Practical Parallel Data Collection for Low-Resource Languages via Images

Anonymous authors
Paper under double-blind review

Abstract

We propose a method of curating high-quality parallel training data for low-resource languages without requiring that the annotators are bilingual. Our method involves using a carefully selected set of images as a pivot between the source and target languages, by getting captions for such images in both languages independently. Human evaluations on the English-Hindi parallel corpora created with our method show that 81.1% of the pairs are acceptable translations, and only 2.47% of the pairs are not a translation at all. We further show the utility of our method by obtaining reasonable performance on two downstream tasks – machine translation and dictionary extraction. All code and data will be released upon acceptance.

1 Introduction

Machine translation (MT) is a natural language processing task that aims to automatically translate text from a source language into a target language. Current state-of-the-art methods for MT are based on neural network architectures (Barrault et al., 2019), and often require large parallel corpora (i.e., the same text in two or more languages) for training.

Creating parallel data between a source and target language usually requires bilingual translators who are fluent in both languages. While there are a few language pairs for which translators are easily available, for many low-resource languages, it is challenging to find *any* speakers sufficiently proficient in both the source and the target language. We propose a method to create data for training machine translation systems in such situations. Our method relies on obtaining captions for images depicting universally relevant concepts. The captions are collected individually in each of the source and target languages, removing the requirement for the human annotators to be well-versed in both languages. The captions for the same image are then paired to create training data.

Notably, as the captions are developed independently for each language, our method creates data that are *comparable*, rather than strictly parallel. Nevertheless, comparable corpora have been proven very useful for machine translation and related applications (Munteanu et al., 2004; Abdul-Rauf & Schwenk, 2009; Irvine & Callison-Burch, 2013). In this work, we evaluate the utility of our collected data as a replacement for parallel corpora in two ways:

- We show that bilingual speakers (of the source and target languages) adjudicate the dataset as containing over 81% acceptable translation pairs.
- We demonstrate reasonable performance when using the dataset to train models for two downstream tasks: unsupervised dictionary induction and machine translation.

Moreover, we also compare our method to the traditional process of parallel corpus creation and show that it is significantly more cost-efficient.

We apply our proposed method and evaluation techniques to a specific language pair: Hindi as the source language and English the target. Hindi is chosen as a testbed because, although it has over 520 million speakers (Chandramouli & General, 2011), far fewer parallel corpora exist as compared to other widely spoken languages like French, German, Spanish, etc. (Arivazhagan et al., 2019). We note that our annotation tools and methodology are not language-specific and can be easily adapted to any very low-resource setting.

Figure 1: A concise, universal concept (left) and a complex culture-specific activity (right).

2 Methodology

Overview Our method requires a set of N images $\mathbb{I} = \{\mathrm{I}_1, \mathrm{I}_2, ..., \mathrm{I}_N\}$. Each image I_i is associated with a set of P captions $\mathbb{C}_i^{en} = \{C_{i,1}^{en}, C_{i,2}^{en}, ..., C_{i,P}^{en}\}$ in the target language en. The P captions are provided by different annotators.

For each image i, Q annotators are then asked to provide captions in the source language fr, yielding $\mathbb{C}_i^{fr} = \{C_{i,1}^{fr}, C_{i,2}^{fr}, ..., C_{i,Q}^{fr}\}$. We experiment with two different ways of obtaining comparable data $\mathbb{D}_{en-fr}$ from the sets $\mathbb{C}_i^{fr}$ and $\mathbb{C}_i^{en}$. In the first method , we take Cartesian product of the set of captions in the two languages.

$$\mathbb{D}_{en-fr} = \bigcup_{i=1}^{N} \mathbb{C}_i^{en} \bigotimes \mathbb{C}_i^{fr} \tag{1}$$

$$\text{Where } \mathbb{C}_i^{en} \bigotimes \mathbb{C}_i^{fr} = \{(C_{i,j}^{en}, C_{i,k}^{fr}) : j \in [1, P], k \in [1, Q]\} \tag{2}$$

Thus, yielding $P * Q$ *comparable* sentences per image. A second way is to randomly assign each of the sentences in source set to the target set, yielding $min(P, Q)$ comparable sentences per image. We refer to these methods as *cross* and *random* assignment respectively.

Selecting the Right Images As mentioned in the introduction, our method requires images that can attract consistent and succinct captions on universal topics. We obtain such a set with a two-step approach. First, we select a dataset of images that depict fairly general themes, because some image sources may not be ideal for our proposed method. For example, images sourced from news articles typically involve a large amount of context and will not be a good fit for this task (Hodosh et al., 2013), or images that show a specific entity that's only popular within a certain geographical or cultural context (like a celebrity) cannot be used for our task.

After selecting the right set of images, we select a subset of those images that have consistent captions. We quantify this notion of caption consistency by calculating a *caption diversity score* d_i for each image i. For an image i, d_i calculates the syntactic variance in its set of captions $\mathbb{C}_i^{en}$. We define $d_i = l_i + w_i + e_i$ where l_i, w_i, and e_i are respectively the total length, the total number of unique words, and the total pairwise edit distance for the captions included in $\mathbb{C}_i^{en}$. In other words, the score helps in finding images with short captions, with fewer unique words and consistent captions provided by different annotators.

For example, consider the images shown in Figure 1 taken from the Flickr8k dataset (Hodosh et al., 2013). The image on the right (with a high caption diversity score) depicts a complex activity and is bound to attract varied captions. This becomes clear from the captions for this image from the dataset: "A holder and kicker for a football team dressed in orange, white and black play while onlookers behind them watch." and "Two young men on the same football team are wearing orange and white uniform and playing on an outside field while coach and other players watch." On the other hand, the image on the left shows a simple, universal concept. The captions for the image are almost identical, and are simple

English Captions from Flickr8k
A bald , shirtless man rock climbing .
A bald man climbing rocks .
A man climbing up a rocky cliff
A man with no shirt on is rock climbing .
A rock climber scales a mountain .

Crowdsourced Hindi Captions
एक आदमी बिना शर्ट पहने चट्टान पर चढ़ रहा है
कुछ लोग पहाड़ी पर चढ़ रहे हैं
एक आदमी पहाड़ पर चढ़ रहा है |
एक आदमी पहाड़ी पे ट्रैकिंग करता हुआ
एक आदमी पहाड़ पर चढ़ाई कर रहा है

Translated Hindi Captions (for reference)
A man is climbing a rock without wearing a shirt
Some people are climbing the hill
A man is climbing a mountain.
A man trekking up a hill
A man is climbing a mountain

Figure 2: An Example Image and the Corresponding English (part of the dataset) and Hindi (generated by the workers using only the image) captions.

variations of "A black dog is running in the water." We focus on such images as they will likely get consistent captions in any language, leading to higher quality comparable data.

3 Experiments and Results

Selecting Suitable Images For our experiments, we use the Flickr8k dataset (Hodosh et al., 2013) as it is comprised of images that show simple, fairly universal concepts. The Flickr8k dataset has 8000 images, and each image has $P = 5$ English captions. We first select 700 images with the lowest caption diversity scores (see section 2). Then, we manually prune the set of images to retain $N = 500$ images to maximize the number of different concepts covered by the images. We prune by deleting consecutive images that cover similar concepts. For example, if five consecutive images portray "horse riding", only one of them is retained.

Obtaining Image Captions We ask five different crowd workers, sourced via Amazon Mechanical Turk, to describe each of the images in the source language. The crowd workers need to be fluent *only* in the source language and no information in the target language is supplied or requested.

For our experiments, we used Hindi as the source language. Apart from being located in India, the workers were not required to have any other skills. Following Hodosh et al. (2013), we provide the annotators with guidelines (Table 2 in the appendix) that are conducive to reducing the variance in image captions. We obtain $Q = 5$ captions per image, for a total of 2500 captions for the 500 images. For quality control, the authors who are native Hindi speakers manually verified all the captions. Note that our method *requires no resources in the source language* apart from the instructions for the annotators. We make no assumptions specific to Hindi in our setup, and the method can be adopted for any other language.

3.1 Evaluating Data Quality

Manual Evaluation Out of the 500 images selected for the experiment, we randomly chose 120 for a final human evaluation on the quality of the data obtained. We used comparable data obtained via the *cross* method (see Section 2). Therefore, a total of $120 * 25 = 3000$ sentence pairs were evaluated. The annotators were asked to rate the translation quality on a Likert scale (Likert, 1932) between 1-5. Table 1 shows the instructions for the evaluators and the % of translation falling under each quality category.

Quality	Criteria	%	Cum. %
Perfect	The translation is flawless.	16.45	16.45
Good	The translation is good. The differences between a perfect and a good translation are not very important to the meaning of the source sentence	35.22	51.67
Acceptable	The translation conveys the meaning adequately but can be improved	29.43	81.10
Bad	The translation conveys the meaning to some degree but is a bad translation	16.42	97.52
Not a translation	There is no relation whatsoever between the source and the target sentence	2.47	100.00

Table 1: Detailed Instructions for the Evaluators and Quality Evaluation Results

As Table 1 shows, over 81% of the parallel sentences were cumulatively rated to be acceptable or better. More importantly, only 2.47% of sentences were rated as not being translations at all, and 16% were deemed perfect translations. For a low resource setting, this means that the data created with our proposed method can be used without any further pruning. An example image and the corresponding captions are shown in Figure 2, and additional examples can be found in the appendix.

3.2 Evaluating Quality of Comparable Data on Downstream Tasks

We present two different tasks to measure the quality of our data. We use the *random* variant of the dataset for both the experiments (Section 2).

Translation Downstream Task We use a transformer based Neural Machine Translation System for all translation experiments (Vaswani et al., 2017). Each transformer has 4 attention heads with a 512 dimensional embedding layer and hidden state. Dropout (Srivastava et al., 2014) with a 0.3 probability is used in all layers. We use BPE tokenization (Sennrich et al., 2015) with a vocabulary of size 8000 for all the experiments.

We use the Ted Hindi-English parallel corpus (Kunchukuttan et al., 2017) in conjunction with our dataset (Ours) to measure the potential improvements when additionally using our collected data. The Ted data has pre-defined training and test splits with 18798 and 1243 sentences, respectively. We create training and test splits for Ours with 2220 training and 248 test sentences. We train two models: i) $\text{Model}_{\text{Ted + Ours}}$ trained on a combination of our data (Ours) and the ted data (Ted), and ii) $\text{Model}_{\text{Ted}}$ trained only on the Ted data. We evaluate our models using BLEU (Papineni et al., 2002) scores on varying amounts of training data on the two test sets, a standard metric used by the MT community which captures the syntactic similarity between the expected and the predicted translations.

As Figure 3 shows, when a limited amount of Ted data is used, adding Ours leads to a 5x improvement in the BLEU scores (0.07 vs. 0.36) for Ted test set. The improvements are also observed with higher amounts of Ted data, showing that Ours data is indeed helpful for machine translation tasks. For the Ours test set, models trained with the Ours data consistently outperform the models trained with Ted data only.

Unsupervised Dictionary Extraction The task of unsupervised dictionary extraction from parallel data aims at creating a table containing pairs of translated words and phrases. Our approach relies on word-level alignments between our collected captions. We first generate word-level alignments using Fast Align (Dyer et al., 2013) to create word alignment data A. From A, we generate a dictionary of word pairs $\{(word_{src}, word_{tgt}) : count_{align}(word_{src}, word_{tgt}) > c \wedge P_{align}(word_{tgt}|word_{src}) > p\}$. Here $count_{align}$ is the number of times $word_{tgt}$ and $word_{src}$ are aligned in A, P_{align} is the probability of alignment as observed in A, and c and p are two tied hyper-parameters. We use $(p, c) \in \{(0.5, 20), (0.6, 5), (0.9, 2)\}$, chosen by tuning on the development set. We manually

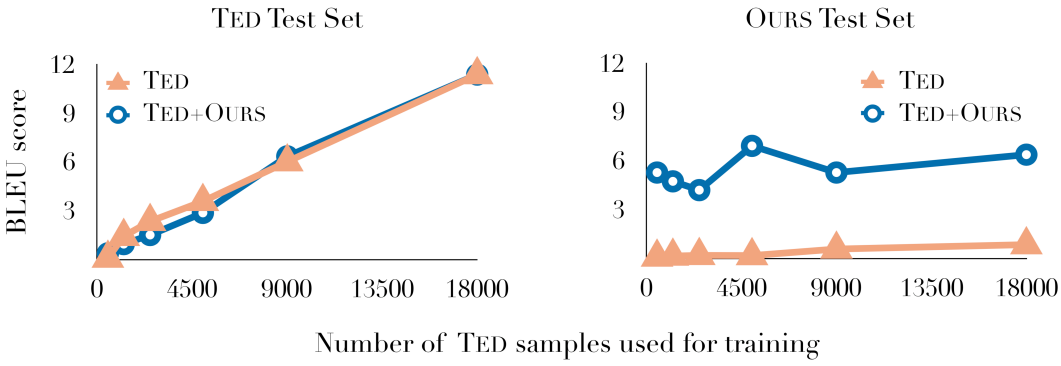

Figure 3: Machine Translation BLEU Scores on TED and OURS Test Set

evaluated the generated word pairs and found that the performance was reasonable for comparable data – out of 75 word pairs generated by our method, 43 (57.3%) were accurate. Some sample translation pairs induced from the dictionary are shown in the Appendix.

3.3 Cost Breakdown

We used Amazon Mechanical Turk (MTurk) to obtain 2500 captions for the 500 images. Seventy-six workers participated in the job for a total cost of 197 USD. The average time required to caption each image was 4.04 minutes, i.e., a total of about 168 hours for 2500 captions. On average, professional translators charge about 0.1 USD/word or 31.56 USD/hour for English-Hindi translation.[1] Given that the 2500 English captions had a total of 114,433 words, professional translation would have charged about 5,539 USD by hourly rates or 12,107 USD by the per-word rate. Thus, our method is **at least about 28 times cheaper**.

Note that as per the recommendation of the Government of India (GOI, 2019), the highest recommended minimum wage in India across all zones is 447 INR/day or 6.27 USD/day. Assuming an 8-hour working day, this is 78.37 cents per hour – we paid ≈50% higher, at 117.2 cents per hour (all workers were required to be located in India).

4 Conclusion and Future Work

In this work, we propose a method that uses images for generating high-quality parallel training data without the need for bilingual translators. More specifically, our technique for image selection and crowdsourcing results in useful training data for scenarios where finding annotators proficient in both the languages is challenging, as demonstrated by human evaluation and downstream task performance.

To the best of our knowledge, we are the first to introduce the idea of crowdsourcing comparable data using images for low resource settings. Hewitt et al. (2018) and (Bergsma & Van Durme, 2011) rely on a corpus of images associated with words (accessed via image search engines) in the languages of interest. Similarities in images are then used to induce bilingual lexicons. In contrast, our method is ideal for settings where absolutely no resources are available for a low resource language. Further, Singhal et al. (2019) use a similar proprietary dataset obtained via Google search to learn multilingual embeddings. Hitschler et al. (2016) and (Chen et al., 2019) improve the quality of statistical machine translation and bilingual lexicon induction by using large monolingual image-captioning corpora. While their work is orthogonal to ours, it underscores the fact that the dataset generated by our method can indeed boost downstream tasks.

In the future, we plan to use our data creation technique on extremely low-resource languages and release parallel corpora that can potentially propel the use of state-of-the-art NLP techniques on these languages.

[1] https://search.proz.com/employers/rates

References

Sadaf Abdul-Rauf and Holger Schwenk. On the use of comparable corpora to improve smt performance. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pp. 16–23, 2009.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*, 2019.

Loïc Barrault, Ondřej Bojar, Marta R Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, et al. Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pp. 1–61, 2019.

Shane Bergsma and Benjamin Van Durme. Learning bilingual lexicons using the visual similarity of labeled web images. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

C Chandramouli and Registrar General. Census of india 2011. *Provisional Population Totals. New Delhi: Government of India*, 2011.

Shizhe Chen, Qin Jin, and Alexander Hauptmann. Unsupervised bilingual lexicon induction from mono-lingual multimodal data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 8207–8214, 2019.

Chris Dyer, Victor Chahuneau, and Noah A Smith. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 644–648, 2013.

GOI. Report of the expert committee on determining the methodology for fixing the national minimum wage. *Report of the Expert Committee on Determining the Methodology for Fixing the National Minimum Wage*, 2019.

John Hewitt, Daphne Ippolito, Brendan Callahan, Reno Kriz, Derry Tanti Wijaya, and Chris Callison-Burch. Learning translations via images with a massively multilingual image dataset. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2566–2576, 2018.

Julian Hitschler, Shigehiko Schamoni, and Stefan Riezler. Multimodal pivots for image caption translation. *arXiv preprint arXiv:1601.03916*, 2016.

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

Ann Irvine and Chris Callison-Burch. Combining bilingual and comparable corpora for low resource machine translation. In *Proceedings of the eighth workshop on statistical machine translation*, pp. 262–270, 2013.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. The iit bombay english-hindi parallel corpus. *arXiv preprint arXiv:1710.02855*, 2017.

Rensis Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

Dragos Stefan Munteanu, Alexander Fraser, and Daniel Marcu. Improved machine translation performance via parallel sentence extraction from comparable corpora. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pp. 265–272, 2004.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311–318. Association for Computational Linguistics, 2002.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

Karan Singhal, Karthik Raman, and Balder ten Cate. Learning multilingual word embeddings using image-text data. *arXiv preprint arXiv:1905.12260*, 2019.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.