Practical Comparable Data Collection for Low-Resource Languages via Images

Aman Madaan, Shruti Rijhwani, Antonios Anastasopoulos, Yiming Yang & Graham Neubig

Language Technologies Institute, Carnegie Mellon University
Pittsburgh, PA 15213, USA
Introduction

• Machine Translation: *Convert text in language A to language B*

• Requires **parallel** data
  - Same sentence in both the source and target languages

• Comparable data (approximate translations) is still effective (Munteanu et al., 2004; Abdul-Rauf & Schwenk, 2009; Irvine & Callison-Burch, 2013)
Objective

• Parallel/Comparable data curation requires bilingual speakers

• For Extremely low-resource languages, bilingual speakers might not be available or proficient in the second language

• To create comparable data without using bilingual speakers via images
Key Idea

• Use simple, universal images to gather captions on both the source and target languages.

• Captions on the same image in the two languages should be comparable.

A dog running on grass

一只狗在草上跑

कुत्ता घास पर दौड़ रहा है

A dog running on grass

A dog running on grass
Methodology

- Given $N$ (simple + universal) images in the target language, each with $P$ captions

- Obtain $Q$ captions for each image in the source language from annotators proficient in (only) the source language

- Instruct the annotators to be concise, use a single sentence

- Cartesian product of the captions to yield $P \times Q$ comparable sentences OR random assignment to get $\min(P, Q)$ comparable sentences

- Method requires no resources in the source language apart from the instructions for the annotators.

- Target language is typically high-resource for practical settings

---

**English Captions for Flickr8k (P=5)**
- 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 (Q = 5)**
- 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 ट्रेकिंग करता हुआ
- एक आदमी पहाड़ पर चढ़ाई कर रहा है

---

**Translated Hindi Captions**
- 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
Desired Images

• Images should be simple + universal
Simplicity and Universalness

• **Simplicity**: Captions for simple images are short, have fewer unique words, and are consistent across annotators.

• Simpler images have lower $d_i$. For $C_{trg}^i$ set of captions for the $i^{th}$ image, calculate $d_i = l_i + w_i + e_i$

$$l_i = \sum_{j=1}^{p} \text{length}(c_{i,j}^{trg}) \quad w_i = \sum_{j=1}^{p} \text{unique_words}(c_{i,j}^{trg}) \quad e_i = \sum_{j=1}^{p} \sum_{k=j+1}^{p} \text{edit_distance}(c_{i,j}^{trg}, c_{i,k}^{trg})$$

• **Universalness**: Hard to quantify, our heuristic is to start from a set of relatively generic images (Hodosh et al., 2013),
Experiments and Results
Dataset Selection

• We used the Flickr8k dataset (Hodosh et al., 2013)
  - Contains images that depict everyday actions and events involving people and animals. (favors universal images)
  - Aims to include images that can be unambiguously described in a sentence (favors simple images)

• Selected 700 images (defined by the caption complexity score), pruned to 500 images manually
Obtaining Captions

• Hindi selected as the source language

• Five captions per image, 2500 captions for 500 images

• Crowd workers sourced via Amazon Mechanical Turk

• We make no assumptions specific to Hindi in our setup, and it can be adopted for any other language

• Workers were required to be in India, were paid 150% more than the highest minimum wage recommended by the Govt. of India
Manual Evaluation

• Manual evaluation for 600 comparable sentences

• 81% acceptable or better, only 2.47% rated as not a translation at all
Downstream Tasks

• Unsupervised Dictionary Extraction with Fast-Align - 57% accuracy

• Machine Translation
Conclusion & Next Steps

• We propose a method that uses images for generating high-quality comparable training data without the need for bilingual translators

• Human evaluation and downstream task performance show that data has comparable characteristics

• We plan to use our data creation technique on extremely low-resource languages

• It would also be interesting to explore methods to quantify the definition of universalness
Thanks!


Code/data: https://github.com/madaan/PML4DC-Comparable-Data-Collection