Supplemental Materials: Learning Translations via Images with a Massively Multilingual Image Dataset

John Hewitt  Daphne Ippolito  Brendan Callahan  Reno Kriz
Derry Wijaya  Chris Callison-Burch
University of Pennsylvania
Computer and Information Science Department
{johnhew, daphnei, derry, ccb}@seas.upenn.edu

1 Image Quality Analysis

In order to ascertain that the images in our dataset were high-quality, it was necessary to manually validate that the images associated with a word were indeed related to that word. We limited our analysis to French, Indonesian, and English.

We selected a random sample of 955 English words that had both French and Indonesian translations in our dataset as well as groundtruth concreteness scores. We gathered the first 10 images for each of these words in each of the three languages. We then crowdsourced annotations from workers on Amazon Mechanical Turk. We did so by providing workers each English word, and the associated 10 images, and asking them to identify which images did NOT relate to that word. For French and Indonesian, we provided workers the translated English word, rather than the original foreign word. Finally, alongside the 10 images, we provided two additional images associated with a different word in our dataset, to provide quality control. An example of the interface we showed the Mechanical Turkers is provided in Figure 2.

2 Corpus Creation Pipeline

Here are the steps in our corpus creation pipeline in more detail:

1. We started with a collection of bilingual dictionaries between 100 foreign languages and English. Most of the dictionaries contain approximately 10,000 foreign words with one or more English translations. We discard dictionary entries where the English word and foreign word are identical.

2. For each word in a bilingual dictionary, we queried Google Image Search using the word as the search term, and setting the language-specific search fields to the appropriate language.

3. We saved the first 100 images returned by Google Image Search for each word, along with the image’s web page and other metadata.

4. We performed language identification on the text from the page that the image appeared on.

5. Based on the result of the language identification, we filtered out images whose text did not seem to be in the expected language.

6. We produced low-dimensional vector representations for each image using a convolutional neural network trained on ImageNet (Deng et al., 2009).

3 Details on the Bilingual Dictionaries

Our images were based on the crowdsourced bilingual dictionaries assembled by Pavlick et al. (2014). Most of the bilingual dictionaries contain approximately 10,000 foreign words, but the exact number varies per language, since Pavlick et al. (2014) filtered the dictionaries based on the estimated quality of the crowd workers making the contribution in order to discard poor translations. As a helpful guide, we have grouped the languages by ranges of word counts, so that readers can get a sense of how large the corpus for each language is.

- 8,000-10,000 words: Afrikaans, Albanian, Amharic, Arabic, Azerbaijani, Basque, Belarusian, Bengali, Bihon, Bosnian, Bulgarian, Catalan, Central Bicolano, Croatian, Danish, Dutch, Esperanto, Filipino, Finnish, French, Galician, German, Greek, Gujarati, Haitian, Hebrew, Hindi, Hungarian, Ilokano, Indonesian, Italian, Japanese, Javanese, Kannada, Kapampangan, Latvian,
Lithuanian, Macedonian, Malay, Malayalam, Marathi, Nepali, Norwegian Nynorsk, Norwegian, Piedmontese, Polish, Portuguese, Punjabi, Romanian, Russian, Serbian, Serbo-Croatian, Slovak, Slovenian, Somali, Spanish, Sundanese, Swedish, Tamil, Telugu, Turkish, Ukrainian, Waray, Welsh, Urdu, Uzbek

- 5,000-8,000 words: Breton, Czech, Frisian, Georgian, Irish, Korean, Low Saxon, Luxembourgish, Swahili, Uighur, Vietnamese
- 1,000-5,000 words: Aragonese, Armenian, Chinese, Neapolitan, Sicilian, Thai, Yoruba
- 100-1,000 words: Icelandic, Malagasy, Newar, Pashto, Persian, Sindhi
- <100 words: Ido, Kazakh, Kurdish, Wolof

Table 1 shows the sizes of the image sets extracted for each language before and after filtering out images that showed up on web pages in languages other than the expected language.

4 Details on the Complementary Text Corpus

The words used on a web page containing an image are likely to be related to the image; this is a heuristic used with great success by search engines. By extracting the text of the web pages from which we drew our image corpus, we are able to accomplish dual goals of ensuring the text is in the language of interest, and enhancing the image dataset with a “comparable corpus.” A comparable corpus is a multilingual dataset with some noisy signal of translation equivalence. In our case, text extracted from web pages with similar images is likely to be topically similar. Because the image similarities are language-independent, we get a noisy multilingual signal.

Due to the vagaries of the internet, we were able to extract text for approximately 78% of the images in the corpus. Tables 2 and 3 show the top-5 most common languages detected in the text corpus, along with the fraction of web pages represented by that language, for 6 high- and low-resource languages of interest, respectively. Note that each web page does not correspond to a single image in the image corpus; instead 1 web page might be shared across many images.

The percent of web pages written in the language of interest varied greatly from language to language, but for high-resource languages was frequently between 50% and 60%. Qualitatively, many pages were from YouTube or other English-speaking sites that happened to rank highly on foreign-language image searches. This motivates the necessity of filtering images used in the bilingual lexicon induction task to those coming from in-language web pages. Figure 1 shows an example of text extracted from a page retrieved from the image search metadata. This text is paired with the image that appeared on its web page, as well as the Indonesian word used in the image search.

4.1 Language-Confidence Heuristic

We used the heuristic that as long as an expected language showed up in the top-3 most likely languages as output by our language detection system on a web page, images on that page were kept. This relatively lenient heuristic is well-motivated because of the nature of automatically-scraped text from the web. English text is pervasive on the internet, even when the primary language of content of the page is not English. Further, many pages with our images have small amounts of text. In all cases, we jointly attempt to detect all languages on
a given page. Thus, when the language of interest shows up on the top 3 guesses, there is reasonable evidence that some of the text on the page is in that language (or, admittedly, a related language), even if there's also a substantial amount of English or some other language. These multilingual web pages are, for our purposes, valid to be kept.

5 Corpus Structure

We have uploaded sample data for two languages (French and Indonesian) to a Dropbox folder so that they may be downloaded. They are available for download at: https://www.dropbox.com/sh/fc31nedbtun3j0p/AACzp2GQBG19pNGmjJVH60wVa?dl=0
Along with the description of the data below, we provide a README with exact commands for easy analysis of the sample.

5.1 Image Sample Files

For each language package, we have constructed a sample file that contains the images for a random selection of 100 words.

The French sample file is 2.68GB and the Indonesian sample file is 1.1GB.

For each language, the download is a .tar file. When extracted, 100 .tar.gz files are given. Each one, when unzipped, is a directory of arbitrary index, representing all the image data for a single word.

Each such directory has the following files:

01-99.ext : approximately 100 image files of varying extensions.
metadata.json Metadata extracted from the Google image search, including the URL of the web page on which the image was found.
word.txt The plain text of the word.

We also include corresponding sample files with the English translations for French and Indonesian. The format is identical, except for an outer directory named English-## that indicates which part (of 27 total) of our "English superset" the English word was in.

Metadata example:

```json
{
  "original_url": "https://shopswell-...
  "page_title": "EEK! A MOUSE! |
  "image_type": "jpg",
  "thumbnail_url": "https://encrypted...
  "referring_url": "https://www.shops...
  "image_site_url": "shopswell.com",
  "subtitle/sentence": "A MOUSE!
  "thumbnail_width": 284,
  "thumbnail_height": 177,
  "original_width": 1680,
  "original_height": 1050,
}
```

5.2 Summary Files

We have provided JSON files that provide summary information for each language. The keys are named as follows.

- total_words - the total number of words
- total_images - the total number of images
- total_file_size - the total aggregate file size of the images
- avg_file_size - the average file size across all images
- avg_width - the average pixel width across all images
- max_images_per_word - the maximum number of images per word
- min_images_per_word - the minimum number of images per word
- median_images_per_word - the median number of images per word
- num_unique_hosts - the number of unique hostnames for the images
- top_10Hostname_counts - the top 10 most frequently seen hostnames, and their counts
- extension_counts - counts of the image file extensions

5.3 Dictionary Files

For each foreign language, we include the bilingual dictionaries from Pavlick et al. (2014). They are named dict. followed by the two letter language code (ie dict.fr for French and dict.id for Indonesian). These text files have one foreign word per line. Each line is tab separated and after the first column are the list of possible English translations.
5.4 Feature files

We provide precomputed Imagenet features for each image. The layout of the folders and feature files mirrors the image package layout.

The features are stored in Python 3 pickle files, whose value upon being read in is the 4,096 dimension numpy array that can be used for image comparison.

5.5 Text Corpus Sample and Language Confidence Filter

We have provided, for each language, the subset of the text corpus that was extracted from the images in the sample. This is given in the form of a file per image, uniquely identified by the language, word index, and image index for that word. The text is tokenized.

Along with the text files, we provide our language identification result for each page, and a Python script to filter the provided images to only the set of images drawn from web pages written in the expected language.

6 Qualitative Examples

Informative qualitative examples from the image dataset are given in Figs. 4, 5, 6, 7, 8, 9.

References


Table 1: Statistics about the number of images and words in our data set. Estimated numbers are rounded to the nearest 1,000. A same translation word (stw) refers to words whose translations are exact matches of the word itself, non-stw words show the count where that is not the case.
<table>
<thead>
<tr>
<th>Source Language</th>
<th>Arabic</th>
<th>Dutch</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detected Language</td>
<td>Arabic</td>
<td>.5543</td>
<td>Dutch</td>
</tr>
<tr>
<td>English</td>
<td>.4189</td>
<td>English</td>
<td>.4539</td>
</tr>
<tr>
<td>Persian</td>
<td>.0051</td>
<td>German</td>
<td>.0078</td>
</tr>
<tr>
<td>French</td>
<td>.0020</td>
<td>French</td>
<td>.0056</td>
</tr>
<tr>
<td>Norwegian</td>
<td>.0015</td>
<td>Norwegian</td>
<td>.0022</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source Language</th>
<th>German</th>
<th>Italian</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detected Language</td>
<td>German</td>
<td>.5946</td>
<td>Italian</td>
</tr>
<tr>
<td>English</td>
<td>.3847</td>
<td>English</td>
<td>.3797</td>
</tr>
<tr>
<td>Dutch</td>
<td>.0038</td>
<td>Spanish</td>
<td>.0109</td>
</tr>
<tr>
<td>French</td>
<td>.0032</td>
<td>Portuguese</td>
<td>.0063</td>
</tr>
<tr>
<td>Norwegian</td>
<td>.0019</td>
<td>French</td>
<td>.0036</td>
</tr>
</tbody>
</table>

Table 2: The top-5 most common languages detected in individual pages for each of 6 high-resource languages. With each language is the fraction of web pages represented by that language.

<table>
<thead>
<tr>
<th>Source Language</th>
<th>Bengali</th>
<th>Cebuano</th>
<th>Indonesian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detected Language</td>
<td>Bengali</td>
<td>.7256</td>
<td>English</td>
</tr>
<tr>
<td>English</td>
<td>.2300</td>
<td>Spanish</td>
<td>.0401</td>
</tr>
<tr>
<td>Russian</td>
<td>.0160</td>
<td>Tagalog</td>
<td>.0264</td>
</tr>
<tr>
<td>Tajik</td>
<td>.0083</td>
<td>Cebuano</td>
<td>.0227</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>.0073</td>
<td>French</td>
<td>.0129</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source Language</th>
<th>Turkish</th>
<th>Uighur</th>
<th>Uzbek</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detected Language</td>
<td>Turkish</td>
<td>.6772</td>
<td>Uighur</td>
</tr>
<tr>
<td>English</td>
<td>.3077</td>
<td>English</td>
<td>.1140</td>
</tr>
<tr>
<td>Spanish</td>
<td>.0013</td>
<td>Inupiaq</td>
<td>.0778</td>
</tr>
<tr>
<td>Indonesian</td>
<td>.0011</td>
<td>Arabic</td>
<td>.0291</td>
</tr>
<tr>
<td>German</td>
<td>.0011</td>
<td>Persian</td>
<td>.0290</td>
</tr>
</tbody>
</table>

Table 3: The top-5 most common languages detected in individual pages for each of 6 low-resource languages. With each language is the percent of web pages represented by that language. Note that we used a separate, unpublished language detection system for Uighur because CLD2 does not support Uighur detection.
Figure 2: An example of the task we gave to the workers on Amazon Mechanical Turk, along with the instructions given. In this example, we can the first two images in the left column are the two controls we put in, which are not associated with the word “mail”.
Figure 3: Our dataset allows translations to be discovered by comparing the images associated with foreign and English words. Shown here are five images for the Indonesian word *kucing*, along with its top 4 ranked translations using CNN features.
Figure 4: Shown here are five images for the abstract Indonesian word berharap, along with its top 4 ranked translations using CNN features. At the bottom are images for the actual translation hope, which was ranked 536.
Figure 5: Shown here are five images for the abstract Indonesian word konsep, along with its top 4 ranked translations using CNN features. The actual translation, concept, was ranked 3,465.

Figure 6: Top 10 images for French word: étoile and the top 4 ranked translations using CNN features.
Figure 7: The top 4 ranked translations of the Indonesian word *naga* using CNN features.

Figure 8: The top 4 ranked translations of the French word *noblesse* using CNN features.

Figure 9: The top 4 ranked translations of the French word *romain* using CNN features.