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

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

[1] Charles University, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics, Malostranské naměstí 25, 118 00
Prague, Czech Republic
{parida,bojar}@ufal.mff.cuni.cz
[2] KIIT University, Bhubaneswar-24, India
sdashfca@kiit.ac.in

Abstract. Visual Genome is a dataset connecting structured image information with English language. We present "Hindi Visual Genome", a multimodal dataset consisting of text and images suitable for English-Hindi multimodal machine translation task and multimodal research. We have selected short English segments (captions) from Visual Genome along with associated images and automatically translated them to Hindi with manual post-editing, taking the associated images into account. We prepared a set of 31599 segments, accompanied by a challenge test set of 1400 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 for multimodal 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.

Keywords: Visual Genome, Multimodal Corpus, Parallel Corpus, Word Embedding, Neural Machine Translation (NMT), Image Captioning

1 Introduction

Multimodal content is gaining popularity in machine translation (MT) community due to its appealing chances to improve translation quality and its usage in commercial applications such as image caption translation for online news articles or machine translation for e-commerce product listings [11, 4, 8, 20]. 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 forms in Hindi based on the context (e.g. either a computer mouse or a small rodent).

* Corresponding author

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	30,000
Development Test Set (D-Test)	1,000
Evaluation Test Set (E-Test)	599
Challenge Test Set (C-Test)	1,400

There is a limited number of multimodal datasets available and even fewer of them are also multilingual. Our aim is to extend the set of languages available for multimodal experiments by adding a Hindi variant of a subset of Visual Genome.

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.

The main portion of our Hindi Visual Genome is intended for training purposes of tools like multimodal 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 multimodal multilingual datasets. Section 3 describes the way we selected and prepared the training set. Section 4 is devoted to the test set: the method to find ambiguous words and the steps taken when constructing the test set. Section 5 provides various statistics on the data and Section 6 discusses our observations. We conclude in Section 7.

Creating such a dataset enables researchers multimodal experimenting with Hindi for various applications and also facilitates the exploration of how the language is grounded in vision.

2 Related Work

Multimodal neural machine translation is an emerging area where translation takes more than text as input. It also uses features from image or sound for generating the translated text. Combining visual features with language modeling has shown better result for image captioning and question answering [14, 19, 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 is “Flickr30k” [7], a multilingual (English-German, English-French, and English-Czech) shared task based on multimodal translation was part of WMT 2018 [3]. [5] proposed a multimodal NMT system using image feature for Hindi-English language pair. Due to the lack of English-Hindi multimodal 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]. [2] 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 multimodal translation by replacing image embedding with an estimated posterior probability prediction for image categories.

3 Training 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 31599 randomly selected images from Visual Genome, with one English-captioned region each. To obtain the Hindi translation, we have followed these steps:

1. We translated all 31599 captions into Hindi using the NMT model (Tensor-to-Tensor, [18]) 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”,[3] which we designed as a simple interface for post-editing the translations. One important feature was the use of a Hindi on-screen keyboard[4] 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.

4 Challenge Test Set Preparations

In addition to the randomly selected 31599 items described above, we prepared a challenge test set of 1400 segments which need images for word sense disambiguation. To achieve this targeted selection, we first found the most ambiguous words from the “Visual Genome” corpus and then extracted segments containing the most ambiguous words. The overall steps for obtaining the ambiguous words are shown in Fig. 1.

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

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

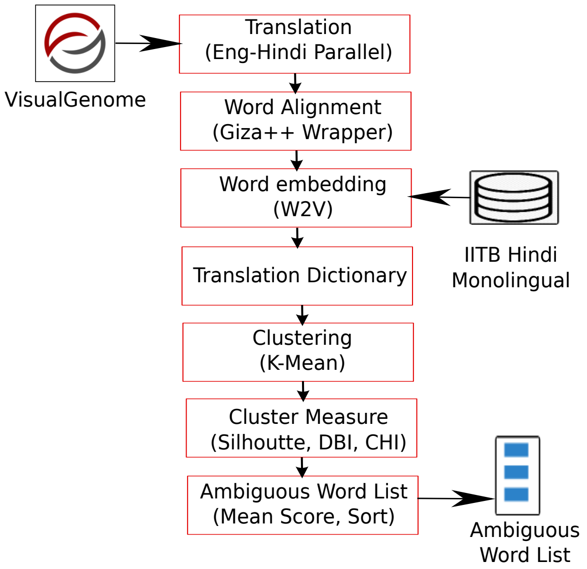

VisualGenome
Translation (Eng-Hindi Parallel)
Word Alignment (Giza++ Wrapper)
Word embedding (W2V)
IITB Hindi Monolingual
Translation Dictionary
Clustering (K-Mean)
Cluster Measure (Silhoutte, DBI, CHI)
Ambiguous Word List (Mean Score, Sort)
Ambiguous Word List

Fig. 1: Overall Pipeline for Ambiguous Word Finding from input Corpus.

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 systems into Hindi, obtaining a synthetic parallel corpus. In our experiment, we used Google Translate.
2. Apply word alignment on the synthetic parallel corpus using GIZA++ [15], in a wrapper[5] 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 ($H_1, H_2, \dots H_n$), 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[6] implementation and trained on IITB Hindi Monolingual Corpus[7] which contains about 45 million Hindi sentences. Using such a large collection of Hindi text 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.

[5] https://github.com/ufal/qtleap/blob/master/cuni_train/bin/gizawrapper.pl
[6] https://radimrehurek.com/gensim/tut1.html
[7] http://www.cfilt.iitb.ac.in/iitb_parallel/iitb_corpus_download/

If we followed a solid definition of word senses and if we knew how many 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). See also the discussion in Section 5.

7. Evaluate the obtained clusters with the Silhouette Score, Davies-Bouldin Index (DBI), and Calinski-Harabaz Index (CHI) [1, 6]. 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 Hindi words.

5 Dataset Statistics

Table 2 shows a sample of cluster measures for a range of clusters. It also highlights the best score (selected) among the range of clusters for the cluster measurement techniques (“Silhouette “DBI”, and “CHI”).

Table 2: Cluster Measure for the Range (K) for the Measurement Techniques. The Highlighted Values are the Best Cluster Measure Scores for the Cluster Measurement Techniques.

Sl No.	Word	K cluster	Silhoutte max	DBI min	CHI max
1	Water	2	.037	.839	2.457
2		3	.046	.355	**2.515**
3		4	.034	.175	2.180
4		5	.033	.119	2.184
5		6	**.049**	**.105**	2.130
6	Galloping	2	**.194**	.148	2.087
7		3	.123	.172	2.253
8		4	.128	.076	2.333
9		5	.118	.020	2.446
10		6	.036	**.016**	**2.468**

Table 3 shows the ambiguous word list along with cluster measure. We have restricted to a small number for display purpose.

Table 3: Ambiguous Word List based on Cluster Measure. The Min Score is Mean Value of the Clustering Techniques

Sl No.	Word	Silhoutte	DBI	CHI	Min
1	Stump	.089	.014	1.714	**0.606**
2	Single	.162	.034	2.853	**1.016**
3	Elderly	.180	.023	2.640	**0.948**
4	Bushes	.080	.065	2.446	**0.863**
5	Gas	.111	.011	2.366	**0.829**

The sample Hindi word embedding for the English word is shown in Fig. 2.

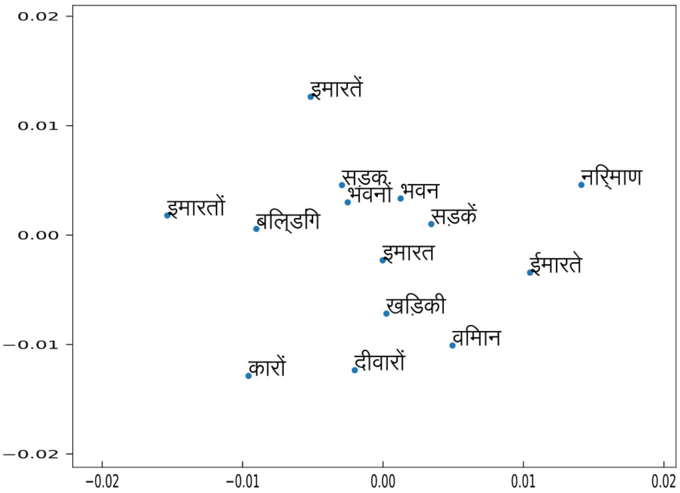


Fig. 2: Hindi Word Embedding for the English Word "Building"

To find the ideal K value for clustering, we tried to plot the elbow as shown in 3. But, there was no clear K value for defining the cluster. So, we focused on cluster measure index/score.

Table 4: Most Ambiguous Words in Sorted Order

Sl No.	Word	Min Score
1	Start	4.539
2	Three	3.553
3	Gathering	2.926
4	Hiding	2.814
5	Up	2.611
6	hide	2.572
7	shown	2.346
8	under	2.319
9	runs	2.269
10	most	2.209

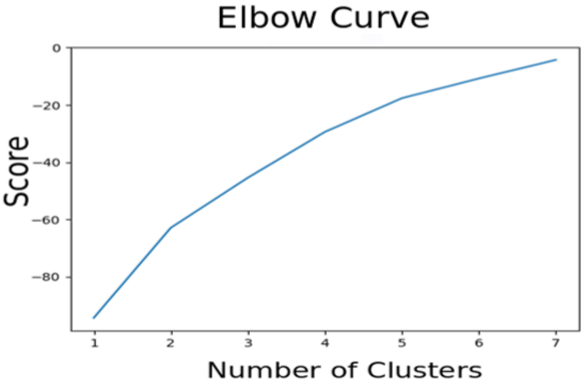


Fig. 3: Elbow Curve for a defined range.

5.1 Dataset Description

The number of segments of the test set distributed based on the occurrence of ambiguous words shown in the Table 5.

Table 5: Test set Distribution based on Ambiguous Word list

Sl No.	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

We tried to make a balance and challenging test set covering ambiguous words based on their occurrences.

6 Analysis and Discussion

We have taken a sample of 30 English words and their Hindi clusters and compare with Both BableNet[8] and IndoWordNet[9] as shown in the Fig. 4.

[8] http://live.babelnet.org/

[9] http://www.cfilt.iitb.ac.in/indowordnet/

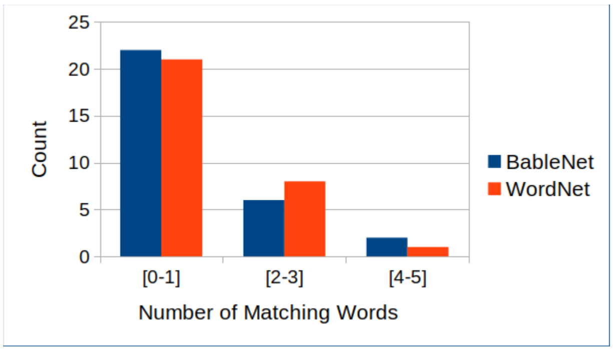


Fig. 4: Number of Matching Words with BableNet and IndoWordNet.

The manual cluster analysis results are shown in the Fig. 5

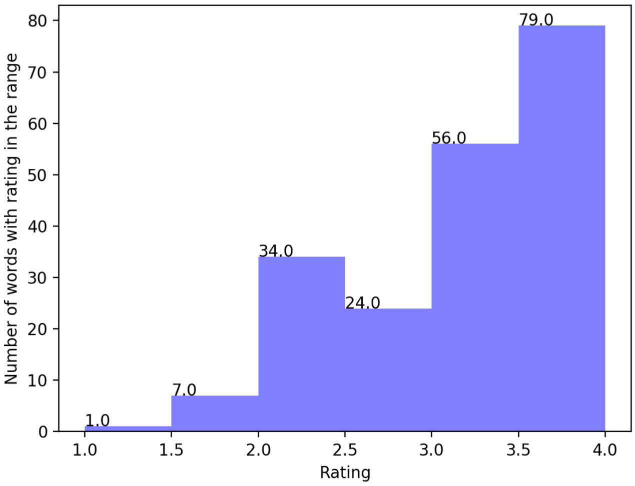
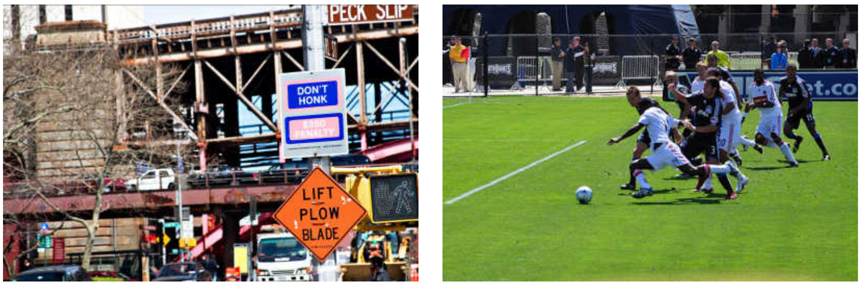


Fig. 5: Manual Cluster Analysis and Rating Graph.

Manually we evaluated a list of 444 ambiguous words along with images and their captions. Our main focus was to find such words which have different senses found in the corresponding images and their captions. For example, "Penalty" which have two senses such as i) Fine, ii) Football kick and the associated captions for the images in the Fig. 6.

The list of ambiguous words based on the manual evaluation of senses and images are shown in the Table 6.



(a) Street Sign Advising of Penalty. (b) The Penalty Box is White Lined.

Fig. 6: Manual Analysis of Ambiguous Words (senses) with Images and Captions

Table 6: Most Ambiguous Words as per Manual Evaluation where Image may help Disambiguate

Sl No.	Word
1	Penalty
2	Block
3	Press
4	Characters
5	Cross
6	Second
7	Fine
8	English
9	Players
10	Stand
11	Court
12	Stamp
13	Fast
14	Fair
15	Models
16	Date
17	Forms
18	Forces
19	Springs

From the ambiguous words from the Table 6, 13 (approx. 70%) words appeared in the first 222 (50%) and 6 (approx. 30%) words in the next 222 (50%) of the evaluated list of 444.

7 Conclusion and Future Work

We presented the multimodal dataset for English-to-Hindi machine translation which is first in its kind to best of our knowledge. It will help the researcher in multimodal research and investigate the usage of the image as input to improve translation quality.

We will release the final version through the LINDAT repository[10], an infrastructure for sharing large and small resources related to language processing. Our release “Hindi Visual Genome” is available for research and non-commercial use under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 License.[11]

Our future work include i) Carry out English-to-Hindi multimodal experiment using NMT model “NeuralMonkey” [9] ii) Enrich the dataset with more images, regions, and captions. iii) Perform a comparative analysis with similar NMT experiment with images.

8 Acknowledgments

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 LM2015-071 and OP VVV VI CZ.02.1.01/0.0/0.0/16013/0001781).

References

1. de Amorim, R.C., Hennig, C.: Recovering the number of clusters in data sets with noise features using feature rescaling factors. Information Sciences **324**, 126–145 (2015)
2. Barnard, K., Johnson, M.: Word sense disambiguation with pictures. Artificial Intelligence **167**(1-2), 13–30 (2005)
3. Barrault, L., Bougares, F., Specia, L., Lala, C., Elliott, D., Frank, S.: Findings of the third shared task on multimodal machine translation. In: Proceedings of the Third Conference on Machine Translation: Shared Task Papers. pp. 304–323 (2018)
4. Belz, A., Erdem, E., Pastra, K., Mikolajczyk, K. (eds.): Proceedings of the Sixth Workshop on Vision and Language, VL@EACL 2017, Valencia, Spain, April 4, 2017. Association for Computational Linguistics (2017), https://aclanthology.info/volumes/proceedings-of-the-sixth-workshop-on-vision-and-language
5. Chowdhury, K.D., Hasanuzzaman, M., Liu, Q.: Multimodal neural machine translation for low-resource language pairs using synthetic data. ACL 2018 p. 33 (2018)
6. Davies, D.L., Bouldin, D.W.: A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence **PAMI-1**(2), 224–227 (April 1979). https://doi.org/10.1109/TPAMI.1979.4766909
7. Elliott, D., Frank, S., Sima’an, K., Specia, L.: Multi30k: Multilingual english-german image descriptions. arXiv preprint arXiv:1605.00459 (2016)
8. Elliott, D., Kádár, Á.: Imagination improves multimodal translation. CoRR **abs/1705.04350** (2017), http://arxiv.org/abs/1705.04350

[10] http://www.lindat.cz

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

9. Helcl, J., Libovický, J.: Neural monkey: An open-source tool for sequence learning. The Prague Bulletin of Mathematical Linguistics (107), 5–17 (2017). https://doi.org/10.1515/pralin-2017-0001, `http://ufal.mff.cuni.cz/pbml/107/art-helcl-libovicky.pdf`
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.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision **123**(1), 32–73 (2017)
11. Lala, C., Madhyastha, P., Wang, J., Specia, L.: Unraveling the contribution of image captioning and neural machine translation for multimodal machine translation. The Prague Bulletin of Mathematical Linguistics **108**(1), 197–208 (2017)
12. Liu, C., Sun, F., Wang, C., Wang, F., Yuille, A.L.: MAT: A multimodal attentive translator for image captioning. CoRR **abs/1702.05658** (2017), `http://arxiv.org/abs/1702.05658`
13. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. CoRR **abs/1301.3781** (2013), `http://arxiv.org/abs/1301.3781`
14. Mostafazadeh, N., Brockett, C., Dolan, B., Galley, M., Gao, J., Spithourakis, G.P., Vanderwende, L.: Image-grounded conversations: Multimodal context for natural question and response generation. CoRR **abs/1701.08251** (2017), `http://arxiv.org/abs/1701.08251`
15. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Computational Linguistics **29**(1), 19–51 (2003)
16. Parida, S., Bojar, O.: Translating Short Segments with NMT: A Case Study in English-to-Hindi. In: Proceedings of EAMT 2018 (2018)
17. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE international conference on computer vision. pp. 2641–2649 (2015)
18. 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.: Tensor2tensor for neural machine translation. CoRR **abs/1803.07416** (2018), `http://arxiv.org/abs/1803.07416`
19. Yang, L., Tang, K.D., Yang, J., Li, L.: Dense captioning with joint inference and visual context. CoRR **abs/1611.06949** (2016), `http://arxiv.org/abs/1611.06949`
20. Zhou, M., Cheng, R., Lee, Y.J., Yu, Z.: A visual attention grounding neural model for multimodal machine translation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 3643–3653 (2018)