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, Bishnupriya Manipuri, 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

Indonesian (extracted from web page)
Kucing merupakan salah satu jenis hewan yang banyak dipelihara oleh kebanyakan orang. Wajahnya yang lucu dan imut menjadika n kucing adalah teman bermain yang menggemaskan. Belum lagi tingkah polah dari kucing yang kerap manja serta menarik perhatian membuat kita betah berlama-lama bermain dengan kucing.

Translation (done for illustrative purposes)
Cats are one of the animals that many people keep. Their funny and cute faces make cats adorable playmates. Not to mention their behaviors are often affectionate and done to attract attention, which makes us enjoy spending a lot of time playing with cats.

Figure 1: Example text extracted from a web page corresponding to an image found for the Indonesian word *kucing* (cat), and the same text manually translated to English.

to language, but for high-resouce 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/AACzpZGQBG19pNGmjJVH60wVa?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:


```
{
"original_url":␣"https://shopswell-.
"page_title":␣"EEK!␣A␣MOUSE!␣|␣Shop.
"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_10_hostname_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

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. IEEE Conference on*, pages 248–255. IEEE.

Ellie Pavlick, Matt Post, Ann Irvine, Dmitry Kachaev, and Chris Callison-Burch. 2014. The language demographics of Amazon Mechanical Turk. *Transactions of the Association for Computational Linguistics*, 2:79–92.

	Words in Dict		Before Filtering			After Filtering		
Language	All	Non-stw	Images	Web Pages	Webcrawl Words	Images	Web Pages	Webcrawl Words
Afrikaans	9,719	8,147	923,000	692,000	456,860,000	324,000	541,000	322,327,000
Albanian	9,600	9,036	923,671	665,880	439,126,000	324,000	520,000	309,815,000
Amharic	8,051	8,051	765,000	574,000	378,324,000	268,000	448,000	266,918,000
Arabic	9,813	9,811	941,011	685,467	421,236,351	329,826	542,544	360,831,818
Aragonese	3,141	2,557	298,000	224,000	147,783,000	105,000	175,000	104,265,000
Armenian	1,652	1,652	157,000	118,000	77,691,000	55,000	92,000	54,813,000
Asturian	3,486	2,535	331,000	248,000	163,828,000	116,000	194,000	115,585,000
Azerbaijani	9,942	9,287	931,691	665,843	439,126,000	327,000	520,000	309,815,000
Basque	9,686	8,728	920,000	690,000	455,171,000	323,000	539,000	321,135,000
Belarusian	10,087	9,973	958,000	719,000	474,594,000	336,000	562,000	334,838,000
Bengali	10,086	10,083	936,666	615,264	584,644,773	401,984	429,246	495,398,451
Bishnupriya Manipuri	9,874	9,871	938,000	704,000	464,460,000	329,000	550,000	327,689,000
Bosnian	10,022	9,425	974,803	765,833	504,995,000	342,000	598,000	356,287,000
Breton	6,895	2,285	655,000	491,000	324,278,000	230,000	384,000	228,786,000
Bulgarian	10,181	10,180	988,989	754,797	498,239,000	347,000	590,000	351,521,000
Catalan	9,999	9,046	950,000	713,000	470,371,000	333,000	557,000	331,860,000
Cebuano	8,510	6,547	791,476	558,039	350,153,208	278,000	441,184	32,841,835
Central Bicolano	9,935	6,406	944,000	708,000	466,993,000	331,000	553,000	329,476,000
Chinese	3,315	3,314	292,521	221,305	146,094,000	103,000	173,000	103,073,000
Croatian	9,903	9,301	941,000	706,000	466,149,000	330,000	552,000	328,881,000
Czech	7,392	6,960	702,000	527,000	347,923,000	246,000	412,000	245,469,000
Danish	8,376	6,998	796,000	597,000	393,524,000	279,000	466,000	277,642,000
Dutch	9,877	8,082	957,554	814,109	507,544,154	335,186	663,886	337,735,314
English	263,098	N/A	24,596,102	18,447,000	12,170,541,000	8,629,000	14,412,000	8,586,641,000
Esperanto	8,024	7,336	762,000	572,000	377,479,000	267,000	447,000	266,322,000
Filipino (Tagalog)	9,436	8,063	906,702	693,036	456,860,000	318,000	541,000	322,327,000
Finnish	10,018	8,700	952,000	714,000	471,216,000	334,000	558,000	332,455,000
French	9,887	8,166	962,222	816,834	613,624,883	374,849	673,147	451,973,804
Frisian	6,383	4,497	606,000	455,000	299,788,000	213,000	355,000	211,508,000
Galician	9,987	8,763	949,000	712,000	469,527,000	333,000	556,000	331,264,000
Georgian	5,315	5,314	505,000	379,000	249,964,000	177,000	296,000	176,356,000
German	9,807	8,175	953,052	826,086	522,788,572	381,520	662,193	406,915,283
Greek	9,899	9,897	940,000	705,000	465,304,000	330,000	551,000	328,285,000
Gujarati	9,979	9,975	945,875	305,097	200,985,000	332,000	238,000	141,800,000
Haitian	9,188	5,865	873,000	655,000	432,370,000	306,000	512,000	305,049,000
Hebrew	8,195	8,195	779,000	584,000	385,080,000	273,000	456,000	271,684,000
Hindi	9,150	9,147	889,789	626,673	413,792,000	312,000	490,000	291,941,000
Hungarian	9,850	9,020	958,540	756,537	499,083,000	336,000	591,000	352,117,000
Icelandic	822	738	78,000	59,000	38,846,000	27,000	46,000	27,407,000
Ido	68	48	6,000	5,000	3,378,000	2,000	4,000	2,383,000
Ilokano	9,333	4,449	887,000	665,000	439,126,000	311,000	520,000	309,815,000
Indonesian	9,773	7,683	946,444	834,041	467,016,410	269,457	612,703	299,680,811
Irish	7,301	6,334	694,000	521,000	343,700,000	243,000	407,000	242,490,000
Italian	9,518	8,310	927,027	814,854	579,321,695	363,685	666,641	449,642,224
Japanese	8,071	8,071	767,000	575,000	379,168,000	269,000	449,000	267,513,000
Javanese	9,877	7,575	938,000	704,000	464,460,000	329,000	550,000	327,689,000
Kannada	9,924	9,921	943,000	707,000	466,149,000	331,000	552,000	328,881,000
Kapampangan	9,870	3,646	938,000	704,000	464,460,000	329,000	550,000	327,689,000
Kazakh	30	30	3,000	2,000	1,689,000	1,000	2,000	1,192,000
Korean	7,435	7,434	706,000	530,000	349,612,000	248,000	414,000	246,660,000
Kurdish	33	33	3,000	2,000	1,689,000	1,000	2,000	1,192,000
Latvian	9,939	9,585	962,034	604,692	398,591,000	337,000	472,000	281,217,000
Lithuanian	9,939	7,741	944,000	708,000	466,993,000	331,000	553,000	329,476,000
Low Saxon	7,344	5,637	698,000	524,000	345,389,000	245,000	409,000	243,681,000
Luxembourgish	6,609	4,545	628,000	471,000	310,766,000	220,000	368,000	219,254,000
Macedonian	10,095	9,972	959,000	719,000	474,594,000	336,000	562,000	334,838,000
Malagasy	164	159	16,000	12,000	7,600,000	6,000	9,000	5,362,000
Malay	9,351	7,823	888,000	666,000	439,126,000	312,000	520,000	309,815,000
Malayalam	10,124	10,124	962,000	722,000	476,283,000	337,000	564,000	336,030,000
Marathi	9,988	9,987	949,000	712,000	469,527,000	333,000	556,000	331,264,000
Neapolitan	4,493	3,441	427,000	320,000	211,118,000	150,000	250,000	148,950,000
Nepali	9,916	9,915	700,479	396,109	260,942,000	246,000	309,000	184,102,000
Newar	262	262	25,000	19,000	12,667,000	9,000	15,000	8,937,000
Norwegian (Nynorsk)	8,473	6,976	805,000	604,000	398,591,000	282,000	472,000	281,217,000
Norwegian	9,083	7,603	863,000	647,000	426,459,000	303,000	505,000	300,878,000
Pashto	331	331	31,000	23,000	15,201,000	11,000	18,000	10,724,000
Persian (Farsi)	921	921	87,843	76,894	50,668,000	31,000	60,000	35,748,000
Piedmontese	9,294	7,308	883,000	662,000	436,592,000	310,000	517,000	308,028,000
Polish	9,764	9,159	928,000	696,000	459,393,000	326,000	544,000	324,114,000
Portuguese	9,873	8,695	938,000	704,000	464,460,000	329,000	550,000	327,689,000
Punjabi	9,827	9,827	934,000	701,000	462,771,000	328,000	548,000	326,497,000
Romanian	9,880	8,819	963,377	758,086	499,928,000	338,000	592,000	352,712,000
Russian	9,962	9,958	946,000	710,000	468,682,000	332,000	555,000	330,668,000
Serbian	10,146	10,039	979,731	764,291	504,150,000	344,000	597,000	355,691,000
Serbo-Croatian	10,057	9,590	955,000	716,000	472,060,000	335,000	559,000	333,051,000
Sicilian	1,751	1,491	166,000	125,000	82,758,000	58,000	98,000	58,388,000
Sindhi	36	36	3,000	2,000	1,689,000	1,000	2,000	1,192,000
Slovak	9,939	9,283	893,962	648,120	427,303,000	314,000	506,000	301,474,000
Slovenian	7,927	7,354	753,000	565,000	372,412,000	264,000	441,000	262,747,000
Somali	9,907	7,177	904,728	579,501	382,546,000	317,000	453,000	269,897,000
Spanish	9,825	8,778	959,099	699,244	450,079,676	305,749	553,368	380,362,877
Sundanese	9,909	4,726	941,000	706,000	466,149,000	330,000	552,000	328,881,000
Swahili	7,019	6,132	666,805	500,104	330,189,000	234,000	391,000	232,957,000
Swedish	9,551	8,086	928,935	720,860	475,438,000	326,000	563,000	335,434,000
Tamil	9,449	9,448	901,355	511,136	336,945,000	316,000	399,000	237,723,000
Telugu	9,751	9,751	933,566	364,407	240,675,000	328,000	285,000	169,802,000
Thai	4,487	4,487	423,768	334,513	220,407,000	149,000	261,000	155,503,000
Turkish	10,007	9,263	984,243	781,376	483,687,039	372,415	609,274	439,236,149
Uighur	5,650	5,650	495,402	176,736	104,323,748	174,000	122,660	54,723,339
Ukrainian	10,027	9,990	972,517	720,864	475,438,000	341,000	563,000	335,434,000
Urdu	9,999	9,998	930,003	625,048	412,103,000	326,000	488,000	290,749,000
Uzbek	9,696	5,630	904,713	579,421	326,562,177	188,211	430,682	108,250,659
Vietnamese	5,911	4,586	558,470	494,680	325,966,000	196,000	386,000	229,978,000
Waray	8,489	5,368	806,000	605,000	399,436,000	283,000	473,000	281,812,000
Welsh	9,923	7,272	917,093	562,247	370,724,000	322,000	439,000	261,555,000
Wolof	45	36	4,000	3,000	1,689,000	1,000	2,000	1,192,000
Yoruba	1,802	1,523	140,134	92,334	60,802,000	49,000	72,000	42,897,000

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.

Source Language	Arabic		Dutch		French	
Detected Language	Arabic	.5543	Dutch	.5139	French	.5822
	English	.4189	English	.4539	English	.4038
	Persian	.0051	German	.0078	Spanish	.0022
	French	.0020	French	.0056	Norwegian	.0017
	Norwegian	.0015	Norwegian	.0022	German	.0013
Source Language	German		Italian		Spanish	
Detected Language	German	.5946	Italian	.5856	Spanish	.6144
	English	.3847	English	.3797	English	.3609
	Dutch	.0038	Spanish	.0109	Portuguese	.0115
	French	.0032	Portuguese	.0063	Galician	.0017
	Norwegian	.0019	French	.0036	Italian	.0015

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.

Source Language	Bengali		Cebuano		Indonesian	
Detected Language	Bengali	.7256	English	.8318	English	.4972
	English	.2300	Spanish	.0401	Indonesian	.4667
	Russian	.0160	Tagalog	.0264	Malay	.0114
	Tajik	.0083	Cebuano	.0227	Turkish	.0034
	Bulgarian	.0073	French	.0129	German	.0026
Source Language	Turkish		Uighur		Uzbek	
Detected Language	Turkish	.6772	Uighur	.6609	English	.6035
	English	.3077	English	.1140	Uzbek	.1882
	Spanish	.0013	Inupiaq	.0778	Russian	.1169
	Indonesian	.0011	Arabic	.0291	Turkish	.0290
	German	.0011	Persian	.0290	Azerbaijani	.0128

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.

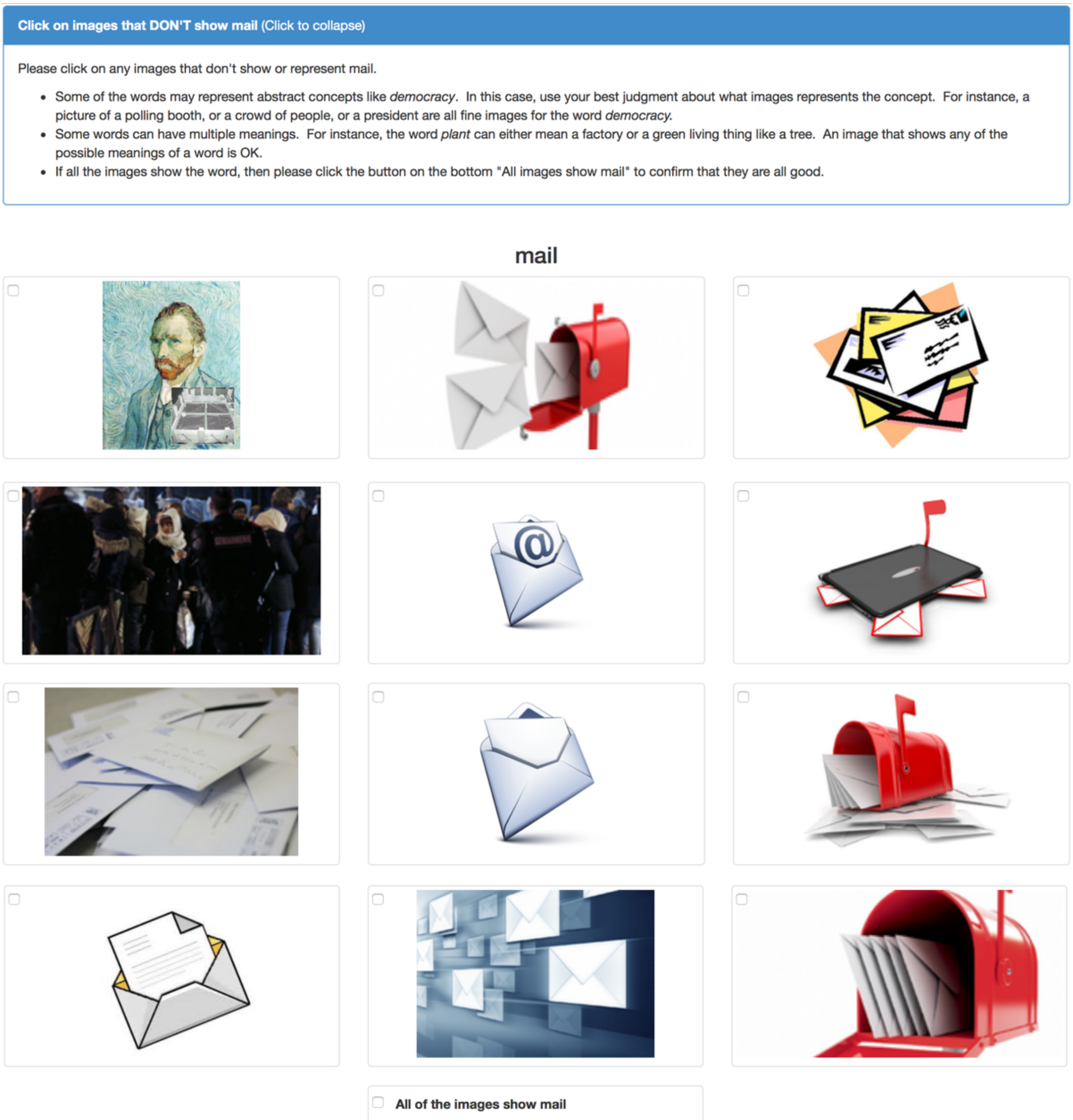

Figure 2: An example of the task we gave to the workers on Amazon Mechanical Turk, along with the instructions given. In this example, we can the first two images in the left column are the two controls we put in, which are not associated with the word “mail”.

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.

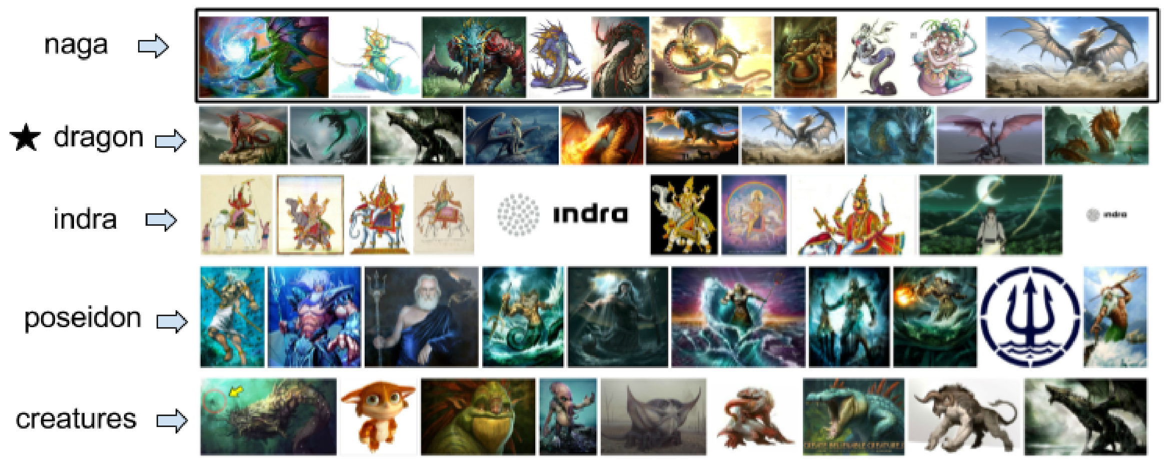

Figure 7: The top 4 ranked translations of the Indonesian word *naga* using CNN features.

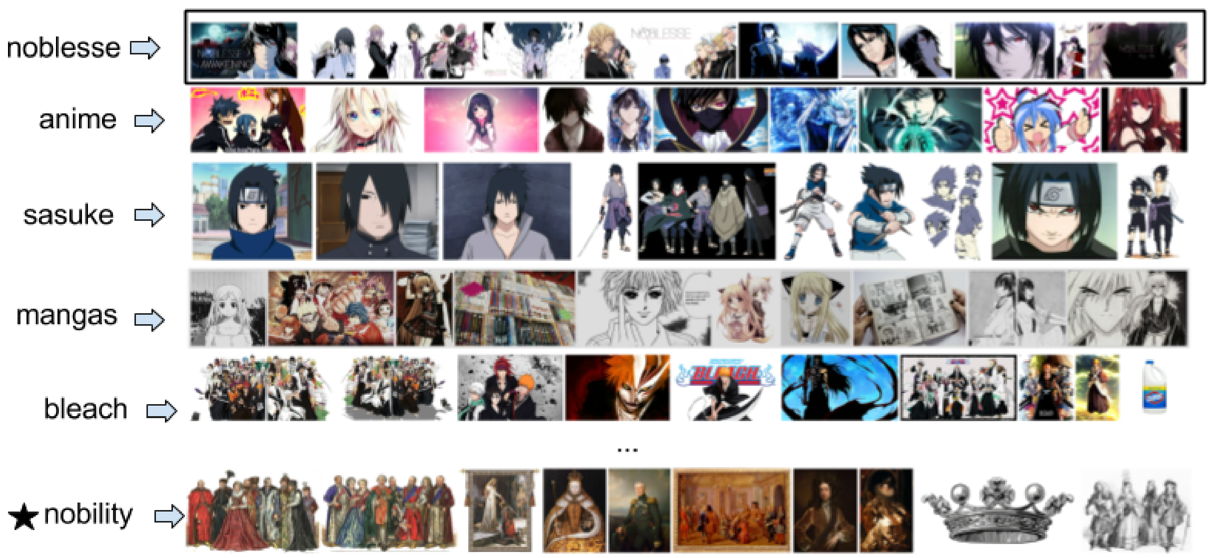

Figure 8: The top 4 ranked translations of the French word *noblesse* using CNN features.

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