Don’t Rule Out Monolingual Speakers: A Method For Crowdsourcing Machine Translation Data

Rajat Bhatnagar, Ananya Ganesh and Katharina Kann
University of Colorado Boulder
{rajat.bhatnagar, ananya.ganesh, katharina.kann}@colorado.edu

Abstract

High-performing machine translation (MT) systems can help overcome language barriers while making it possible for everyone to communicate and use language technologies in the language of their choice. However, such systems require large amounts of parallel sentences for training, and translators can be difficult to find and expensive. Here, we present a data collection strategy for MT which, in contrast, is cheap and simple, as it does not require bilingual speakers. Based on the insight that humans pay specific attention to movements, we use graphics interchange formats (GIFs) as a pivot to collect parallel sentences from monolingual annotators. We use our strategy to collect data in Hindi, Tamil and English. As a baseline, we also collect data using images as a pivot. We perform an intrinsic evaluation by manually evaluating a subset of the sentence pairs and an extrinsic evaluation by finetuning mBART (Liu et al., 2020) on the collected data. We find that sentences collected via GIFs are indeed of higher quality.

1 Introduction

Machine translation (MT) – automatic translation of text from one natural language into another – provides access to information written in foreign languages and enables communication between speakers of different languages. However, developing high performing MT systems requires large amounts of training data in the form of parallel sentences – a resource which is often difficult and expensive to obtain, especially for languages less frequently studied in natural language processing (NLP), endangered languages, or dialects. For some languages, it is possible to scrape data from the web (Resnik and Smith, 2003), or to leverage existing translations, e.g., of movie subtitles (Zhang et al., 2014) or religious texts (Resnik et al., 1999). However, such sources of data are only available for a limited number of languages, and it is impossible to collect large MT corpora for a diverse set of languages using these methods. Professional translators, which are a straightforward alternative, are often rare or expensive.

In this paper, we propose a new data collection strategy which is cheap, simple, effective and, importantly, does not require professional translators or even bilingual speakers. It is based on two assumptions: (1) non-textual modalities can serve as a pivot for the annotation process (Madaan et al., 2020); and (2) annotators subconsciously pay increased attention to moving objects, since humans are extremely good at detecting motion, a crucial skill for survival (Albright and Stoner, 1995). Thus, we propose to leverage graphics interchange formats (GIFs) as a pivot to collect parallel data in two or more languages.

We prefer GIFs over videos as they are short in duration, do not require audio for understanding and describe a comprehensive story visually. Furthermore, we hypothesize that GIFs are better pivots than images – which are suggested by Madaan et al. (2020) for MT data collection – based on our second assumption. We expect that people who

Figure 1: Sentences written by English and Hindi annotators using GIFs or images as a pivot.
are looking at the same GIF tend to focus on the main action and characters within the GIF and, thus, tend to write more similar sentences. This is in contrast to using images as a pivot, where people are more likely to focus on different parts of the image and, hence, to write different sentences, cf. Figure 1.

We experiment with collecting Hindi, Tamil and English sentences via Amazon Mechanical Turk (MTurk), using both GIFs and images as pivots. As an additional baseline, we compare to data collected in previous work (Madaan et al., 2020). We perform both intrinsic and extrinsic evaluations – by manually evaluating the collected sentences and by training MT systems on the collected data, respectively – and find that leveraging GIFs indeed results in parallel sentences of higher quality as compared to our baselines.\(^1\)

\section{Related Work}

In recent years, especially with the success of transfer learning (Wang et al., 2018) and pretraining in NLP (Devlin et al., 2019), several techniques for improving neural MT for low-resource languages have been proposed (Sennrich et al., 2016; Fadaee et al., 2017; Xia et al., 2019; Lample et al., 2017; Lewis et al., 2019; Liu et al., 2020).

However, supervised methods still outperform their unsupervised and semi-supervised counterparts, which makes collecting training data for MT important. Prior work scrapes data from the web (Lai et al., 2020; Resnik and Smith, 2003), or uses movie subtitles (Zhang et al., 2014), religious texts (Resnik et al., 1999), or multilingual parliament proceedings (Koehn, 2005). However, those and similar resources are only available for a limited set of languages. A large amount of data for a diverse set of low-resource languages cannot be collected using these methods.

For low-resource languages, Hasan et al. (2020) propose a method to convert noisy parallel documents into parallel sentences. Zhang et al. (2020) filter noisy sentence pairs from MT training data.

The closest work to ours is Madaan et al. (2020). The authors collect (pseudo-)parallel sentences with images from the Flickr8k dataset (Hodosh et al., 2013) as a pivot, filtering to obtain images which are simplistic and do not contain culture-specific references. Since Flickr8k already contains 5 English captions per image, they select images whose captions are short and of high similarity to each other. Culture-specific images are manually discarded. We compare to the data from Madaan et al. (2020) in Section 4, denoting it as M20.

\section{Experiments}

\subsection{Pivot Selection}

We propose to use GIFs as a pivot to collect parallel sentences in two or more languages. As a baseline, we further collect parallel data via images as similar to our GIFs as possible. In this subsection, we describe our selection of both mediums.

\textbf{GIFs} We take our GIFs from a dataset presented in Li et al. (2016), which consists of 100k GIFs with descriptions. Out of these, 10k GIFs have three English one-sentence descriptions each, which makes them a suitable starting point for our experiments. We compute the word overlap in F1 between each possible combination of the three sentences, take the average per GIF, and choose the highest scoring 2.5k GIFs for our experiments. This criterion filters for GIFs for which all annotators focus on the same main characters and story, and it eliminates GIFs which are overly complex. We thus expect speakers of non-English languages to focus on similar content.

\textbf{Images} Finding images which are comparable to our GIFs is non-trivial. While we could compare our GIFs’ descriptions to image captions, we hypothesize that the similarity between the images obtained thereby and the GIFs would be too low for a clean comparison. Thus, we consider two alternatives: (1) using the first frame of all GIFs, and (2) using the middle frame of all GIFs.

In a preliminary study, we obtain two Hindi one-sentence descriptions from two different annotators for both the first and the middle frame for a subset of 100 GIFs. We then compare the BLEU (Papineni et al., 2002) scores of all sentence pairs. We find that, on average, sentences for the middle frame have a BLEU score of 7.66 as compared to 4.58 for the first frame. Since a higher BLEU score indicates higher similarity and, thus, higher potential suitability as MT training data, we use the middle frames for the image-as-pivot condition in our final experiments.

\footnote{\textsuperscript{1}All data collected for our experiments is available at https://nala-cub.github.io/resources.}
### Sentences from the GIF-as-Pivot Setting

| Rating | Sentences from the GIF-as-Pivot Setting |
|--------|----------------------------------------|
| 1      | A child flips on a trampoline.          |
|        | A girl enjoyed while playing.           |
| 3      | A man in a hat is walking up the stairs holding a bottle of water. |
|        | A man is walking with a plastic bottle. |
| 5      | A man is laughing while holding a gun.  |
|        | A man is laughing while holding a gun.  |

### Sentences from the Image-as-Pivot Setting

| Rating | Sentences from the Image-as-Pivot Setting |
|--------|------------------------------------------|
| 1      | A woman makes a gesture in front of a group of other women. |
|        | This woman is laughing.                   |
| 3      | An older woman with bright lip stick lights a cigarette in her mouth. |
|        | This woman is lighting a cigarette.       |
| 5      | A woman wearing leopard print dress and a white jacket is walking forward. |
|        | A woman is walking with a leopard print dress and white coat. |

Table 1: Sentences obtained in English and Hindi for each setting where both annotators agree on the rating. The first sentence is the sentence written in English and the second sentence is the corresponding English translation of the Hindi sentence, translated by the authors.

### 3.2 Data Collection

We use MTurk for all of our data collection. We collect the following datasets: (1) one single-sentence description in Hindi for each of our 2,500 GIFs; (2) one single-sentence description in Hindi for each of our 2,500 images, i.e., the GIFs’ middle frames; (3) one single-sentence description in Tamil for each of the 2,500 GIFs; (4) one single-sentence description in Tamil for each of the 2,500 images; and (5) one single-sentence description in English for each of our 2,500 images. To build parallel data for the GIF-as-pivot condition, we randomly choose one of the available 3 English descriptions for each GIF.

For the collection of Hindi and Tamil sentences, we restrict the workers to be located in India and, for the English sentences, we restrict the workers to be located in the US. We use the instructions from Li et al. (2016) with minor changes for all settings, translating them for Indian workers.\(^2\)

Each MTurk human intelligence task (HIT) consists of annotating five GIFs or images, and we expect each task to take a maximum of 6 minutes. We pay annotators in India $0.12 per HIT (or $1.2 per hour), which is above the minimum wage of $1 per hour in the capital Delhi.\(^3\) Annotators in the US are paid $1.2 per HIT (or $12 per hour). We have obtained IRB approval for the experiments reported in this paper (protocol #: 20-0499).

\(^2\)Our instructions can be found in the appendix.

\(^3\)https://paycheck.in/salary/minimumwages/16749-delhi

### 3.3 Test Set Collection

For the extrinsic evaluation of our data collection strategy we train and test an MT system. For this, we additionally collect in-domain development and test examples for both the GIF-as-pivot and the image-as-pivot setting.

Specifically, we first collect 250 English sentences for 250 images which are the middle frames of previously unused GIFs. We then combine them with the English descriptions of 250 additional unused GIFs from Li et al. (2016). For the resulting set of 500 sentences, we ask Indian MTurk workers to provide a translation into Hindi and Tamil. We manually verify the quality of a randomly chosen subset of these sentences. Workers are paid $1.2 per hour for this task. We use 100 sentence pairs from each setting as our development set and the remaining 300 for testing.

### 4 Evaluation

#### 4.1 Intrinsic Evaluation

In order to compare the quality of the parallel sentences obtained under different experimental conditions, we first perform a manual evaluation of a subset of the collected data. For each lan-
Table 3: Cumulative percentages with respect to each setting; GIF-as-pivot shows the best results;

Table 4: BLEU for different training and test sets; All denotes a weighted average over all test sets; all models are obtained by finetuning mBART; best scores for each training set size and test set in bold.

4The definitions of each score as given to the annotators can be found in the appendix.
Table 5: BLEU for different training and test sets; All denotes a weighted average over all test sets; all models are obtained by finetuning mBART; best scores for each training set size and test set in bold.

| Test Set | Training Set | 500 | 1000 | 1500 | 2000 | 2500 |
|----------|--------------|-----|------|------|------|------|
| GIF      | GIF          | 2.63| 4.46 | 8.26 | 9.27 | 4.99 |
| GIF      | Image        | 2.33| 3.34 | 3.00 | 4.77 | 3.83 |
| Image    | GIF          | 0.95| 2.42 | 3.15 | 3.67 | 2.74 |
| Image    | Image        | 6.65| 5.62 | 6.02 | 7.75 | 7.22 |
| All      | GIF          | 1.79| 3.44 | 5.71 | 6.47 | 3.87 |
| All      | Image        | 4.49| 4.48 | 4.51 | 6.26 | 5.53 |

| Test Set | Training Set | 500 | 1000 | 1500 | 2000 | 2500 |
|----------|--------------|-----|------|------|------|------|
| GIF      | GIF          | 0.54| 1.00 | 0.83 | 0.98 | 0.86 |
| GIF      | Image        | 0.5 | 0.18 | 0.96 | 0.43 | 0.48 |
| Image    | GIF          | 0  | 0.31 | 0.36 | 0.62 | 0.7  |
| Image    | Image        | 0.41| 0.35 | 0.51 | 0.36 | 0.29 |
| All      | GIF          | 0  | 0.43 | 0.68 | 0.73 | 0.77 |
| All      | Image        | 0.46| 0.27 | 0.74 | 0.4  | 0.39 |

We conclude that 2,500 examples are not enough to train an MT system for these directions, and, while we report all results here for completeness, we believe that the intrinsic evaluation paints a more complete picture. We leave a scaling of our extrinsic evaluation to future work.

## 5 Conclusion

In this work, we made two assumptions: (1) that a non-textual modality can serve as a pivot for MT data collection, and (2) that humans tend to focus on moving objects. Based on this, we proposed to collect parallel sentences for MT using GIFs as pivots, eliminating the need for bilingual speakers and reducing annotation costs. We collected parallel sentences in English, Hindi and Tamil using our approach and conducted intrinsic and extrinsic evaluations of the obtained data, comparing our strategy to two baseline approaches which used images as pivots. According to the intrinsic evaluation, our approach resulted in parallel sentences of higher quality than either baseline.

## Acknowledgments

We would like to thank the anonymous reviewers, whose feedback helped us improve this paper. We are also grateful to Aman Madaan, the first author of the M20 paper, for providing the data splits and insights from his work. Finally, we thank the members of CU Boulder’s NALA Group for their feedback on this research.

## References

- Thomas D Albright and Gene R Stoner. 1995. Visual motion perception. *Proceedings of the National Academy of Sciences*, 92(7):2433–2440.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Marzieh Fadaee, Arianna Biazzza, and Christof Monz. 2017. Data augmentation for low-resource neural
### A Sentence Rating Instructions

| Score | Title       | Description                                                   |
|-------|-------------|---------------------------------------------------------------|
| 1     | Not a translation | There is no relation whatsoever between the source and the target sentence |
| 2     | Bad         | Some word overlap, but the meaning isn’t the same             |
| 3     | Acceptable  | The translation conveys the meaning to some degree but is a bad translation |
| 4     | Good        | The translation is missing a few words but conveys most of the meaning adequately |
| 5     | Perfect     | The translation is perfect or close to perfect                 |

Table 6: Description of the ratings for the manual evaluation of translations.
Instructions for English Image Task

Below you will see five images. Your task is to describe each image in one English sentence. You should focus solely on the visual content presented in the image. Each sentence should be grammatically correct. It should describe the main characters and their actions, but NOT your opinions, guesses or interpretations.

- **DOs**
  - Please use only English words. No digits allowed (spell them out, e.g., three).
  - Sentences should neither be too short nor too long. Try to be concise.
  - Each sentence must contain a verb.
  - If possible, include adjectives that describe colors, size, emotions, or quantity.
  - Please pay attention to grammar and spelling.
  - Each sentence must express a complete idea, and make sense by itself.
  - The sentence should describe the main characters, actions, setting, and relationship between the objects.

- **DONTs**
  - The sentence should NOT contain any digits.
  - The sentence should NOT mention the name of a movie, film, or character.
  - The sentence should NOT mention invisible objects.
  - The sentence should NOT contain any digits.
  - The sentence should NOT make subjective judgments about the image.

Remember, please describe only the visual content presented in the images. Focus on the main characters and their actions.

Figure 2: Instructions for the data collection via images in English, via GIFs in Hindi and Tamil.