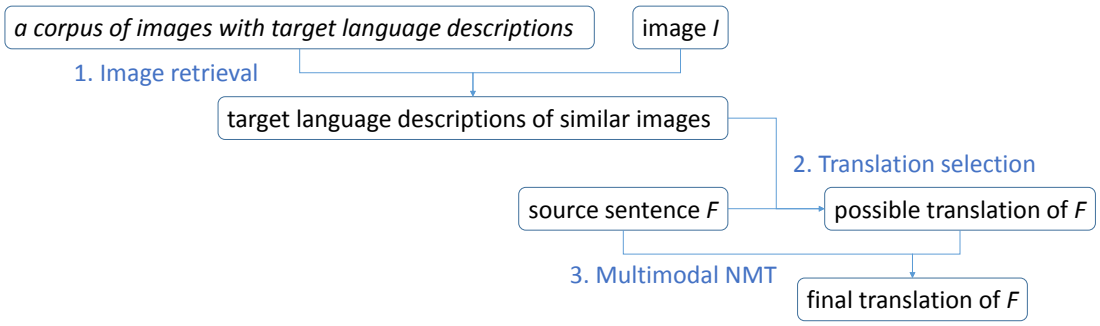

Figure 1: A overview of our multimodal method.

sentence, a decoder that generates the target sentence from left to right and an alignment model that learns which parts of the source sentence to focus on when the decoder generates each target word.

2.3 SMT reranked by NMT

The hierarchical phrase-based SMT model generates a n-best list for each source sentence. We use the attentional NMT model to assign a score to each output in the n-best list. This new NMT score together with the original SMT features is used to rerank the n-best list. The weight of the new NMT score is tuned together with other feature weights on the n-best lists of the development set.

3 Our Multimodal Approach

We propose a method to integrate the visual information into the NMT model.

Originally, the encoder of the NMT model only encodes the information of source sentence F. Our method integrates the visual information of image I into the encoder. Figure 1 is a overview of our multimodal method, which contains three steps.

Image retrieval Given image I, we search the 100 most similar images $\mathcal{I}$ from the training set and get the target language descriptions of these similar images as possible descriptions of I. When calculating image similarity, we used the Euclidean distance between averaged pooled feature vectors provided by the organizers.

Translation selection We select the most probable target word e for each source word f in sentence F as follows:

$$score\left(e, f, I\right) = score\left(e, f\right) + \lambda \cdot score\left(e, I\right). \quad (1)$$

Here $score\left(e, f\right)$ measures the probability of f being translated into e as follows:

$$score\left(e, f\right) = \frac{align\left(e, f\right)}{\sum_{e' \in V} align\left(e', f\right)}, \quad (2)$$

where $align\left(e, f\right)$ is how many times f and e are aligned in the word-aligned training set.[1] $score\left(e, I\right)$ measures how related e and I are as follows:[2]

$$score\left(e, I\right) = idf\left(e\right) \cdot \sum_{I' \in \mathcal{I}} \frac{is\_in\left(e, I'\right)}{dis\left(I, I'\right)}, \quad (3)$$

where $idf\left(e\right)$ is the inverse document frequency of e to punish high-frequency words, $dis\left(I, I'\right)$ is the Euclidean distance between I and I', $is\_in\left(e, I'\right)$ is 1/0 when e is/isn't contained in the description of I'. λ is the weight that can be tuned on the development set.

Multimodal NMT The original NMT model projects each source word f into a vector. We add an additional embedding matrix to project the selected target word e for f into a new vector. Then we concatenate both vectors and use them as the input for the bi-directional recurrent neural network of the NMT encoder.

4 Experiments

4.1 Text-only systems

We use training, development and test sets provided by the organizers (Elliott et al., 2016; Elliott

[1] When counting the alignment $align\left(e, f\right)$, we only use the intersection of the bi-directional GIZA++ alignments, so the alignments are more reliable.

[2] For e that does not occur in $\mathcal{I}$, $score\left(e, I\right)$ is 0.

	Description	Distance
Query 1	**a group of men are loading cotton onto a truck**	
Results	a baby camel going towards a woman , while a man takes a picture .	35.43
	person sitting in a chair selling goods outside of a building .	36.04
Query 2	**a man sleeping in a green room on a couch .**	
Results	a baby and three cats are resting on a bed .	29.85
	a man in a white t-shirt and beige shorts lies asleep on a black sofa .	29.86
Query 3	**a boy wearing headphones sits on a woman 's shoulders .**	
Results	a young blond girl in pink shirt and pigtails is sitting atop a man 's shoulders in a crowd .	28.88
	a man dressed in blue is juggling in front of an audience .	29.10
Query 4	**two men setting up a blue ice fishing hut on an iced over lake**	
Results	a man is drilling through the frozen ice of a pond .	23.49
	an inline skater in red pants and blue shirt skates between green cones .	30.05
Query 5	**a balding man wearing a red life jacket is sitting in a small boat .**	
Results	man with shawl praying by a large lake and small boat .	31.53
	four people standing on a raft sailing away on the water .	31.54

Table 1: Image retrieval examples (two most similar images for each query image). Description is the English descriptions for query and result images. Distance is the Euclidean distance between image vectors.

	Flickr		COCO	
Method	en-de	en-fr	en-de	en-fr
Hiero	27.86	50.38	24.57	41.88
NMT	30.52	50.46	24.27	41.26
Reranking	31.98	55.25	28.05	45.17

Table 2: Results of text-only approaches (BLEU).

	en-de	en-fr
$\lambda = 0$	52.17	65.60
$\lambda = 0.2$	52.93	66.31

Table 3: 1-gram BLEU score of selected target words on the development set.

et al., 2017). We lowercase, normalise punctuation and tokenise all sentences. The Hiero translation system was based on Moses (Koehn et al., 2007). We used GIZA++ (Och and Ney, 2003) and *grow-diag-final-and* heuristic (Koehn et al., 2003) to obtain symmetric word alignments. For decoding, we used standard features: direct/inverse phrase translation probability, direct/inverse lexical translation probability and a 5-gram language model, which was trained on the target side of the training corpus by IRSTLM Toolkit[3] with improved Kneser-Ney smoothing.

Attentional encoder-decoder networks were trained with Lamtram[4]. Word embedding size and hidden layer size are both 512. Training data was reshuffled between epochs. Validation was done after each epoch. We used the Adam optimization algorithm (Kingma and Ba, 2014). Because the training set is only 29K sentence pairs, we used dropout (0.5) and a small learning rate (0.0001) to reduce overfitting, which yielded improvements of $3 - 4$ BLEU on the development set. For training the NMT model, we replace words that occur less than twice in the training set as UNK. When decoding, we find the most probable source word for each UNK and replace the UNK using a lexicon extracted from the word-aligned training set.

We used the NMT model to rerank the unique $10,000$-best output of the Hiero system. The NMT score was used as an additional feature for the Hiero system. Feature weights were tuned by MERT (Och, 2003).

Table 2 shows results of the Hiero system, the NMT system and using the NMT model to rerank the Hiero outputs. The reranking system had the best performance on both language pairs. It is straightforward that using the NMT feature to rerank the Hiero outputs can achieve improvements over the pure Hiero system. The reason why the reranking method outperformed the NMT system should be that the training corpus is relatively small and the NMT system did not outperform the Hiero system largely. Therefore, the reranking method that takes advantages of both the Hiero and NMT systems worked the best on this task.

4.2 Multimodal systems

For the multimodal method, we found when $\lambda = 0.2$, the selected target words for the development

[3]http://hlt.fbk.eu/en/irstlm

[4]https://github.com/neubig/lamtram

	Flickr		COCO	
Method	en-de	en-fr	en-de	en-fr
NMT	30.52	50.46	24.27	41.26
MNMT	29.56	49.83	23.60	40.77

Table 4: Comparison of the NMT model and the MNMT model (BLEU).

		BLEU	Meteor	TER
en-de	Our system	31.9	53.9	48.1
	Best system	33.4	54.0	48.5
en-fr	Our system	55.3	72.0	28.4
	Best system	55.9	72.1	28.4

Table 5: Official evaluation results on the 2017 Flickr test sets.

set had the highest 1-gram BLEU score[5] for both language pairs as shown in Table 3, which shows that visual features did help to select more accurate translations than using only translation probabilities in the translation selection step.

However, on both language pairs, our multimodal NMT model did not improve, but decreased the test set BLEU score compared to the baseline NMT model as shown in Table 4. And using the multimodal NMT as an additional feature for reranking the Hiero system did not further improve the Hiero system that had integrated the text-only NMT model. Table 5 and 6 show the official evaluation results of our system and the best system for the multimodal task (with METEOR as the primary metric). Our system is very competitive, especially with METEOR, even though only text features helped in our system, which shows with finely tuned parameters, the text-only approach that uses the NMT model to rerank the output of the Hiero system can give a strong result for this task. In addition, our system obtained the first place for the English-to-French task according to human evaluation (Elliott et al., 2017).

To further analyze the results of our multimodal method, we give some output examples for each step in Figure 1.

Table 1 gives some image retrieval results. As we can see, in the descriptions of the retrieved images, there is a lot of noise that is not useful for helping the translation of the source sentence, which is why we used 100 images with the high-

[5]Because the selected target words are not reordered, so we only calculate 1-gram BLEU score.

		BLEU	Meteor	TER
en-de	Our system	28.1	48.5	52.9
	Best system	28.7	48.9	52.5
en-fr	Our system	45.1	65.6	34.7
	Best system	45.9	65.9	34.2

Table 6: Official evaluation results on the 2017 COCO test sets.

est similarities and the translation selection step to select useful information for our multimodal NMT model. Note that we used the target language (German or French) descriptions of similar images in our method, but Table 1 shows the source language (English) descriptions for easy understanding. In addition, for this image retrieval step, a large image corpus can be helpful to find more similar images and only target descriptions are needed for this image corpus.

Table 7 shows some examples for our multimodal method. For the first two examples, the visual information helped to improve the translations. In Example 1, "running" is translated into "rennt" by the NMT model incorrectly. The translation selection step selected the correct translation "läuft" for "running" and helped the MNMT model translate it correctly. In Example 2, "home" should be translated into "hauses", but it is missing in the NMT translation. The translation selection step selected "haus" as the translation for "home", which then appeared in the translation of the MNMT model.

However, for the last two examples in Table 7, the additional target descriptions decreased the translation quality. In Example 3, "looking" was correctly translated into "blickt" by the NMT model. But "schaut" was selected as the translation of "looking" at the translation selection step, which led the MNMT model translated it incorrectly. In Example 4, "flying" was correctly translated into "fliegenden" by the NMT model. But "fliegt" was selected as the translation of "flying" by the translation selection step, which led to "flying" being missing in the MNMT translation. Here, "fliegenden" and "fliegt" are different forms of the German word "fliegen", which are very difficult to distinguish using visual information. Using only the original form for these selected target words can be helpful to solve this problem.

As shown in Table 7, the target descriptions used as additional inputs for the multimodal NMT

Example 1	
Src	an adult australian shepherd follows behind a **running** australian shepherd puppy .
Ref	ein ausgewachsener australian shepherd folgt einem welpen , der vor ihm **läuft** .
NMT	ein erwachsener australischer fängt hinter einem **rennt** australischer .
TS	ein erwachsener australischer schäferhund folgt hinter ein **läuft** australischer schferhund welpe .
MNMT	ein erwachsener australischer schäferhund folgt einem **läuft** australischer hund .
Example 2	
Src	woman and child outside the front door of their scenic **home** .
Ref	eine frau und ein kind vor der tür ihres idyllischen **hauses** .
NMT	eine frau und ein kind vor der tür des malerische .
TS	frau und kind freien der vor tür von ihren malerische **haus** .
MNMT	eine frau und ein kind vor der tür eines malerische **haus** .
Example 3	
Src	a little girl is **looking** through a telescope at the beach .
Ref	ein kleines mädchen **blickt** durch ein teleskop auf den strand .
NMT	ein kleines mädchen **blickt** durch ein teleskop am strand .
TS	einem kleines mädchen ist **schaut** durch einem teleskop auf der strand .
MNMT	ein kleines mädchen **schaut** durch ein teleskop am strand .
Example 4	
Src	a dog turns on the grass to persue a **flying** ball .
Ref	ein hund dreht sich auf dem gras um einem **fliegenden** ball nachzulaufen .
NMT	ein hund dreht sich auf dem gras , um einen **fliegenden** ball zu persue .
TS	ein hund dreht auf der gras zu persue ein **fliegt** ball .
MNMT	ein hund dreht sich auf dem gras , um den ball zu persue .

Table 7: Translation examples. NMT: the translation by the NMT model; TS: the selected words for each source word in the translation selection step; MNMT: the translation by the MNMT model.

model helped the translation for some cases, but also introduced new noise, which hurt the translation performance in some other cases. In future work, we will work on how to use these target description information more effectively.

5 Conclusion

We described our system for the WMT17 multimodal translation task, including text-only approaches and a multimodal method that first searches for some possible target language descriptions of the image and then integrates these target descriptions into the NMT model to help the translation of the source sentence. Results show the text-only approach that uses a NMT model to rerank the output of a Hiero system gave a strong result for this task and the MNMT model did not further improve the text-only system, but the target descriptions did contain some useful information that can help the translations. In future work, we will work on how to make use of these related target descriptions more effectively. In addition, a larger corpus of images with only target language descriptions can be useful for our method to obtain more accurate target descriptions.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*.

Ozan Caglayan, Walid Aransa, Yaxing Wang, Marc Masana, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, and Joost van de Weijer. 2016. Does multimodality help human and machine for translation and image captioning? In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 627–633. http://www.aclweb.org/anthology/W16-2358.

Iacer Calixto, Desmond Elliott, and Stella Frank. 2016. Dcu-uva multimodal mt system report. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 634–638. http://www.aclweb.org/anthology/W16-2359.

Iacer Calixto, Daniel Stein, Evgeny Matusov, Pintu Lohar, Sheila Castilho, and Andy Way. 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: Volume 2, Short Papers*. Association for Computational Linguistics, Valencia, Spain, pages 637–643. http://www.aclweb.org/anthology/E17-2101.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting*

of the Association for Computational Linguistics (ACL'05). Association for Computational Linguistics, Ann Arbor, Michigan, pages 263–270. https://doi.org/10.3115/1219840.1219873.

D. Elliott, S. Frank, K. Sima'an, and L. Specia. 2016. Multi30k: Multilingual english-german image descriptions pages 70–74.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the Second Shared Task on Multimodal Machine Translation and Multilingual Image Description. In *Proceedings of the Second Conference on Machine Translation*. Copenhagen, Denmark.

Julian Hitschler, Shigehiko Schamoni, and Stefan Riezler. 2016. Multimodal pivots for image caption translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 2399–2409. http://www.aclweb.org/anthology/P16-1227.

Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. 2016. Attention-based multimodal neural machine translation. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 639–645. http://www.aclweb.org/anthology/W16-2360.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. pages 177–180. http://www.aclweb.org/anthology/P07-2045.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. pages 48–54.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. pages 160–167. https://doi.org/10.3115/1075096.1075117.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics* 29(1):19–51.