ODIANLP's Participation in WAT2020

Shantipriya Parida and **Petr Motlíček**
Idiap Research Institute, Switzerland
{firstname.lastname}@idiap.ch

Amulya Ratna Dash
BITS, Pilani, India
amulyaratnadash@gmail.com

Satya Ranjan Dash and **Debasish Kumar Mallick**
KIIT University, India
{sdashfca,1939014}@kiit.ac.in

Satya Prakash Biswal
University Of Chicago, USA
satyapb2002@gmail.com

Priyanka Pattnaik
COE AI LAB, India
priyankapattanaik2013@gmail.com

Biranchi Narayan Nayak
Vettons, Malaysia
biranchi125@gmail.com

Ondřej Bojar
Charles University, MFF, ÚFAL
bojar@ufal.mff.cuni.cz

Abstract

This paper describes the team ("ODIANLP")'s submission to WAT 2020. We have participated in the English→Hindi Multimodal task and Indic task. We have used the state-of-the-art *Transformer* model for the translation task and *InceptionResNetV2* for the Hindi Image Captioning task. Our submission tops in English→Hindi Multimodal task in its track and Odia↔English translation tasks. Also, our submissions performed well in the Indic Multilingual tasks.

1 Introduction

Although machine translation (MT) has proven very successful for many high resource languages, it is still challenging for low resource languages and translation effectively utilizing other modalities (e.g. image; Parida et al., 2020a, 2019b).

Workshop on Asian Translation (WAT)[1] is a recurring event focusing on the Asian languages (Nakazawa et al., 2020) since 2013. This year, WAT introduced the translation task for one of the low resource Indian languages, Odia.[2] Odia is nowadays spoken by 50 million speakers. It is heavily influenced by the Dravidian languages as well as Arabic, Persian, and English. Odia's inflectional morphology is rich with a three-tier tense system. The prototypical word order is subject-object-verb (SOV) (Parida et al., 2020a,b).

In this system description paper, we explain our approach for the participated tasks. Section 2 describes the datasets used in our experiment. Section 3 presents the model and experimental setups used in our approach. Section 5 provides the official evaluation results of WAT2020[3] followed by the conclusion in Section 6.

2 Dataset

We have used the official datasets provided by the WAT2020 organizers for the tasks and also used additional datasets recommended by the organizers.

[1] http://lotus.kuee.kyoto-u.ac.jp/WAT/

[2] https://www.britannica.com/topic/Odia-language

[3] http://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2020/index.html

Set	Sentences	Tokens English	Tokens Hindi
HVG Train	28930	143164	145448
IITB Train	1.5 M	20.6 M	22.1 M
D-Test	998	4922	4978
E-Test	1595	7853	7852
C-Test	1400	8186	8639

Table 1: Statistics of our data used in the English→Hindi Multimodal task: the number of sentences and tokens.

Task 1: English→ Hindi Multimodal Translation For this task, the organizers provided HindiVisualGenome 1.1 (Parida et al., 2019a)[4] dataset (HVG for short). 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). Our WAT submissions were for E-Test (denoted "EV" in WAT official tables) and C-Test (denoted "CH" in WAT tables). Additionally, we used the IITB Corpus[5] which is supposedly the largest publicly available English-Hindi parallel corpus (Kunchukuttan et al., 2017). This corpus contains 1.59 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.

Task 2: Indic Odia↔English Translation For this task, the organizers provided OdiEnCorp 2.0 (Parida et al., 2020b).[6] To train the model, we used an additional dataset (OdiEnMonoCorp[7]) suggested by the organizers (Parida et al., 2020a). The statistics of the datasets are shown in Table 2.

Task 3: Indic Multilingual Translation For this task, the organizers provided filtered data of the PMIndia dataset (Haddow and Kirefu, 2020).[8] We have not used any additional resources in this task. The statistics of the dataset are shown in Table 3.

[4] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3267

[5] http://www.cfilt.iitb.ac.in/iitb_parallel/

[6] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3211

[7] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2879

[8] http://data.statmt.org/pmindia/

Set	Sentences	Tokens English	Tokens Odia
Train	69370	1.34 M	1.16 M
Train (Monolingual)	71663	2.8 M	2.64 M
Dev	13544	158188	140726
Test	14344	186320	165274

Table 2: Statistics of our data used in Odia↔English translation task: the number of sentences and tokens.

Set	Sentences Train	Sentences Dev	Sentences Test	Tokens English	Tokens *Target*
Bengali	26239	2 K	3522	0.46 M	0.39 M
Hindi	52718	2 K	3169	1 M	1.09 M
Malayalam	29704	2 K	2886	0.54 M	0.33 M
Tamil	35224	2 K	3637	0.64 M	0.47 M
Telugu	35989	2 K	3049	0.64 M	0.53 M
Gujarati	44083	2 K	4463	0.84 M	0.77 M
Marathi	31669	2 K	3760	0.57 M	0.46 M
Total	255626	14 K	24486	4.69 M	4.03 M

Table 3: Statistics of PMIndia used in Indic Multilingual translation task: the number of sentences pairs and tokens for English and the respective target language.

3 Experiment

We focussed only on the text translation and image captioning task.

In the English→Hindi Multimodal task, for the 'Text-Only' subtask, we used the *Transformer* model (Vaswani et al., 2018) which is popular for machine translation and other text-processing tasks, such as low resource text summarization (Parida and Motlicek, 2019). We have used the *Transformer* model as implemented in OpenNMT-py (Klein et al., 2017).[9] We used *InceptionResNetV2*[10] for 'Hindi-only' image captioning subtask.

In the Odia↔English translation task, we used the Transformer model as implemented in OpenNMT-tf.[11]

In the Indic Multilingual translation task, we used the *Transformer (big)* model with relative position representations (Shaw et al., 2018) as implemented in OpenNMT-tf (Klein et al., 2017).

[9] http://opennmt.net/OpenNMT-py/quickstart.html

[10] https://keras.io/api/applications/inceptionresnetv2/

[11] http://https://opennmt.net/OpenNMT-tf/quickstart.html

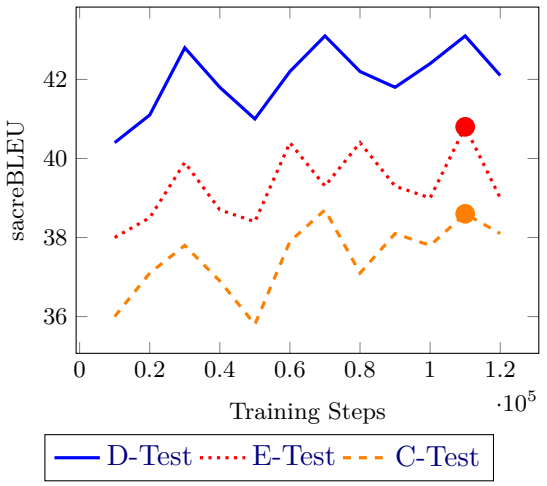


Figure 1: English→Hindi Multimodal task ('text-only') learning curves in terms of sacreBLEU score. The big round dots indicate which training iteration was used when producing our final submissions to the WAT manual and automatic evaluation for E-Test and C-Test.

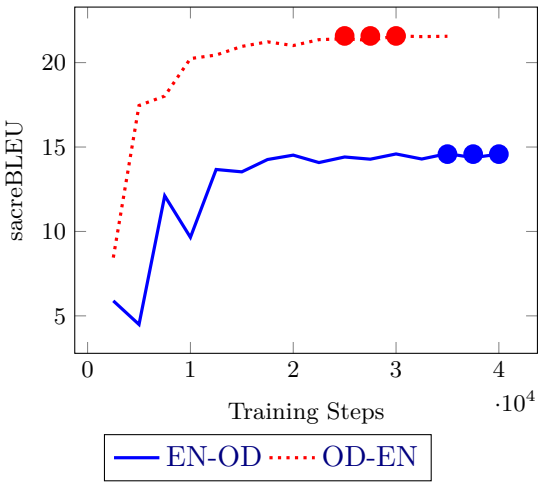


Figure 2: Odia↔English Translation task learning curves in terms of sacreBLEU score on Dev set. The big round dots indicate which training iterations were averaged into a single model when producing our final submissions to the WAT manual and automatic evaluation.

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.

In the English→Hindi Multimodal task, we generated the vocabulary of 32k sub-word types jointly for both the source and target languages. In Odia↔English task and Indic Multilingual task, we generated the vocabulary of 24k sub-word types jointly for both the source and target languages. The vocabulary is shared between the encoder and decoder for all the tasks.

3.2 Training

English→Hindi Multimodal task: To train the model, we used a single GPU and followed the standard "Noam" learning rate decay,[12] 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. For the 'text-only' subtask of the English→Hindi Multimodal task, we concatenated HVG and IITB training data and shuffled it at the level of sentences. The learning curve is shown in Figure 1.

[12] https://nvidia.github.io/OpenSeq2Seq/html/api-docs/optimizers.html

Indic Odia↔English translation task For the Odia↔English translation task, we back-translated the Odia sentences from the OdiEnMonoCorp to increase the training set size. OdiEnMonoCorp is distributed in the form of paragraphs which we split into individual sentences and equipped each Odia sentence with synthetic English produced using Google Translate. We used *LazyAdam* optimizer and 8000 warm-up steps. We used averaging of checkpoints at an interval of 2500 steps as the final model. The final model for Odia→English used the checkpoints at 25K, 27.5K, and 30K training steps and the English→Odia model used the check-points at 35K, 37.5K, and 40K steps. The learning curve on the development sets is shown in Figure 2.

Indic Multilingual translation task: We trained two multi-lingual models for the translation of English from and into Bengali, Hindi, Malayalam, Tamil, Telugu, Marathi, and Gujarati. We followed the solution proposed in Multilingual Neural Machine Translation (Johnson et al., 2017) and prefixed the input sentence with an artificial token to specify the required target language. We used *LazyAdam* optimizer and 8000 warm-up steps. The final One-to-Many model for English into Indic languages used the checkpoint at 32.5K training

System and WAT Task Label	WAT BLEU ODIANLP	Best competitor
English→Hindi Multimodal task		
MMEVTEXT20en-hi	**40.85***	38.84
MMEVHI20en-hi	**0.78**	-
MMCHTEXT20en-hi	**38.50**	27.75
MMCHHI20en-hi	0.0	-
Indic Odia↔English translation task		
ODIAENen-od	**11.07***	9.85
ODIAENod-en	**18.31***	17.89

Table 4: WAT2020 Automatic Evaluation Results for English→Hindi and Odia↔English. The scores marked with '∗' indicate the best performance in its track among all competitors. For each task, we show the score of our system (ODIANLP) and the score of the best competitor in the respective task. The scores marked with '∗' indicate the best performance in its track among all competitors.

	From English				Into English			
WAT Task	ODIANLP	Human	Best Comp	Human	ODIANLP	Human	Best Comp	Human
INDIC20en-bn	16.38	3.1	**19.64**	**3.9**	19.71	1.7	**23.38**	**3.9**
INDIC20en-hi	21.05	2.7	**24.48**	**3.8**	21.88	1.7	**28.51**	**3.7**
INDIC20en-gu	11.24	-	**14.66**	-	20.47	-	**30.26**	-
INDIC20en-ml	3.41	-	**6.32**	-	15.30	-	**20.87**	-
INDIC20en-mr	8.79	-	**11.52**	-	16.85	-	**24.05**	-
INDIC20en-ta	4.94	-	**7.21**	-	14.53	-	**20.16**	-
INDIC20en-te	4.09	-	**6.93**	-	14.94	-	**19.03**	-

Table 5: WAT2020 Automatic and Manual Evaluation Results for Indic Multilingual Task. For each task, we show the score of our system (ODIANLP) and the score of the best competitor ('Best Comp') in the respective task. WAT2020 performed human evaluation for the INDIC20en-bn, INDIC20bn-en, INDIC20en-hi, and INDIC20hi-en task.

steps. The final Many-to-One model for Indic languages into English used the checkpoint at 57.5K training steps.

4 Official Results

We report the official automatic and human evaluation results of our models for all the participated tasks here in Table 4 and Table 5.

5 Discussion

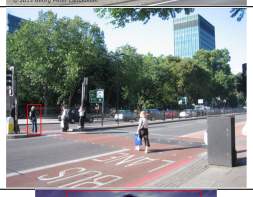
We have analyzed the result and report some observations based on the automatic evaluation scores. For the English→Hindi Multimodal sub-task, our text-only submission (with an additional resource for training) obtains the best result compared to utilizing both text and image (by competitors in the multimodal track). A sample output generated from our model is shown in Figure 3. Although our NMT model able to translate many ambiguous words (e.g. *Cross*) in many instances, still it fails in some instances which could be resolved using an image as shown in Figure 3.

Our submission to the image captioning task ('Hindi only', denoted with 'MM*HI20en-hi' in Table 4) failed. The system generated generally fluent text segments but they were not related to the image.

BLEU scores for the Odia→English translation are higher than English→Odia in Table 4, but this does not necessarily indicate that the translation into English would be better, because cross-language comparison of BLEU scores is generally not possible. The situation is different in the Indic Multilingual task (Table 5) where translation into English benefits from the English target side of other languages in the mixed corpus, but again, this claim should not be made based on cross-language BLEU comparison but rather by comparing multilingual with standard bilingual baseline.

Sample outputs generated from our Odia model are shown in Figures 4 and 5.

English Input: a man trying to **cross**
Translated Output: एक आदमी क्रॉस करने की कोशिश कर रहा है
Gloss: A man trying to cross



English Input: the woman is waiting to **cross** the street
Translated Output: महिला सड़क पार करने की प्रतीक्षा कर रही है।
Gloss: The woman is waiting to cross the street

English Input: the lady appears to be going **cross** country skiing
Translated Output: लगता है कि महिला क्रॉस कंट्री स्कीइंग जा रही है
Gloss: It seems that the lady is going for cross country skiing

English Input: a **cross** sign on top of the tower
Translated Output: टॉवर के शीर्ष पर एक पार संकेत
Gloss: A par sign on top of tower

Figure 3: Sample Hindi output as generated for the challenge test set. The ambiguous source word is bolded in the English input. We illustrate one error in the last example, 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.

English to Odia			
Translation Type	English (Source)	Odia (Translation)	Gloss/Remark
Correct Translation	It is located on the bank of the River Sone which merges with River Ganges at Digha a few kilometers from Danapur.	ଏହା ଦାନପୁର ଠାରୁ କିଛି କିଲୋମିଟର ଦୂରରେ ଦିଘାଠାରେ ଗଙ୍ଗା ନଦୀ ସହ ମିଳିତ ହେଉଥିବା ସୋନ ନଦୀର କୂଳରେ ଅବସ୍ଥିତ ।	It is located on the bank of the river Sone which merges with river Ganges a few kilometer away from Danapur
Partial Correct Translation	The temple is maintained by the Bengal, Bihar and Odisha Digambara Jaina Tirthankara Committee Bimala Devi Jain is the local caretaker.	ଏହି ମନ୍ଦିରର ରକ୍ଷଣାବେକ୍ଷଣ ବଙ୍ଗଳା, ବିହାର ଓ ଓଡ଼ିଶା ଦିଗମ୍ବରା ଜୈନ ତୀର୍ଥଙ୍କରର କମିଟି ବିମଲା ଦେବୀ ଜୈନ ତାଙ୍କର ସ୍ଥାନୀୟ କ୍ୱାରିଅର ।	This temple is maintained by the Bengal, Bihar, and Odisha Digambara Jain Tirthankara committee Bimala Devi is the local career. (the word "caretaker" mistranslated into Odia)
Incorrect Translation	donator	ଦାନମାଦାତ	Mistranslated the English word "donator" into Odia

Figure 4: Sample English→Odia output generated by our NMT model including correct, partial correct, and incorrect translation.

6 Conclusion and Future Scope

In this system description paper, we presented our systems for three tasks in WAT 2020 in which we participated: i) English→Hindi Multimodal task, ii) Indic Odia↔English, and iii) Indic Multilingual translation task.

As the next steps, i) we plan to explore more on the Indic Multilingual task utilizing additional resources for training, ii) analyze the image captioning task which didn't work, and iii) utilize image features for improving the translation quality.

Acknowledgments

At Idiap, the work was supported by the EU H2020 project "Real-time network, text, and speaker analytics for combating orga-

Odia to English			
Translation Type	Odia (Source)	English (Translation)	Gloss/Remark
Correct Translation	ତେଣୁ ତମେ ତାଙ୍କର ଶତ୍ରୁ।	So you are his enemy.	Therefore you are his enemy
Partial Correct Translation	'ଭାରତୀୟ ସିନେମା ର ଜନକ' ବୋଲି ଅଭିହିତ କରାଯାଉଥିବା ଦାଦାସାହେବ ଫାଳକେ ଭାରତ ର ପ୍ରଥମ ପୂର୍ଣ୍ଣ ଦୀର୍ଘ ଚଳଚିତ୍ର ' ରାଜା ହରିଶ୍ଚନ୍ଦ୍ର' ନିର୍ମାଣ କରିଥିଲେ ୧୯୧୩ ମସିହା ରେ	Dadasaheb Phalke, who is described as the "father of Indian Cinema", built the first long film of India" in 1913.	Dadasaheb Phalke who called as "The father of Indian Cinema" build the first full length cinema "Raja Harishchandra" in 1913 (the movie name "Raja Harishchandra" missing)
Incorrect Translation	ଉଦ୍ଧୃତି ଚିହ୍ନ ପ୍ରାରମ୍ଭ	Open number	Begin of quotation mark (mistranslated the Odia word "ଉଦ୍ଧୃତି ଚିହ୍ନ")

Figure 5: Sample Odia→English output generated by our NMT model including correct, partial correct, and incorrect translation.

nized 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

Barry Haddow and Faheem Kirefu. 2020. PMIndia – A Collection of Parallel Corpora of Languages of India. *arXiv e-prints*, page arXiv:2001.09907.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, 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. OpenNMT: Open-source toolkit for neural machine translation. In *Proc. ACL*.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2017. The IIT Bombay English-Hindi Parallel Corpus. *arXiv preprint arXiv:1710.02855*.

Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, and Sadao Kurohashi. 2020. Overview of the 7th workshop on Asian translation. In *Proceedings of the 7th Workshop on Asian Translation*, Suzhou, China. Association for Computational Linguistics.

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 multi-modal english to hindi machine translation. *Computación y Sistemas*, 23(4).

Shantipriya Parida, Ondřej Bojar, and Satya Ranjan Dash. 2020a. Odiencorp: Odia–english and odia-only corpus for machine translation. In *Smart Intelligent Computing and Applications*, pages 495–504. Springer.

Shantipriya Parida, Ondřej Bojar, and Petr Motlicek. 2019b. Idiap nmt system for wat 2019 multimodal translation task. In *Proceedings of the 6th Workshop on Asian Translation*, pages 175–180.

Shantipriya Parida, Satya Ranjan Dash, Ondřej Bojar, Petr Motlicek, Priyanka Pattnaik, and Debasish Kumar Mallick. 2020b. Odiencorp 2.0: Odia-english parallel corpus for machine translation. In *Proceedings of the WILDRE5–5th Workshop on Indian Language Data: Resources and Evaluation*, pages 14–19.

Shantipriya Parida and Petr Motlicek. 2019. Abstract text summarization: A low resource challenge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5996–6000.

Martin Popel and Ondřej Bojar. 2018. Training tips for the transformer model. *The Prague Bulletin of Mathematical Linguistics*, 110(1):43–70.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468. Association for Computational Linguistics.

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.