ISSN 2007-9737

Hindi Visual Genome: A Dataset for Multi-Modal English to Hindi Machine Translation

Shantipriya Parida[1], Ondřej Bojar[1], Satya Ranjan Dash[2]

[1] Charles University, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics,
Czech Republic

[2] KIIT University, School of Computer Application,
India

{parida,bojar}@ufal.mff.cuni.cz, sdashfca@kiit.ac.in

Abstract. Visual Genome is a dataset connecting structured image information with English language. We present "Hindi Visual Genome", a multi-modal dataset consisting of text and images suitable for English-Hindi multi-modal machine translation task and multi-modal research. We have selected short English segments (captions) from Visual Genome along with the associated images and automatically translated them to Hindi. A careful manual post-editing followed which took the associated images into account. Overall, we prepared a set of 31,525 segments (which we conveniently split into training, development and test data), accompanied by a challenge test set of 1,400 segments. This challenge test set was created by searching for particularly ambiguous English words based on the embedding similarity and manually selecting those where the image helps to resolve the ambiguity. Our dataset is the first manually revised dataset for multi-modal English-Hindi machine translation, freely available for non-commercial research purposes. Our Hindi version of Visual Genome also allows to create Hindi image labelers or other practical tools. Hindi Visual Genome also served in Workshop on Asian Translation (WAT) 2019 Multi-Modal Translation Task.

Keywords. Visual genome, Multi-modal corpus, parallel corpus, word embedding, neural machine translation (NMT), image captioning.

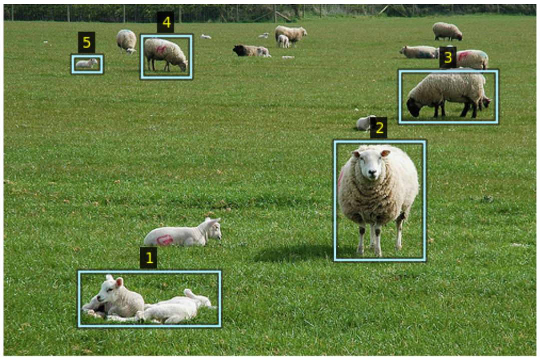
1 Introduction

Multi-modal content is gaining popularity in machine translation (MT) community due to its appealing chances to improve translation quality by considering information available only in the visual context of a sentence. Handling multi-modal content also has its usage in commercial applications such as translation of image captions in online news articles or machine translation for e-commerce product listings [11, 3, 9, 22]. Although the general performance of neural machine translation (NMT) models is very good given large amounts of parallel texts, some inputs can remain genuinely ambiguous, especially if the input context is limited.

One example is the word "mouse" in English (source) which can be translated into different words in Hindi based on the context (e.g. either a computer mouse or a small rodent), another example is given in Fig. 1 below. There is a limited number of multi-modal datasets available and even fewer of them are also multi-lingual. Our aim is to extend the set of languages available for multi-modal experiments by adding a Hindi variant of a subset of Visual Genome.

Creating such a dataset enables multi-modal experimenting with Hindi for various applications

doi: 10.13053/CyS-23-4-3294



1: *Two lambs lying in the sun.*
Hindi MT literal translation (Gloss):
Two baby sheep are ***telling lies*** *. . .*

Selected surrounding captions:
2. *Sheep standing in the grass*
3. *Sheep with black face and legs*
4. *Sheep eating grass*
5. *Lamb sitting in grass.*

Fig. 1. A sample image from Visual Genome with several rectangular regions annotated with English captions. The caption of region 1 was processed by baseline MT into Hindi, resulting in a translation error (underlined)

and it could also facilitate the exploration of how the language is grounded in vision.

Visual Genome (`http://visualgenome.org/`, [10]) is a large set of real-world images, each equipped with annotations of various regions in the image. The annotations include a plain text description of the region (usually sentence parts or short sentences, e.g. "a red ball in the air") and also several other formally captured types of information (objects, attributes, relationships, region graphs, scene graphs, and question-answer pairs). We focus only on the textual descriptions of image regions and provide their translations into Hindi.

An example of a Visual Genome image with captions is given in Fig. 1. Region 1 is one of the captions in Visual Genome ("Two lambs lying in the sun.") and we illustrate how MT into Hindi mistranslated the word "lying" as "telling lies". While such a meaning is possible (in fairy tales), the image clearly hints towards to correct meaning. In this particular case, the surrounding captions could help but no direct evidence for or against the two meanings is available in the textual information.

Table 1. Hindi Visual Genome corpus details. One item consists of an English source segment, its Hindi translation, the image and a rectangular region in the image

Data Set	Items
Training Set	28,932
Development Test Set (D-Test)	998
Evaluation Test Set (E-Test)	1,595
Challenge Test Set (C-Test)	1,400

The main portion of our Hindi Visual Genome is intended for training purposes of tools like multi-modal translation systems or Hindi image labelers. Every item consists of an image, a rectangular region in the image, the original English caption from Visual Genome and finally our Hindi translation. Additionally, we create a challenge test set with the same structure but a different sampling that promotes the presence of ambiguous words in the English captions with respect to their meaning and thus their Hindi translation. The final corpus statistics of the "Hindi Visual Genome" are in Table 1.

The paper is organized as follows: In Section 2, we survey related multi-modal multi-lingual datasets. Section 3 describes the way we selected and prepared the training set. Section 4 is devoted to the challenge test set: the method to find ambiguous words and the steps taken when constructing the test set, its final statistics and a brief discussion of our observations. We conclude in Section 5.

2 Related Work

Multi-modal neural machine translation is an emerging area where translation takes more than text as input. The additional information sources could be features from images or sound. Combining visual features with language modeling has shown better result for image captioning and question answering [14, 21, 12].

Many experiments were carried out considering images to improve machine translation, i.a. for resolving ambiguity due to different senses of words in different contexts. One of the starting points were multi-modal shared tasks at WMT [19, 7, 2] organized around “Flickr30k” [8], a multi-modal multi-lingual (English-German, English-French, and English-Czech) dataset.

[4] proposed a multi-modal NMT system using image feature for Hindi-English language pair. Due to the lack of English-Hindi multi-modal data, they used a synthetic training dataset and manually curated development and test sets for Hindi derived from the English part of Flickr30k corpus [17]. [1] proposed a probabilistic method using pictures for word prediction constrained to a narrow set of choices, such as possible word senses. Their results suggest that images can help word sense disambiguation.

Different techniques then followed, using various neural network architectures for extracting and using the contextual information. One of the approaches was proposed by [11] for multi-modal translation by replacing image embedding with an estimated posterior probability prediction for image categories.

3 Training and Test Set Preparations

To produce the main part of our corpus, we have automatically translated and manually post-edited the English captions of “Visual Genome” corpus into Hindi.

The starting point were 31,525 randomly selected images from Visual Genome. Of all the English-captioned regions available for each of the images, we randomly select one. To obtain the Hindi translation, we have followed these steps:

1. We translated all 31,525 captions into Hindi using the NMT model (Tensor-to-Tensor, [20]) specifically trained for this purpose as described in [16].
2. We uploaded the image, the source English caption and its Hindi machine translation into a “Translation Validation Website”,[1] which we designed as a simple interface for post-editing the translations. One important feature was the use of a Hindi on-screen keyboard[2] to enable proper text input even for users with limited operating systems.
3. Our volunteers post-edited all the Hindi translations. The volunteers were selected based on their Hindi language proficiency.
4. We manually verified and finalized the post-edited files to obtain the training and test data.

To facilitate comparison of future experiments with this dataset, we create a dedicated development and test set. The split of the 31,525 items into the training, development and test sets with sizes as listed in Table 1 was random.

4 Challenge Test Set Preparations

In addition to the randomly selected 31,525 items described above, we prepared a challenge test set of 1,400 segments which are more likely to benefit from visual information for word sense disambiguation. To achieve this targeted selection, we first found the most ambiguous words from the whole “Visual Genome” corpus and then extracted segments containing these words. The overall steps for obtaining the ambiguous words are shown in Fig. 2.

The detailed sequence of processing steps was as follows:

1. Translate all English captions from the Visual Genome dataset (3.15 millions unique strings) using a baseline machine translation system into Hindi, obtaining a synthetic parallel corpus. In this step, we used Google Translate.

[1] `http://ufallab.ms.mff.cuni.cz/~parida/index.html`

[2] `https://hinkhoj.com/api/`

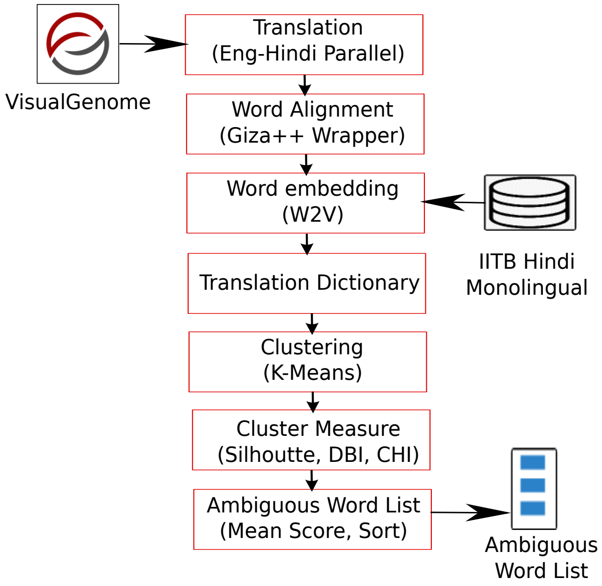

VisualGenome
Translation (Eng-Hindi Parallel)
Word Alignment (Giza++ Wrapper)
Word embedding (W2V)
IITB Hindi Monolingual
Translation Dictionary
Clustering (K-Means)
Cluster Measure (Silhoutte, DBI, CHI)
Ambiguous Word List (Mean Score, Sort)
Ambiguous Word List

Fig. 2. Overall pipeline for ambiguous word finding from input corpus

2. Apply word alignment on the synthetic parallel corpus using GIZA++ [15], in a wrapper[3] that automatically symmetrizes two bidirectional alignments; we used the intersection alignment.

3. Extract all pairs of aligned words in the form of a “translation dictionary”. The dictionary contains key/value pairs of the English word (E) and all its Hindi translations ($H_1, H_2, \ldots H_n$), i.e. it has the form of the mapping $E \mapsto \{H_1, ..., H_n\}$.

4. Train Hindi word2vec (W2V) [13] word embeddings. We used the gensim[4] [18] implementation and trained it on IITB Hindi Monolingual Corpus[5] which contains about 45 million Hindi sentences. Using such a large collection of Hindi texts improves the quality of the obtained embeddings.

5. For each English word from the translation dictionary (see Step 3), get all Hindi translation words and their embeddings (Step 4).

6. Apply K-means clustering algorithm to the embedded Hindi words to organize them according to their word similarity.

   If we followed a solid definition of word senses and if we knew how many senses there are for a given source English word and how they match the meanings of the Hindi words, the K would correspond to the number of Hindi senses that the original English word expresses. We take the pragmatic approach and apply K-means for a range of values (K from 2 to 6).

7. Evaluate the obtained clusters with the Silhouette Score, Davies-Bouldin Index (DBI), and Calinski-Harabaz Index (CHI) [6, 5]. Each of the selected scores reflects in one way or another the cleanliness of the clusters, their separation. For the final sorting (Step 8), we mix these scores using a simple average function.

   The rationale behind using these scores is that if the word embeddings of the Hindi translations can be clearly clustered into 2 or more senses, then the meaning distinctions are big enough to indicate that the original English word was ambiguous. The exact *number* of different meanings is not too important for our purpose.

8. Sort the list in descending order to get the most ambiguous words (as approximated by the mean of clustering measures) at the top of the list.

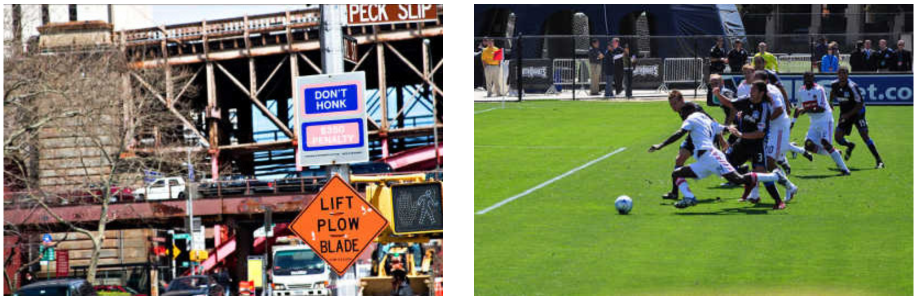
9. Manually check the list to validate that the selected ambiguous words indeed potentially need an image to disambiguate them. Select a cutoff and extract the most ambiguous English words.

The result of this semi-automatic search and manual validation of most ambiguous words was a list of 19 English words, see Table 2.

[3] https://github.com/ufal/qtleap/blob/master/cuni_train/bin/gizawrapper.pl

[4] https://radimrehurek.com/gensim/tut1.html

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



Street sign advising of penalty

The penalty box is white lined

Fig. 3. An illustration of two meanings of the word "penalty" exemplified with two images

Table 2. Challenge test set: distribution of the ambiguous words

	Word	Segment Count
1	Stand	180
2	Court	179
3	Players	137
4	Cross	137
5	Second	117
6	Block	116
7	Fast	73
8	Date	56
9	Characters	70
10	Stamp	60
11	English	42
12	Fair	41
13	Fine	45
14	Press	35
15	Forms	44
16	Springs	30
17	Models	25
18	Forces	9
19	Penalty	4
	Total	1400

For each of these words, we selected and extracted a number of items available in the original Visual Genome and provided the same manual validation of the Hindi translation as described in Section 3 for the training and regular test sets. We tried to make a balance and the frequencies of the ambiguous words in the challenge test set roughly correspond to the original frequencies in Visual Genome.

Incidentally, 7 images and English captions occur in both the training set and the challenge test set.[6] The overlap in images (but using different regions and captions) is larger: 359.

Fig. 3 illustrates two sample items selected for the word "penalty" (Hindi translation omitted here). We see that for humans, the images are clearly disambiguating the meaning of the word: the fine to be paid for honking vs. the kick in a soccer match.

Arguably, the surrounding English words in the source segments (e.g. "street" vs. "white lined") can be used by machine translation systems to pick the correct translation even without access to the image. The size of the original dataset of images with captions however did not allow us to further limit the selection to segments where the text alone is not sufficient for the disambiguation.

5 Conclusion and Future Work

We presented a multi-modal English-to-Hindi dataset. To the best of our knowledge, this

[6] The English segments appearing in both the training data and the challenge test set are: A round concert block, Man stand in crane, Street sign on a pole in english and chinese, a fast moving train, a professional tennis court, bird characters on top of a brown cake, players name on his shirt.

is the first such dataset that includes an Indian language. The dataset can serve e.g. in Hindi image captioning but our primary intended use case was research into the employment of images as additional input to improve machine translation quality.

To this end, we created also a dedicated challenge test set with text segments containing ambiguous words where the image can help with the disambiguation.

All parts of our dataset served in WAT 2019[7] shared task on multi-modal translation.[8]

We illustrated that the text-only information in the surrounding words could be sufficient for the disambiguation. One interesting research direction would be thus to ignore all the surrounding words and simply ask: given the image, what is the correct Hindi translation of this ambiguous English word. Another option we would like to pursue is to search larger datasets for cases where even the whole segment does not give a clear indication of the meaning of an ambiguous word.

Our "Hindi Visual Genome" is available for research and non-commercial use under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 License[9] at `http://hdl.handle.net/11234/1-2997`.

Acknowledgments

We are grateful to Vighnesh Chenthil Kumar, a summer intern from IIIT Hyderabad at Charles University for his help with the semi-automatic search for the most ambiguous words. The work was carried out during Shantipriya Parida's post-doc funded by Charles University.

This work has been 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

1. **Barnard, K. & Johnson, M. (2005).** Word sense disambiguation with pictures. *Artificial Intelligence*, Vol. 167, No. 1-2, pp. 13–30. DOI: 10.1016/j.artint.2005.04.009.

2. **Barrault, L., Bougares, F., Specia, L., Lala, C., Elliott, D., & Frank, S. (2018).** Findings of the third shared task on multimodal machine translation. *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pp. 304–323. DOI: 10.18653/v1/W18-6402.

3. **Belz, A., Erdem, E., Pastra, K., & Mikolajczyk, K.**, editors **(2017).** *Proceedings of the Sixth Workshop on Vision and Language, VL@EACL17, Valencia, Spain,.* Association for Computational Linguistics. DOI: 10.18653/v1/W17-20.

4. **Chowdhury, K. D., Hasanuzzaman, M., & Liu, Q. (2018).** Multimodal neural machine translation for low-resource language pairs using synthetic data. *ACL 2018*, pp. 33. DOI: 10.18653/v1/W18-3405.

5. **Davies, D. L. & Bouldin, D. W. (1979).** A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. PAMI-1, No. 2, pp. 224–227. DOI: 10.1109/TPAMI.1979.4766909.

6. **de Amorim, R. C. & Hennig, C. (2015).** Recovering the number of clusters in data sets with noise features using feature rescaling factors. *Information Sciences*, Vol. 324, pp. 126–145. DOI: 10.1016/j.ins.2015.06.039.

7. **Elliott, D., Frank, S., Barrault, L., Bougares, F., & Specia, L. (2017).** Findings of the second shared task on multimodal machine translation and multilingual image description. *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, Association for Computational Linguistics, Copenhagen, Denmark, pp. 215–233. DOI: 10.18653/v1/W17-4718.

8. **Elliott, D., Frank, S., Sima'an, K., & Specia, L. (2016).** Multi30k: Multilingual english-german image descriptions. *arXiv preprint arXiv:1605.00459*. DOI: 10.18653/v1/W16-3210.

9. **Elliott, D. & Kádár, Á. (2017).** Imagination improves multimodal translation. *CoRR*, Vol. abs/1705.04350.

[7] `http://lotus.kuee.kyoto-u.ac.jp/WAT/WAT2019/index.html`

[8] `https://ufal.mff.cuni.cz/hindi-visual-genome/wat-2019-multimodal-task`

[9] `https://creativecommons.org/licenses/by-nc-sa/4.0/`

10. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, Vol. 123, No. 1, pp. 32–73.

11. Lala, C., Madhyastha, P., Wang, J., & Specia, L. (2017). Unraveling the contribution of image captioning and neural machine translation for multimodal machine translation. *The Prague Bulletin of Mathematical Linguistics*, Vol. 108, No. 1, pp. 197–208. DOI: : 10.1515/pralin-2017-0020.

12. Liu, C., Sun, F., Wang, C., Wang, F., & Yuille, A. L. (2017). MAT: A multimodal attentive translator for image captioning. *CoRR*, Vol. abs/1702.05658.

13. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *CoRR*, Vol. abs/1301.3781.

14. Mostafazadeh, N., Brockett, C., Dolan, B., Galley, M., Gao, J., Spithourakis, G. P., & Vanderwende, L. (2017). Image-grounded conversations: Multimodal context for natural question and response generation. *CoRR*, Vol. abs/1701.08251.

15. Och, F. J. & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, Vol. 29, No. 1, pp. 19–51. DOI: 10.1162/089120103321337421.

16. Parida, S. & Bojar, O. (2018). Translating Short Segments with NMT: A Case Study in English-to-Hindi. *Proceedings of EAMT 2018*.

17. Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., & Lazebnik, S. (2015). Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649.

18. Řehůřek, R. & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, ELRA, Valletta, Malta, pp. 45–50. `http://is.muni.cz/publication/884893/en`. DOI: 10.13140/2.1.2393.1847.

19. Specia, L., Frank, S., Sima'an, K., & Elliott, D. (2016). A shared task on multimodal machine translation and crosslingual image description. *Proceedings of the First Conference on Machine Translation*, Association for Computational Linguistics, Berlin, Germany, pp. 543–553. DOI: 10.18653/v1/W16-2346.

20. Vaswani, A., Bengio, S., Brevdo, E., Chollet, F., Gomez, A. N., Gouws, S., Jones, L., Kaiser, L., Kalchbrenner, N., Parmar, N., Sepassi, R., Shazeer, N., & Uszkoreit, J. (2018). Tensor2tensor for neural machine translation. *CoRR*, Vol. abs/1803.07416.

21. Yang, L., Tang, K. D., Yang, J., & Li, L. (2016). Dense captioning with joint inference and visual context. *CoRR*, Vol. abs/1611.06949.

22. Zhou, M., Cheng, R., Lee, Y. J., & Yu, Z. (2018). A visual attention grounding neural model for multimodal machine translation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3643–3653. DOI: 10.18653/v1/D18-1400.

Article received on 22/12/2018; accepted on 05/03/2019.
Corresponding author is Ondřej Bojar.