Silo NLP's Participation at WAT2022

Shantipriya Parida[*], **Subhadarshi Panda**[†], **Stig-Arne Grönroos** [*‡], **Mark Granroth-Wilding** [*], **Mika Koistinen** [*]
[*]Silo AI, Helsinki, Finland
{firstname.lastname}@silo.ai
[†]Graduate Center, City University of New York, USA
spanda@gradcenter.cuny.edu
[‡]University of Helsinki, Finland

Abstract

This paper provides the system description of "Silo NLP's" submission to the Workshop on Asian Translation (WAT2022). We have participated in the Indic Multimodal tasks (English→Hindi, English→Malayalam, and English→Bengali, Multimodal Translation). For text-only translation, we trained Transformers from scratch and fine-tuned mBART-50 models. For multimodal translation, we used the same mBART architecture and extracted object tags from the images to use as visual features concatenated with the text sequence.

Our submission tops many tasks including English→Hindi multimodal translation (evaluation test), English→Malayalam text-only and multimodal translation (evaluation test), English→Bengali multimodal translation (challenge test), and English→Bengali text-only translation (evaluation test).

1 Introduction

Machine translation (MT) is a classic sub-field in NLP which investigates the usage of computer software to translate text or speech from one language to another without human involvement (Yang et al., 2020). Although MT performance has reached near the level of human translators for many high-resource languages, it remains challenging for many low-resource languages (Popel et al., 2020; Costa-jussà et al., 2022). Also, effective usage of other modalities (e.g. image) in MT is an important research area in the past few years (Sulubacak et al., 2020; Parida et al., 2021b,a).

The WAT is an open evaluation campaign focusing on Asian languages since 2013 (Nakazawa et al., 2020, 2022). In the WAT2022 Multimodal track, a new Indian language *Bengali* was introduced for English→Bengali text, multimodal translation, and Bengali image captioning task.[1]

The multimodal translation tasks in WAT2022 consist of image caption translation, in which the input is a descriptive source language caption together with the image it describes, while the output is a target language caption. The multimodal input enables the use of image context to disambiguate source words with multiple senses.

In this system description paper, we explain our approach for the tasks (including the sub-tasks) we participated in:

Task 1: English→Hindi (EN-HI) Multimodal Translation

- EN-HI text-only translation
- EN-HI multimodal translation

Task 2: English→Malayalam (EN-ML) Multimodal Translation

- EN-ML text-only translation
- EN-ML multimodal translation

Task 3: English→Bengali (EN-BN) Multimodal Translation

- EN-BN text-only translation
- EN-BN multimodal translation

2 Data sets

We used the data sets specified by the organizer for the related tasks along with additional synthetic data for performance improvement. The use of additional data places some[2] of our submissions in the unconstrained track.

Task 1: English→Hindi Multimodal Translation For this task, the organizers provided HindiVisualGenome 1.1 (Parida et al., 2019)[3] 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

[1] https://ufal.mff.cuni.cz/bengali-visual-genome/wat-2022-english-bengali-multim

[2] All except the EN–HI and EN–ML text-only systems.

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

Proceedings of the 9th Workshop on Asian Translation, pages 99–105
October 17, 2022.

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).

For the synthetic image features, we use the Flickr8k data set (Hodosh et al., 2013). Even though it is an image captioning data set, we discard the images, treating the data set as in-domain monolingual data. We use a machine translation into Hindi (Rathi, 2020) as the target side, and generate image features using the procedure described in Section 3.4.

The statistics of the datasets are shown in Table 1.

Task 2: English→Malayalam Multimodal Translation For this task, the organizers provided MalayalamVisualGenome 1.0 dataset[4] (MVG for short). MVG is an extension of the HVG dataset for supporting Malayalam, which belongs to the Dravidian language family (Kumar et al., 2017). The dataset size and images are the same as HVG. While HVG contains bilingual English–Hindi segments, MVG contains bilingual English–Malayalam segments, with the English, shared across HVG and MVG, see Table 1.

Task 3: English→Bengali Multimodal Translation For this task, the organizers provided BengaliVisualGenome 1.0 dataset[5] (BVG for short). BVG is an extension of the HVG dataset for supporting Bengali. The dataset size and images are the same as HVG, and MVG, see Table 1.

3 Experimental Details

This section describes the experimental details of the tasks we participated in.

3.1 EN-HI, EN-ML, EN-BN text-only translation

For EN–HI and EN–ML text-only (E-Test and C-Test) translation, we fine-tuned a pre-trained mBART-50 model (Tang et al., 2020) without using any additional resources.

For EN-BN text-only (E-Test) translation, we used the Transformer base model as implemented in Open-NMT-Py[6] using Bangla Natural Language Image to Text (BNLIT) (Jishan et al., 2019) as an additional dataset. The BNLIT is an image-to-text dataset containing 8743 images and their corresponding text in Bengali. For our experiment, we used Bengali text and translated it into English.

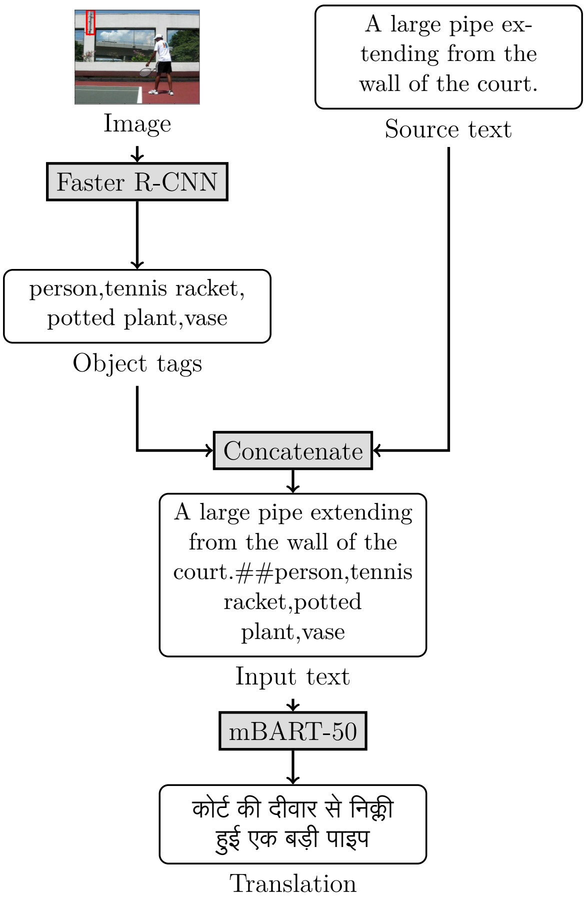

A large pipe extending from the wall of the court.
Image
Source text
Faster R-CNN
person,tennis racket, potted plant,vase
Object tags
Concatenate
A large pipe extending from the wall of the court.##person,tennis racket,potted plant,vase
Input text
mBART-50
कोर्ट की दीवार से निक्ली हुई एक बड़ी पाइप
Translation

Figure 1: Multimodal translation pipeline.

Subword units were constructed using the word pieces algorithm (Johnson et al., 2017). Tokenization is handled automatically as part of the preprocessing pipeline of word pieces. We generated the vocabulary of 32k sub-word types jointly for both the source and target languages. The training steps are defined for 300K with 5K as validation steps and checkpoint steps. For translation, we decoded using beam search with beam size 5.

For C-Test we fine-tuned a pre-trained m-BART-50 model without any additional resources.

3.2 EN-HI, EN-ML, EN-BN Multimodal translation

Our multimodal translation pipeline is shown in Figure 1. For EN-HI multimodal (E-Test and C-Test) translation, we used the object tags extracted from the HVG dataset images (see Section 3.3) for image features and concatenated them with the text.

[4] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3533

[5] http://hdl.handle.net/11234/1-3722

[6] https://opennmt.net/OpenNMT-py/

Set	**Sentences**	**Tokens**			
		English	Hindi	Malayalam	Bengali
Train	28930	143164	145448	107126	113978
D-Test	998	4922	4978	3619	3936
E-Test	1595	7853	7852	5689	6408
C-Test	1400	8186	8639	6044	6657

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

Additionally, we used synthetic image features (see Section 3.4. The combined data set was used to fine-tune a pre-trained mBART-50 model.

For EN-ML multimodal (E-Test and C-Test) translation, we used object tags extracted from the MVG dataset images and concatenated with the text, and fine-tuned on the mBART-50 model.

For EN-BN (E-Test and C-Test) translation, we used object tags extracted from the BVG dataset images and concatenated with the text, and fine-tuned on the mBART-50 model.

For all the multimodal translation experiments using mBART, the decoding beam size was set to 5. Since we used the pre-trained mBART model and fine-tuned it on the visual genome datasets, we did not build our own vocabulary but rather used the pre-trained mBART vocabulary without any modifications.

3.3 Extracted image features

We derive the list of object tags for a given image using the pre-trained Faster R-CNN with ResNet-101-C4 backbone. It can recognize 80 object types from the COCO data set (Lin et al., 2014). Based on their confidence scores, we pick the top 10 object tags. In cases where less than 10 object tags are detected, we consider all the detected tags. Figure 2 shows examples of object tags detected for images from the challenge test set. The detected object tags are then concatenated to the English sentence which needs to be translated to Hindi, Malayalam, and Bengali. The concatenation is done using the special token '##' as the separator. The separator is followed by comma-separated object tags. Adding objects enables the model to utilize visual concepts which may not be readily available in the original sentence. The English sentences along with the object tags are fed to the encoder of the mBART model.

3.4 Synthetic image features

We generate synthetic training data for the multimodal translation task by enriching text-only data using synthetic image features (Grönroos et al., 2018). Grönroos et al. (2018) use continuous image features and generate the synthetic dummy features by taking the average vector of the features in the training data. We improve on this procedure by generating discrete features individually for each enriched training example by decoding from a sequence-to-sequence (s2s) model. The s2s model is trained using the multimodal training data (HVG), but instead of training the normal way, we use both source language and target language text as input (with a separator token between), and our object tags as the output. The model is a Transformer-base (Vaswani et al., 2017) model trained using the Marian-NMT (Junczys-Dowmunt et al., 2018) framework. The trained model is then used to enrich text-only parallel data with synthetic features.

Our method applies to any text-only parallel data, even though in this experiment we use it to enrich the text from an image captioning data set. Due to the relatively small size of the multimodal training data sets, domain match of additional training data has great importance for multimodal translation (Grönroos et al., 2018). It is therefore important to apply domain adaptation techniques when using general-domain text-only parallel data.

4 Results

We report the official automatic evaluation results of our models for all the participating tasks in Table 2.

On the E-Test sets, our multimodal systems receive the highest BLEU scores for English→Hindi and English→Malayalam, and our text-only systems outperform other text-only systems for English→Malayalam and English→Bengali. Our multimodal systems consistently outperform our text-only systems, with increases between +1.1 and +10.2 BLEU.

On the C-Test challenge sets, we have the highest BLEU score for English→Bengali multimodal translation. Again, our multimodal systems outper-

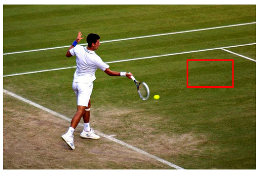
English input: Red dragon fruit on a fruit stand.
Object tags: person, banana, apple, broccoli
Text-only translation: एक स्वाद के फल पर लाल नारंगी फल
Gloss: Red orange fruit on a flavored fruit
Multimodal translation: एक फल स्टैंड पर लाल ड्रैगन फल।
Gloss: Red dragon fruit on a fruit stand.
Reference: फल स्टैंड पर लाल ड्रैगन फल।
Gloss: Red dragon fruit on a fruit stand.

English input: A metal and stone column holding a bell and cross.
Object tags: clock
Text-only translation: एक धातु और घड़ी पकड़े हुए
Gloss: Holding a metal and a watch
Multimodal translation: एक धातु और पत्थर के स्तंभ एक घंटी और क्रॉस पकड़े हुए।
Gloss: A metal and stone pillar holding a bell and a cross.
Reference: एक धातु और पत्थर के स्तंभ एक घंटी और क्रॉस के आधार बनें हुए।
Gloss: A metal and stone pillar forming the basis of a bell and cross.

English input: date the photo was taken
Object tags: teddy bear, cat
Text-only translation: ছবি ছবি চিত্র করা হয়
Gloss: Picture is picture picture
Multimodal translation: ছবি তোলা হয়েছিল
Gloss: The picture was taken
Reference: তারিখটি ছবি তোলা হয়েছিল
Gloss: The date it was photographed

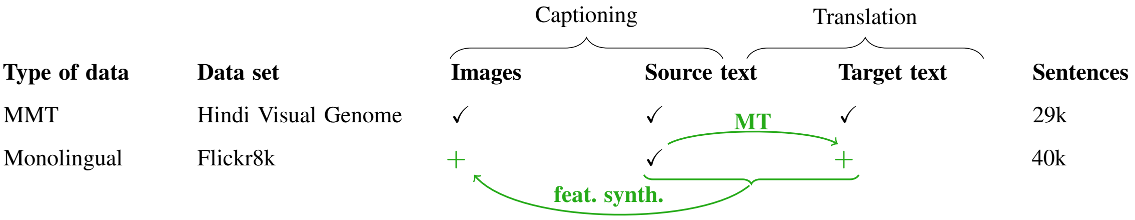
English input: open door of a second bus
Object tags: suitcase, person, bus, backpack, truck, traffic light
Text-only translation: একটি দুটি বাসের খোলা দরজা
Gloss: An open door of two buses
Multimodal translation: একটি দ্বিতীয় বাসের দরজা
Gloss: A second bus door
Reference: দ্বিতীয় বাসের দরজা খোলা
Gloss: The door of the second bus is open

English input: the two male players are after the ball
Object tags: person, sports ball, motorcycle, umbrella, tennis racket, backpack
Text-only translation: ഒരു ജിറാഫ് പുല്ല് തിന്നുന്നു
Gloss: A giraffe eats grass
Multimodal translation: രണ്ട് പുരുഷ കളിക്കാർ പന്തിന് ശേഷം
Gloss: Two male players after the ball
Reference: രണ്ട് പുരുഷ കളിക്കാർ പന്തിന് പുറകേയാണ്
Gloss: Two male players are behind the ball



English input: Grass growing on the grass tennis court.
Object tags: person, tennis racket, sports ball
Text-only translation: പുല്ല് ടെന്നീസ് കോർട്ടിൽ വളരുന്ന പുല്ല്.
Gloss: Grass is the grass that grows on a tennis court.
Multimodal translation: പുല്ല് ടെന്നീസ് കോർട്ടിൽ പുല്ല് വളരുന്നു.
Gloss: Grass grows on a grass tennis court.
Reference: ഗ്രാസ് ടെന്നീസ് കോർട്ടിൽ പുല്ല് വളരുന്നു.
Gloss: Grass grows on a grass tennis court.

Figure 2: Sample translations for the challenge test set. Both the text-only and multimodal translations are shown. The object tags detected and used in the multimodal translation setup are also shown. Hindi translations are shown for the top two images, Bengali translations are shown for the middle two images, and Malayalam translations are shown for the bottom two images.



		Captioning		Translation	
Type of data	**Data set**	**Images**	**Source text**	**Target text**	**Sentences**
MMT	Hindi Visual Genome	✓	✓	✓	29k
Monolingual	Flickr8k	+	✓	+	40k

Figure 3: Process for generating synthetic image features. Green arrows indicate processes to synthesize data, and green plus signs indicate the resulting synthetic data. MT is short for machine translation, feat. synth. for image feature synthesis.

	WAT BLEU		HUMAN		NOTE
System and WAT Task Label	Silo NLP	Best Comp	Silo NLP	Best Comp	
English→Hindi MM Task					
MMEVTEXT21en-hi	36.2	**42.9**			
MMEVMM22en-hi	**42.0**	39.4			
MMCHTEXT22en-hi	29.6	**41.8**	2.715	**3.078**	
MMCHMM22en-hi	39.1	**39.3**			
English→Malayalam MM Task					
MMEVTEXT21en-ml	**30.8**	30.6			
MMEVMM22en-ml	**41.0**	–			
MMCHTEXT22en-ml	14.6	**19.5**	2.013	Not Available	
MMCHMM22en-ml	**20.4**	–			
English→Bengali MM Task					
MMEVTEXT22en-bn	**41.0**	**41.0**			
MMEVMM22en-bn	42.1	**43.9**			
MMCHTEXT22en-bn	22.6	32.9	2.658	**3.525**	Competitor used external data
MMCHMM22en-bn	**28.7**	**28.7**			

Table 2: WAT2022 Automatic and Manual Evaluation Results for English→Hindi, English→Malayalam, and English→Bengali. Rows containing “TEXT" in the task label name denote text-only translation track, and the rows representing “MM" denote multimodal translation including text and images. “–" indicates single submission for the task. For each task, we show the score of our system (Silo NLP) and the score of the best competitor in the respective task.

form our text-only systems, with increases between +5.7 and +9.5 BLEU.

It should be noted that our English→Hindi multimodal system and English→Bengali text-only system are unconstrained, making use of substantial additions of in-domain data. For these language pairs, the increase in translation quality can not be attributed entirely to multimodality. However, for English→Malayalam both systems use the same training sentences, making these systems comparable. The English→Malayalam multimodal system outperforms the text-only system by +10.2 BLEU for E-test and +5.7 BLEU for C-test. As the C-test challenge set is constructed to contain translational ambiguity, the improvement is an indication that the image features are useful for disambiguation. The human evaluation scores from the WAT organizers for the available sub-tasks are updated in Table 2.

We demonstrate examples of translations obtained for the challenge test set in Figure 2. The extracted image features in the form of object tags are also shown for each image in the figure. We observe that the multimodal translations are notably better than the text-only translations. This is consistent with the pattern of the BLEU scores in Table 2.

5 Conclusions

In this system description paper, we presented our system for three tasks in WAT2022: (a) English→Hindi, (b) English→Malayalam, and (c) English→Bengali Multimodal Translation. We released the code through Github for research[7].

In the next step, we will explore the usage of synthetic features in multimodal translation for the Malayalam and Bengali languages, to make use of available text-only corpora for these language pairs.

[7] https://github.com/shantipriyap/SiloNLP_WAT2022

Acknowledgements

We are thankful to Silo AI[8], Helsinki, Finland for the computation resources and the necessary support for participating in WAT2022.

References

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv e-prints*, pages arXiv–2207.

Stig-Arne Grönroos, Benoit Huet, Mikko Kurimo, Jorma Laaksonen, Bernard Merialdo, Phu Pham, Mats Sjöberg, Umut Sulubacak, Jörg Tiedemann, Raphaël Troncy, and Raúl Vázquez. 2018. The MeMAD submission to the WMT18 multimodal translation task. In *Proceedings of the Third Conference on Machine Translation (WMT)*. The Association for Computational Linguistics.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Md. Asifuzzaman Jishan, Khan Raqib Mahmud, and Abul Kalam Al Azad. 2019. Bangla natural language image to text (bnlit).

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.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Arun Kumar, Ryan Cotterell, Lluís Padró, and Antoni Oliver. 2017. Morphological analysis of the Dravidian language family. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 217–222.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Toshiaki Nakazawa, Hideya Mino, Isao Goto, Raj Dabre, Shohei Higashiyama, Shantipriya Parida, Anoop Kunchukuttan, Makoto Morishita, Ondřej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, Yusuke Oda, and Sadao Kurohashi. 2022. Overview of the 9th workshop on Asian translation. In *Proceedings of the 9th Workshop on Asian Translation (WAT2022)*, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, et al. 2020. Overview of the 7th workshop on Asian translation. In *Proceedings of the 7th Workshop on Asian Translation*, pages 1–44.

Shantipriya Parida, Ondřej Bojar, and Satya Ranjan Dash. 2019. Hindi visual genome: A dataset for multi-modal English to Hindi machine translation. *Computación y Sistemas*, 23(4).

Shantipriya Parida, Subhadarshi Panda, Satya Prakash Biswal, Ketan Kotwal, Arghyadeep Sen, Satya Ranjan Dash, and Petr Motlicek. 2021a. Multimodal neural machine translation system for English to Bengali. In *Proceedings of the First Workshop on Multimodal Machine Translation for Low Resource Languages (MMTLRL 2021)*, pages 31–39, Online (Virtual Mode). INCOMA Ltd.

Shantipriya Parida, Subhadarshi Panda, Ketan Kotwal, Amulya Ratna Dash, Satya Ranjan Dash, Yashvardhan Sharma, Petr Motlicek, and Ondřej Bojar. 2021b. Nlphut's participation at wat2021. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 146–154.

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtskỳ. 2020. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature communications*, 11(1):1–15.

Ankit Rathi. 2020. Deep learning apporach for image captioning in Hindi language. In *2020 International Conference on Computer, Electrical & Communication Engineering (ICCECE)*, pages 1–8. IEEE.

Umut Sulubacak, Ozan Caglayan, Stig-Arne Grönroos, Aku Rouhe, Desmond Elliott, Lucia Specia, and Jörg Tiedemann. 2020. Multimodal machine translation through visuals and speech. *Machine Translation*, 34(2):97–147.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all

[8] https://silo.ai

you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010.

Shuoheng Yang, Yuxin Wang, and Xiaowen Chu. 2020. A survey of deep learning techniques for neural machine translation. *arXiv preprint arXiv:2002.07526*.