

# Idiap NMT System for WAT 2019 Multi-Modal Translation Task

Shantipriya Parida

Idiap Research Institute,  
Rue Marconi 19,  
1920 Martigny,  
Switzerland

firstname.lastname@idiap.ch

Petr Motlíček

Charles University,  
Faculty of Mathematics and Physics,  
Institute of Formal and Applied Linguistics,  
Malostranské náměstí 25, 118 00,  
Prague, Czech Republic  
bojar@ufal.mff.cuni.cz

Ondřej Bojar\*

## Abstract

This paper describes the Idiap submission to WAT 2019 for the English-Hindi Multi-Modal Translation Task. We have used the state-of-the-art Transformer model and utilized the IITB English-Hindi parallel corpus as an additional data source. Among the different tracks of the multi-modal task, we have participated in the “Text-Only” track for the evaluation and challenge test sets. Our submission tops in its track among the competitors in terms of both automatic and manual evaluation. Based on automatic scores, our text-only submission also outperforms systems that consider visual information in the “multi-modal translation” task.

## 1 Introduction

In recent years, significant research has been done to address problems that require joint modelling of language and vision (Specia et al., 2016). The popular applications involving Natural Language Processing (NLP) and Computer Vision (CV) include image description generation (Bernardi et al., 2016), video captioning (Li et al., 2019), or visual question answering (Antol et al., 2015).

In the past few decades, multi-modality has received critical attention in translation studies, although the benefit of visual modality in machine translation is still in debate (Caglayan et al., 2019). The main motivation in multi-modal research in machine translation is the intuition that information from other modalities could help to find the correct sense of ambiguous words in the source sentence, which could potentially lead to more accurate translations (Lala and Specia, 2018).

\* Corresponding author

Set	Sentences	Tokens	
		English	Hindi
HVG Train	28932	143178	136722
IITB Train	1.4 M	20.6 M	22.1 M
D-Test	998	4922	4695
E-Test	1595	7852	7535
C-Test	1400	8185	8665

Table 1: Statistics of our data: the number of sentences and tokens.

Despite the lack of multi-modal datasets, there is a visible interest in using image features even for machine translation for low-resource language. For instance, Chowdhury et al. (2018) train a multi-modal neural MT system for Hindi→English using synthetic parallel data only.

In this system description paper, we explain how we used additional resources in the text-only track of WAT 2019 Multi-Modal Translation Task. Section 2 describes the datasets used in our experiment. Section 3 presents the model and experimental setups used in our approach. Section 4 provides the official evaluation results of WAT 2019 followed by the conclusion in Section 6.

## 2 Dataset

The official training set was provided by the task organizers: Hindi Visual Genome (HVG for short, Parida et al., 2019a,b). The training part consists of 29k English and Hindi short captions of rectangular areas in photos of various scenes and it is complemented by three test sets: development (D-Test), evaluation (E-Test) and challenge test set (C-Test). We did not make any use of the images. Our WAT submissions were for E-Test (denoted “EV” in WAT official tables) and C-Test (denoted “CH” in WAT tables).

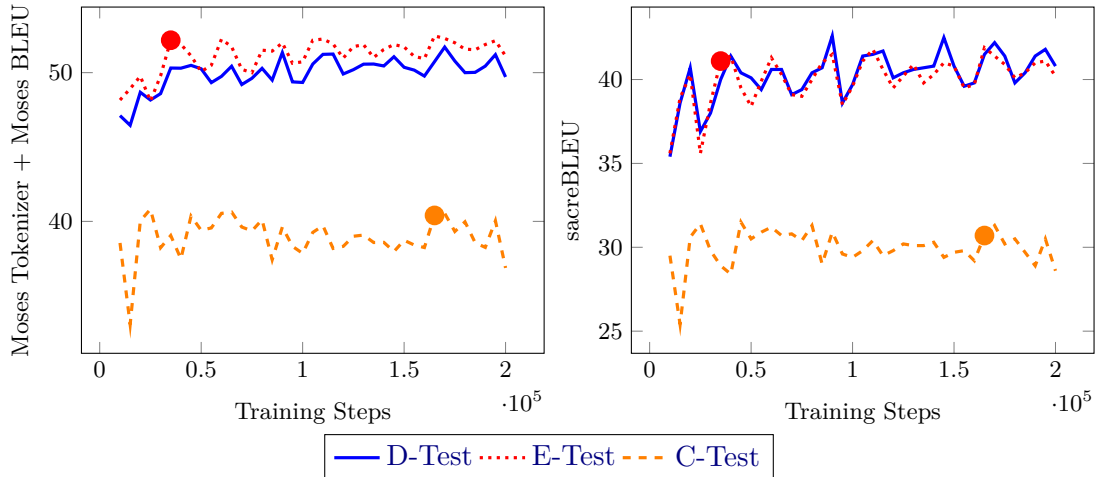


Figure 1: Learning curves in terms of BLEU score. The left plot is based on Moses tokenizer and BLEU score as implemented in Moses scorer. The right plot is sacreBLEU. The big round dots indicate which training iteration was used when producing our final submissions to WAT manual and automatic evaluation for E-Test and C-Test.

Additionally, we used the IITB Corpus (Kunchukuttan et al., 2017) which is supposedly the largest publicly available English-Hindi parallel corpus. This corpus contains 1.49 million parallel segments and it was found very effective for English-Hindi translation (Parida and Bojar, 2018).

The statistics of the datasets are shown in Table 1.

### 3 Experiments

We focussed only on the text translation task.

We used the Transformer model (Vaswani et al., 2018) as implemented in OpenNMT-py (Klein et al., 2017).<sup>1</sup>

#### 3.1 Tokenization and Vocabulary

Subword units were constructed using the word pieces algorithm (Johnson et al., 2017). Tokenization is handled automatically as part of the pre-processing pipeline of word pieces.

We generated the vocabulary of 32k subword types jointly for both the source and target languages. The vocabulary is shared between the encoder and decoder.

#### 3.2 Training

To train the model, we used a single GPU and followed the standard “Noam” learning rate

decay,<sup>2</sup> see Vaswani et al. (2017) or Popel and Bojar (2018) for more details. Our starting learning rate was 0.2 and we used 8000 warm up steps.

We ran only one training run.

We concatenated HVG and IITB training data and shuffled it at the level of sentences.

We let the model train for up to 200K steps, interrupted a few times due to GPU queuing limitations of our cluster. Following the recommendation of Popel and Bojar (2018), we present the full learning curves on D-Test, E-Test and C-Test in Figure 1.

We observed a huge difference between BLEU (Papineni et al., 2002) scores as implemented in the Moses toolkit (Koehn et al., 2007) and the newer implementation in sacreBLEU (Post, 2018). The discrepancy is very likely caused by different tokenization but the best choice in terms of linguistic plausibility still has to be made. In Figure 1, we show both implementations and see that the Moses implementation gives scores higher by 10 (!) points absolute. More importantly, it is a little less peaked, which we see as evidence for better robustness and thus hopefully the linguistic adequacy.

All of the test sets (D-, E- and C-Test) are independent of the training data and the training itself is not affected by them in any way.

<sup>1</sup><http://opennmt.net/OpenNMT-py/quickstart.html>

<sup>2</sup><https://nvidia.github.io/OpenSeq2Seq/html/api-docs/optimizers.html>

System and WAT Task Label	WAT BLEU	Our sacreBLEU	Our Moses BLEU	WAT Human
Our MMEV <b>TEXT</b> en-hi	<b>41.32</b>	41.1	52.18	<b>72.84</b>
Best competitor in MMEV <b>MM</b> en-hi	40.55	–	–	69.17
Our MMCH <b>TEXT</b> en-hi	<b>30.94</b>	30.7	40.40	<b>59.81</b>
Best competitor in MMCH <b>MM</b> en-hi	20.37	–	–	54.50

Table 2: WAT 2019 official automatic and manual evaluation results for English→Hindi (HINDEN) tasks on the E-Test (EV, upper part) and C-Test (CH, lower part), complemented with our automatic scores. Our scores are from the “TEXT”, i.e. text-only, track while the “Best competitor” lines are from the “MM” (multi-modal) track. On each test set, the automatic scores are comparable, because the set of reference translations is identical for the two tracks. The manual scores are comparable to a lower extent because the text-only and multi-modal tracks were manually evaluated in two separate batches.

In other words, they all can be seen as interchangeable, only the choice which particular iteration to run must be done on one of them and evaluated on a different one.

At the submission deadline for E-Test, our training has only started, so we submitted the latest result available, namely E-Test translated with the model at 35K training steps. When submitting the translations of C-Test for the WAT official evaluation, we already knew the full training run and selected the step 165K where E-Test reached its maximum score. In other words, the choice of the model for the C-Test was based on E-Test serving as a validation set.

## 4 Official Results

We report the official automatic as well as manual evaluation results of our models for the evaluation and challenge test dataset here in Table 2. All the scores are available on the WAT 2019 website<sup>3</sup> and in the WAT overview paper (Nakazawa et al., 2019).

According to both automatic and manual scores, our submissions were the best in the text-only task (MM\*\*TEXT), see the tables in Nakazawa et al. (2019).

Since the text-only and multi-modal tracks differ only in the fact whether the image is available and the underlying set of sentences is identical, we can also compare our result with the scores of systems participating in the multi-modal track (MM\*\*MM). We show only the best system of the multi-modal track. Both on the E-Test and C-Test, our (text-only) candidates scored better in BLEU than the best competitor in the multi-modal track (41.32 vs. 40.55 on E-Test and 30.94 vs. 20.37

<sup>3</sup><http://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/>

on C-Test). Manual judgments also indicate that our translations are better than those of the best multi-modal system, but here the comparison has to be taken with a grain of salt. The root of the trouble is that the manual evaluation for the text-only and multi-modal tracks ran separately. While the underlying method (Direct Assessment, DA, Graham et al., 2013) in principle scores sentences in absolute terms, it has been observed by Bojar et al. (2017) that DA scores from independent runs are not reliably comparable. We indicate this by the additional horizontal lines in Table 2.

Figure 2 illustrates of our translation output.

## 5 Discussion

We did not explore the space of possible configurations much, we just ran training and observed the development of the learning curve. Our final results are nevertheless good, indicating that reasonably clean data and baseline settings of the Transformer architecture deliver good translations.

The specifics of the task have to be taken into account. The “sentences” in Hindi Visual Genome are quite short, only 4.7 Hindi and 4.9 English tokens per sentence. This is substantially less than the IITB corpus where the average number of tokens is 15.8 (Hindi) and 14.7 (English). With IITB mixed in the training data, the model gets a significant advantage, not only because of the better coverage of words and phrases but also due to the length. As observed by Kocmi and Bojar (2017) and Popel and Bojar (2018), NMT models struggle to produce outputs longer than the training data was. Our situation is the reverse, so our model “operates within its comfortable zone”.





	<p>English Input: gold religious <b>cross</b> on top of golden ball</p> <p>Translated Output: सोने की गेंद के शीर्ष पर स्वर्ण धार्मिक क्रॉसैं .</p> <p>Gloss: Gold religious cross on top of golden ball</p>
	<p>English Input: a blue wall beside tennis <b>court</b></p> <p>Translated Output: टेनिस कोर्ट के पास एक नीली दीवार हैं ।</p> <p>Gloss: Blue wall near the tennis court</p>
	<p>English Input: the tennis <b>court</b> is made up of sand and dirt</p> <p>Translated Output: टेनिस कोर्ट रेत और गंदगी से बनी है।</p> <p>Gloss: Tennis court is made of sand and dirt</p>
	<p>English Input: A crack on the <b>court</b></p> <p>Translated Output: <u>अदालत</u> पर एक crack</p> <p>Gloss: A crack on the <u>judicial court</u></p>

Figure 2: Sample Hindi output as generated for the challenge test set. The ambiguous source word is bolded in the English input, errors are underlined in the MT output and the gloss. The associated source images are given for the reference purpose only to judge our NMT system translation quality, we have not used any image features in our experiment.

Comparing the scores of D- and E-Test on the one hand and C-Test on the other hand, we see that D- and E-Test are much easier for the system. This can be attributed to the identical distributional properties of D-Test and E-Test as the model observed for HVG in the training data. According to Parida et al. (2019a), C-Test also comes from the Visual Genome but the sampling is different, each sentence illustrating one of 19 particularly ambiguous words (*focus* words in the following).

As shown in Figure 2, our system has generally no trouble in figuring out the correct sense of the focus words, thanks to the surrounding words in the context. The BLEU scores on C-Test are nevertheless much lower than on E-Test or D-Test. We attribute this primarily to the slight mismatch between HVG training data and C-Test. As can be confirmed in Table 1, the average sentence length in C-Test is 6.2 (Hindi) and 5.8 (English) tokens, i.e. 0.9–1.5 longer than the training data. Indeed, the model produces shorter outputs than expected and BLEU brevity penalty affects C-Test more (BP=0.907) than E-Test (BP=0.974).

By a quick visual inspection of the outputs, we notice that some rare words were not translated at all, for example, “dugout”, “skiing”, or “celtic”. Most of the non-translated words are not the focus words of the challenge test set but simply random words in the sentences. The focus words that were not translated include: “springs”, “cross” and some instance of the word “stand”. We did not have the human capacity to review the translations of all the focus words but our general impression is that they were mostly correct. One example, the mistranslation of the (tennis) court is given at the bottom of Figure 2.

Finally, we would like to return to the issue of BLEU implementation pointed out in Section 3.2. The main message to take from this observation is that many common tools are not really polished and well tested for use on less-researched languages and languages not using Latin script. No conclusions can be thus drawn by comparing *numbers* reported across papers. A solid comparison can be only made with the evaluation tool fixed, as is the practice of WAT shared task.

## 6 Conclusion and Future Plans

In this system description paper, we presented our English→Hindi NMT system. We have highlighted the benefits of using additional text-only training data. Our system performed best among the competitors for the submitted track (“text-only”) and also performs better than systems that did consider the image in the “multi-modal” track according to automatic evaluation. We conclude that for the general performance, more parallel data are more important than the visual features available in the image. A targeted manual evaluation would be however necessary to see if the translation of the particularly ambiguous words is better when MT systems consider the image.

As the next step, we plan to utilize image features and carry out a comparison study with the current setup. Also, we plan to experiment with the image captioning variant of the task.

## Acknowledgments

At Idiap, the work was supported by an innovation project (under an InnoSuisse grant) oriented to improve the automatic speech recognition and natural language understanding technologies for German. Title: “SM2:

Extracting Semantic Meaning from Spoken Material” funding application no. 29814.1 IP-ICT. And also supported by the EU H2020 project “Real-time network, text, and speaker analytics for combating organized crime” (ROXANNE), grant agreement: 833635.

At Charles University, the work was supported by the grants 19-26934X (NEUREM3) of the Czech Science Foundation and “Progress” Q18+Q48 of Charles University, and using language resources distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (projects LM2015071 and OP VVV VI CZ.02.1.01/0.0/0.0/16013/0001781).

## References

Stanislaw Antol, Aishwarya Agrawal, Ji-  
asen Lu, Margaret Mitchell, Dhruv Batra,  
C Lawrence Zitnick, and Devi Parikh. 2015.  
Vqa: Visual question answering. In *Proceed-*

*ings of the IEEE international conference on  
computer vision*, pages 2425–2433.

Raffaella Bernardi, Ruket Cakici, Desmond El-  
liott, Aykut Erdem, Erkut Erdem, Nazli Iki-  
zler-Cinbis, Frank Keller, Adrian Muscat, and Bar-  
bara Plank. 2016. Automatic description gener-  
ation from images: A survey of models, datasets,  
and evaluation measures. *Journal of Artificial  
Intelligence Research*, 55:409–442.

Ondřej Bojar, Jindřich Helcl, Tom Kocmi, Jindřich  
Libovický, and Tomáš Musil. 2017. Results  
of the WMT17 Neural MT Training Task. In  
*Proceedings of the Second Conference on Ma-  
chine Translation*, Copenhagen, Denmark. As-  
sociation for Computational Linguistics.

Ozan Caglayan, Pranava Madhyastha, Lucia Spe-  
cia, and Loïc Barrault. 2019. Probing the  
Need for Visual Context in Multimodal Machine  
Translation. *arXiv preprint arXiv:1903.08678*.

Koel Dutta Chowdhury, Mohammed Hasanuzza-  
man, and Qun Liu. 2018. Multimodal neural  
machine translation for low-resource language  
pairs using synthetic data. In *Proceedings of  
the Workshop on Deep Learning Approaches for  
Low-Resource NLP*, pages 33–42.

Yvette Graham, Timothy Baldwin, Alistair Mof-  
fat, and Justin Zobel. 2013. [Continuous Measurement Scales in Human Evaluation of Machine Translation](#). In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le,  
Maxim Krikun, Yonghui Wu, Zhifeng Chen,  
Nikhil Thorat, Fernanda Viégas, Martin Wat-  
tenberg, Greg Corrado, Macduff Hughes, and  
Jeffrey Dean. 2017. [Google’s multilingual neural machine translation system: Enabling zero-shot translation](#). *Transactions of the Association for Computational Linguistics*, 5:339–351.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean  
Senellart, and Alexander M. Rush. 2017. [Open-NMT: Open-source toolkit for neural machine translation](#). In *Proc. ACL*.

Tom Kocmi and Ondřej Bojar. 2017. Curriculum  
Learning and Minibatch Bucketing in Neural  
Machine Translation. In *Proceedings of Recent  
Advances in NLP (RANLP 2017)*.

Philipp Koehn, Hieu Hoang, Alexandra Birch,  
Chris Callison-Burch, Marcello Federico, Nicola  
Bertoldi, Brooke Cowan, Wade Shen, Christine  
Moran, Richard Zens, Chris Dyer, Ondřej Bojar,  
Alexandra Constantin, and Evan Herbst. 2007.  
[Moses: Open Source Toolkit for Statistical Machine Translation](#). In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for*

- Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.
- Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2017. The IIT Bombay English-Hindi Parallel Corpus. *arXiv preprint arXiv:1710.02855*.
- Chiraag Lala and Lucia Specia. 2018. Multimodal lexical translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.
- Sheng Li, Zhiqiang Tao, Kang Li, and Yun Fu. 2019. Visual to text: Survey of image and video captioning. *IEEE Transactions on Emerging Topics in Computational Intelligence*.
- Toshiaki Nakazawa, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Nobushige Doi, Yusuke Oda, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, and Sadao Kurohashi. 2019. Overview of the 6th workshop on Asian translation. In *Proceedings of the 6th Workshop on Asian Translation*, Hong Kong. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *ACL 2002, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania.
- Shantipriya Parida and Ondřej Bojar. 2018. Translating short segments with nmt: A case study in english-to-hindi. In *21st Annual Conference of the European Association for Machine Translation*, page 229.
- Shantipriya Parida, Ondřej Bojar, and Satya Ranjan Dash. 2019a. Hindi Visual Genome: A Dataset for Multimodal English-to-Hindi Machine Translation. *Computación y Sistemas*. In print. Presented at CICLing 2019, La Rochelle, France.
- Shantipriya Parida, Ondřej Bojar, and Satya Ranjan Dash. 2019b. Hindi Visual Genome: A Dataset for Multimodal English-to-Hindi Machine Translation. *arXiv preprint arXiv:1907.08948*.
- Martin Popel and Ondřej Bojar. 2018. Training tips for the transformer model. *The Prague Bulletin of Mathematical Linguistics*, 110(1):43–70.
- Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.
- Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 543–553.
- Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. **Tensor2tensor for neural machine translation**. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 193–199. Association for Machine Translation in the Americas.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.