Learning Bilingual Lexicons Using the Visual Similarity of Labeled Web Images

Shane Bergsma and Benjamin Van Durme
Department of Computer Science and Human Language Technology Center of Excellence
Johns Hopkins University
sbergsma@jhu.edu, vandurme@cs.jhu.edu

Abstract

Speakers of many different languages use the Internet. A common activity among these users is uploading images and associating these images with words (in their own language) as captions, filenames, or surrounding text. We use these explicit, monolingual, image-to-word connections to successfully learn implicit, bilingual, word-to-word translations. Bilingual pairs of words are proposed as translations if their corresponding images have similar visual features. We generate bilingual lexicons in 15 language pairs, focusing on words that have been automatically identified as physical objects. The use of visual similarity substantially improves performance over standard approaches based on string similarity: for generated lexicons with 1000 translations, including visual information leads to an absolute improvement in accuracy of 8-12% over string edit distance alone.

1 Introduction

Bilingual lexicon induction is the task of finding words or phrases across natural languages that share a common meaning. In the machine translation (MT) community, such translations are usually obtained from aligned parallel text. For most language pairs, and most domains, parallel data is unavailable, and therefore a range of methods have been developed to find translations directly from monolingual text [Fung and Yee, 1998; Rapp, 1999; Koehn and Knight, 2002; Haghighi *et al.*, 2008]. Bilingual lexicons have many uses beyond MT, e.g. in cross-language information retrieval.

To find translations using monolingual data, words are associated with information that is preserved across languages. Previous systems have exploited the similar spelling of translations in related languages [Koehn and Knight, 2002; Haghighi *et al.*, 2008], and their similar frequency distribution over time [Schafer and Yarowsky, 2002; Klementiev and Roth, 2006]. A seed lexicon has also been used to project context words from one language into another; translations are then identified as bilingual pairs of words with high contextual similarity [Fung and Yee, 1998; Rapp, 1999].

We exploit the universality of *visual* information to build bilingual lexicons. Billions of images are added to sites like Facebook and Flickr every month.[1] Users naturally label their images as they post them online, providing an explicit link between a word and its visual representation. Since images are labeled with words in many languages, we propose to generate word translations by finding pairs of words that have a high visual similarity between their respective image sets.

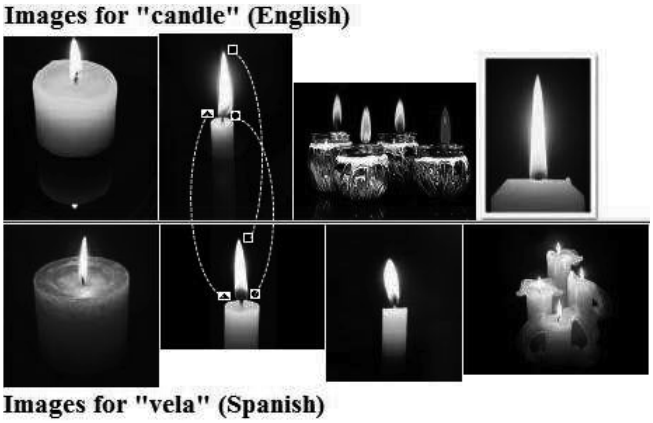

Figure 1: Matching words through their images: Images retrieved from the web for the English word *candle* (top) and the Spanish word *vela* (bottom). The matching between detected SIFT keypoints is shown for a pair of images.

Figure 1 illustrates our approach for a particular word pair. We use Google's image search to automatically acquire images for the words *candle* in English and *vela* in Spanish. We then use computer vision techniques to detect scale-invariant *keypoints* in each image. These keypoints are used to produce a visual similarity score for every *candle*/*vela* image pair. We generate a single score for *candle*/*vela* by combining the visual similarity across all image pairs. Using 20 images for each word, our approach ranks *vela* as the most likely translation for *candle* out of 500 translation candidates, despite there being no identical images shared by the two image sets.

To our knowledge, this is the first work to induce word translations through labeled images. An unexplored alternative to our approach would be to have (monolingual) speakers of different languages provide words for the *same* images.

[1] Facebook recently tweeted that over 750 million images were uploaded over the recent New Year's weekend alone: twitter.com/facebook/status/22372857292005376

For example, the monolingual speakers could play the ESP game [von Ahn and Dabbish, 2004] in different languages, but with the same set of images. Or, we might pay annotators to label images in their native language using online annotation services such as Amazon's Mechanical Turk. Unlike these alternatives, our approach can make use of the many billions of web images and labels that already exist.[2]

We show that visual similarity enables improvements over standard approaches to bilingual lexicon induction. We automatically determine a large class of *physical object* words where one would expect consistent visual representations across languages. We evaluate our method in a realistic and large-scale lexicon induction task using these words. We also show how our method can provide useful semantic information for resolving other, monolingual, linguistic ambiguities.

2 The visual similarity of bilingual words

For a given word, we automatically: (1) acquire a corresponding set of images, (2) extract visual features from these images, (3) compute the visual similarity of two words using their associated image sets, and (4) use this similarity to rank translation pairs for bilingual lexicon induction.[3]

2.1 Using image search engines

Search engines provide a natural way to collect labeled images, given the vast effort that has been expended to refine their widely-used image retrieval services. Search engines retrieve images based on the image caption, file-name, and surrounding text [Feng and Lapata, 2010]. To automatically retrieve images, we provide a word or phrase as an HTTP query to the search engine, and directly download the uniformly-sized thumbnails that are returned (rather than downloading the source images directly). For English words, we used Google's Image Search (`www.google.com/imghp`), while for foreign words, we used the corresponding foreign Google website (all with default settings). For experiments using W images for a given word (e.g., Figure 2(a) below), we take the first W images returned by Google. We used Google because previous research has shown that its results are competitive with "*hand prepared datasets*" [Fergus *et al.*, 2005]. Also, in related ongoing work, we achieve higher accuracy using Google images than using images obtained from Flickr.

2.2 Visual features

We convert each image to a representation based on a finite set of visual features. A range of visual features have been explored in the vision literature, usually in the context of supporting content-based image retrieval [Deselaers *et al.*, 2008]. Often such features correspond only to local parts of the image, and the spatial relationship between these parts is not modeled, analogous to the bag-of-words representation familiar to NLP researchers. We adopt this bag-of-words approach for our two types of features: color features and SIFT features.

[2]Our approach is also independent of the verbosity of a given annotator. Knowledgeable web users will naturally label pictures of *orioles*, *magpies* and *cockatoos*, whereas a solicited annotator might be inclined to tag all these image with the simple label *bird*.

[3]Scripts and experimental data are publicly available at: `www.clsp.jhu.edu/~sbergsma/LexImg/`

Color histogram

Deselaers et al. [2008] note that for image retrieval, the "*color histogram performs well ... and can be recommended as a simple baseline for many applications.*" To create a color histogram, we partition the color space and count the number of image pixels that occur in each partition. We partition colors using the first hexadecimal digit in each pixel's R, G and B values. This results in a 16^3=4096-dimensional vector space. Each color partition and its count is used as a feature dimension and its value, respectively, in this color vector space.

SIFT keypoints

SIFT keypoints are distinctive local image features that are invariant to scaling and rotation, and robust to illumination, noise and distortion [Lowe, 2004]. They are widely used in vision research, including work that intersects with NLP [Feng and Lapata, 2010]. We identify SIFT keypoints using David Lowe's publicly-available software: `www.cs.ubc.ca/~lowe/keypoints/`. SIFT features are taken from images converted to gray-scale. Figure 1 shows the location of SIFT keypoints detected in two images. We added arcs to illustrate keypoints that are close in key-point space.

Each SIFT keypoint is itself a multi-dimensional vector. We convert this bag-of-vectors into a bag-of-words representation by mapping each keypoint to a dimension in a quantized SIFT feature space. First, we cluster a random selection of 430 thousand keypoints (from our English image data) into K cluster centroids using the K-means algorithm. We found the final clustering distortion to be robust to different random initializations. Using the signal processing terminology, each resulting cluster centroid is a *codeword* in the K-dimensional SIFT *codebook*. To quantize the keypoints for a particular image, we map each keypoint to its nearest-neighbor codeword. Each dimension in the resulting feature vector corresponds to a codeword; each value is the count of the number of keypoints mapping to that word.

2.3 Combining image similarities

Let $\mathbf{e}$ and $\mathbf{f}$ be visual feature vectors for a pair of images. We measure the distance between these vectors using their cosine similarity: $\text{cosine}(\mathbf{e},\mathbf{f}) = \frac{\mathbf{e}\cdot\mathbf{f}}{|\mathbf{e}||\mathbf{f}|}$. Many distance functions have been used in the literature and improving this function could be fruitful future work (*cf.* [Deselaers *et al.*, 2008]).

Each word has a corresponding set of images. Let $\mathcal{E}$ and $\mathcal{F}$ denote two such sets in a source and target language. To produce a single word-to-word visual similarity score, $\text{sim}(\mathcal{E},\mathcal{F})$, we combine the similarities of all image pairs using one of two scoring functions: AVGMAX or MAXMAX.

For each $\mathbf{e} \in \mathcal{E}$, AVGMAX finds the best matching image in $\mathcal{F}$. It averages these top-matches to produce a single score:

$$\text{AVGMAX}(\mathcal{E},\mathcal{F}) = \frac{1}{|\mathcal{E}|}\sum_{\mathbf{e}\in\mathcal{E}} \max_{\mathbf{f}\in\mathcal{F}}(\text{cosine}(\mathbf{e},\mathbf{f})) \qquad (1)$$

MAXMAX, on the other hand, takes the single best matching image-to-image similarity as the word-to-word score:

$$\text{MAXMAX}(\mathcal{E},\mathcal{F}) = \max_{\mathbf{e}\in\mathcal{E}} \max_{\mathbf{f}\in\mathcal{F}}(\text{cosine}(\mathbf{e},\mathbf{f})) \qquad (2)$$

3 Creating a lexicon of physical objects

We assume that words for concrete objects, such as machines, tools and living things, will have consistent color and keypoint features in their associated images. Words that represent more abstract concepts, such as *procrastination*, *forgot* and *intolerant*, could be visually represented in myriad ways, or might have many irrelevant images in their automatically-compiled image sets. The latter words might therefore be problematic to visually-align across languages.

We therefore propose to initially focus on finding translations for *physical objects*: words that are both likely to occur in image labels and to have consistent visual representations. A multilingual lexicon of physical objects would have one obvious application: it could be used to extend the reach of multilingual image search engines [Etzioni *et al.*, 2007].

We propose automatic methods for creating a lexicon of physical objects. We first explore a precise but low-coverage pattern-based approach and then a higher-coverage but noisier approach based on distributional similarity with a seed lexicon. While our experiments use single-token words, extending our approach to phrases is straightforward.

3.1 Physical objects via pattern matching

We first collect English words filling the following pattern:

{image,photo,photograph,picture} of {a,an} ____

We require the filler to have a noun part-of-speech tag and the word after the filler to *not* have a noun part-of-speech tag.

We count how often each word fills this pattern in Lin et al [2010]'s web-scale, part-of-speech-tagged N-gram corpus. We rank words by their conditional probability of co-occurring with this pattern. We filter words that occur in the corpus as nouns less than 50% of the time; we also manually filtered 29 potentially offensive terms. After filtering, the top 500 remaining words were taken as our English lexicon.

The resulting lexicon contains many physical objects (like *helicopter*, *finger*, and *sword*), but also some more general or more abstract concepts: *organization*, *situation*, *logo*, and *product*. Matching these words based on their visual features represents a challenging task for our approach.

While it would be possible to apply this same process to other languages, we want to first evaluate the power of visual similarity independently of the quality of our approach's linguistic components. We thus built corresponding lexicons in foreign languages by directly translating the English words using Google Translate (`translate.google.com/`). We take the one-best translation returned by Google Translate and create lexicons in Spanish, German, French, Italian and Dutch. Since different English words may have the same foreign translation, the foreign lexicons can be less than 500 words.

We use Google Translate because it gives high-coverage translations for the 15 language pairs we experimented with.[4] However, note that using a single translation from Google Translate might miss translations for words with multiple senses, and thus make our task more difficult.

[4]We did not previously have electronic dictionaries for all these pairs. In Section 5 we also make use of in-house electronic dictionaries for evaluation in Spanish-English and French-English.

3.2 Physical objects via distributional similarity

The above patterns only identify a small fraction of the physical objects that might be amenable to visual representation. We create a larger list by finding words that occur in similar contexts to a seed list of physical objects, i.e., words that are distributionally similar. For example, our English seed list has the words *helicopter*, *motorcycle* and *truck*; the larger list has similar words *submarine*, *tractor*, and *lorry*.

We use a seed lexicon of 100 physical objects in each language. Our English seeds are the top 100 words as ranked by the pattern-based approach (excluding words occurring fewer than 50 times in the N-gram data). The foreign seed lists consist of the Google translations of the English seed list.

We exploit the availability of large corpora in each language to rank a list of unigrams by their contextual similarity with the seeds. Contextual similarity is defined as the cosine similarity between context vectors, where each vector gives the counts of words to the left and right of the target unigram. We get counts from English and foreign Google N-gram data [Lin *et al.*, 2010; Brants and Franz, 2009]. Rather than building the vectors explicitly, we use the locality-sensitive hash algorithm of Van Durme and Lall [2010] to build low-dimensional bit signatures in a streaming fashion. This allows for fast, approximate cosine computation. We rank the unigrams by their average similarity with their ten most-similar seeds. The top 20,000 highest-ranked unigrams comprise the final physical object lexicon in each language.

4 Experiments Part 1: 500-word lists

4.1 Set-up

Evaluation We first test on the 500-word lists created via pattern-matching (§ 3.1). Here, each source word, indexed by i, has a translation in each target lexicon; let this be at position $tr(i)$. For each source word's image set, $\mathcal{E}_i$, we rank all foreign image sets, $\mathcal{F}_j$, by their similarity with $\mathcal{E}_i$. The goal is to have $\mathcal{F}_{tr(i)}$ ranked highest, i.e., $\text{rank}_{\mathcal{E}_i}(\mathcal{F}_{tr(i)})=1$.

We use the following evaluation measures:

- **MRR**: Mean-reciprocal rank of correct translation: MRR = $\frac{1}{500}\sum_{i=1}^{500}\frac{1}{\text{rank}_{\mathcal{E}_i}(\mathcal{F}_{tr(i)})}$ (closer to 1 is better).
- **Top-N accuracy**: Proportion of instances where the correct translation occurs within the top N highest-ranked translations. We use N=1, 5 and 20.

Data We use our English-Spanish lists to perform preliminary experiments and to set the parameters of our algorithm (including the λ parameters described below). Our final results are the average MRR and Top-N accuracies across all pairs from English, Spanish, German, French, Italian and Dutch, excluding English-Spanish. Images for each language are collected and processed as described in § 2. The proposed rankings are evaluated against the Google translations.

Comparison approaches Let $w_{\mathcal{E}}$ and $w_{\mathcal{F}}$ be source and target word strings which have corresponding image sets $\mathcal{E}$ and $\mathcal{F}$. We compare the following similarity functions:

1. **Random**: Randomly score each $\mathcal{E}, \mathcal{F}$ pair.

System	MRR	Top-1	Top-5	Top-20
AVGMAX	**36.0**	**31.0**	**40.8**	**48.8**
MAXMAX	31.5	27.0	35.2	42.0

Table 1: 500-word lists experiment (%): AVGMAX performs better than MAXMAX on English-Spanish bilingual lexicon induction.

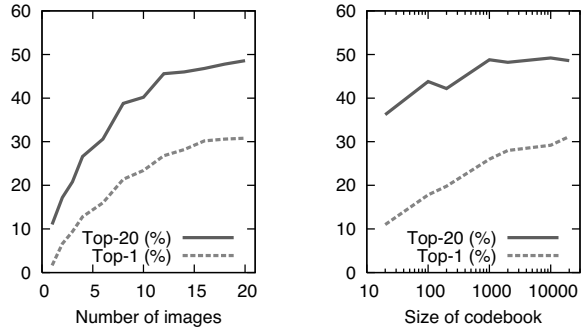

Figure 2: 500-word lists experiment: Performance of English-Spanish lexicon induction improves with (a) more images per word and (b) more codewords (clusters of SIFT keypoints).

2. **Color Histogram**: Compute visual similarity using color features only: $\text{sim}_{\text{color}}(\mathcal{E}, \mathcal{F})$.
3. **SIFTs**: Compute visual similarity using SIFT features only: $\text{sim}_{\text{SIFT}}(\mathcal{E}, \mathcal{F})$.
4. **SIFTs+Color**: Use a linear combination of the SIFT and color histogram similarities: $\text{sim}_{\text{SIFT}}(\mathcal{E}, \mathcal{F}) + \lambda_0 \text{sim}_{\text{color}}(\mathcal{E}, \mathcal{F})$.
5. **Normalized Edit Dist. (NED)**: Compute the character-level (orthographic) similarity of $w_{\mathcal{E}}$ and $w_{\mathcal{F}}$ using the widely-used edit distance measure. NED uses dynamic programming to compute the minimum number of insertions, deletions and substitutions needed to transform the source string $w_{\mathcal{E}}$ into the target string $w_{\mathcal{F}}$. It normalizes this edit distance by the length of the longer string.
6. **SIFTs+Color+NED**: Use a linear combination of the two visual and one orthographic measure: $\text{sim}_{\text{SIFT}}(\mathcal{E}, \mathcal{F}) + \lambda_1 \text{sim}_{\text{color}}(\mathcal{E}, \mathcal{F}) + \lambda_2 \text{NED}(w_{\mathcal{E}}, w_{\mathcal{F}})$

4.2 Part 1 results

We first provide results on our English-Spanish development data. We use this data to investigate three key components of our algorithm: the scoring function (default AVGMAX), the number of images in each image set (default 20) and the SIFT codebook dimensionality (default 20,000). For simplicity, we investigate these components using only SIFT features.

Table 1 shows that we get a consistent gain using AVGMAX rather than MAXMAX scoring. Our approach therefore leverages not just the exact image matches in the image sets, but aggregate information over many weaker matches.

The number of images that we use in each image set has a strong impact on both performance and efficiency (computing AVGMAX increases quadratically with the number of images in each image set). While the Top-1 accuracy plateaus around 20 images (Figure 2(a)), the Top-20 scores are still increasing, showing there is some value in later image results. Performance (and computation) also increases with the number of codewords in the SIFT codebook (Figure 2(b)), up to around 1000 codewords (note the x-axis is on a logarithmic scale). Beyond 1000 codewords, Top-20 accuracy plateaus while Top-1 accuracy increases. Using more codewords results in a more specific visual representation, meaning that more general similarities between keypoints might be missed, but false positive matches are reduced.

System	MRR	Top-1	Top-5	Top-20
Random	1.4	0.2	0.9	4.1
Color Histogram	19.6	14.4	23.2	35.6
SIFTs	32.1	27.4	35.7	45.3
SIFTs+Color	36.7	31.1	41.4	*53.7*
Normalized Edit Dist.	*41.7*	*37.3*	*45.8*	52.9
SIFTs+Color+NED	**53.6**	**48.0**	**59.5**	**68.7**

Table 2: 500-word lists experiment: Average lexicon induction performance (%) across all pairs within {German, English, Spanish, French, Italian, Dutch}. Top score in **bold**, second-highest in *italics*. Combining visual and orthographic similarity performs best.

Table 2 provides final results averaged over the other 14 language pairs, using default settings for the above components. SIFT features are more powerful than colors, but their combination achieves even better results. The full visual system (SIFTs+Color) is competitive with Normalized Edit Dist., and even exceeds its Top-20 accuracy. Since visual and orthographic similarity provide such complementary information, the SIFTs+Color+NED combination works much better than either visual or orthographic similarity on its own, and achieves the top result on all measures (in **bold**). Note the Top-1 accuracy of this system: across 14 language pairs, the correct translation is the first one proposed (of 500 candidates) in nearly half the cases.

5 Experiments Part 2: 20,000-word lists

5.1 Set-up

Evaluation We now create bilingual lexicons using the 20,000-word lists. These lists consist of words that are distributionally similar to a seed list of 100 physical objects in each language (§ 3.2). We conduct experiments to generate English-Spanish and English-French lexicons. For these experiments, it is not the case that every English word has a translation in the foreign lexicon. According to our gold standard lexicons (below), only 24% of the English words have a Spanish translation, and only 21% have a French translation. The task is now to detect these correct translations within the 400 million possible pairs. We therefore choose a different evaluation: Given a proposed list of the M most-confident translations, what proportion are correct? We compare systems by plotting these proportions for different values of M.[5]

[5]Note how the scale of our Part 2 experiments compares to previous work. Koehn and Knight [2002] evaluate on the 1,000 most-frequent English and 1,000 most-frequent German nouns, while Haghighi et al [2008] evaluate on the 2,000 most-frequent English and foreign nouns. By focusing on only the most-frequent nouns, these approaches use data where lots of distributional information

Data The 20,000-word lists are generated as described in § 3.2. For each word in each language, 20 images are downloaded and processed as described in § 2, resulting in a target of 400,000 images for each language (but not all words return a full set of 20 images). For efficiency reasons, we use the SIFT features, but not the color features, in these experiments. For computing similarity, we parallelize the roughly $400K^2$=160 billion cosine computations.

We compile a gold-standard translation lexicon for evaluation via two sources. First, we include all entries in several in-house electronic Spanish-English and French-English dictionaries. Second, we use Google translate in two directions: (A) to convert every English-list word to its foreign translation, and (B) to translate every foreign-list word to its English translation. Unfortunately, the 20,000 lists include many typos and other rare strings. Since Google translate passes out-of-vocabulary words verbatim, we exclude any verbatim translations from our gold standard as unreliable. To prevent these exclusions from distorting our results, we only include a proposed translation in our results if both the English and foreign words occur at least once in our gold standard (of course, they need not occur together in a translation pair). To be clear: this only removes pairs where both the English and foreign words are translated verbatim by Google translate, and neither occurs in our in-house lexicons.

Comparison approaches We compare visual similarity using SIFT features (SIFTs) and orthographic similarity using Normalized Edit Distance (NED). For efficiency, we only retain the top 1000 most-similar words for each English word. For the SIFT similarity, we use the default scoring function, number of images, and codebook size from Part 1. We compare systems based on the visual and orthographic measures on their own, and a joint system that simply sums the two similarities over their individual top-1000 lists (SIFTs+NED).

5.2 Part 2 results

Figures 3(a) and 3(b) show results on English-Spanish and English-French lexicon induction. Here, NED strongly outperforms visual similarity alone (SIFTs), reflecting the smaller proportion of physical objects in the 20,000-word lists, and hence the greater difficulty of visual matching. However, when we combine visual and orthographic similarity, we achieve substantial improvements: when proposing 1000 translations, we get an absolute improvement of 12% (Spanish) and 8% (French) over using orthographic similarity alone. Remarkably, without any manual involvement beyond the 100 seed words, we are able to generate 1000 translations with 80% precision in French and 70% in Spanish.

Table 3 provides some specific examples of similarities computed using the visual features. Note that being able to propose correct translations for low-frequency nouns like *rosary* and *fishhook* is a major advance over previous work.

is available for the lexical items (hence datasets favorable to their methods). We increase the scope of the lexicons by *an order-of-magnitude*: finding matches across 20,000 English and foreign nouns. While we focus on physical objects, we actually attempt something that is much wider in scope than previous work.

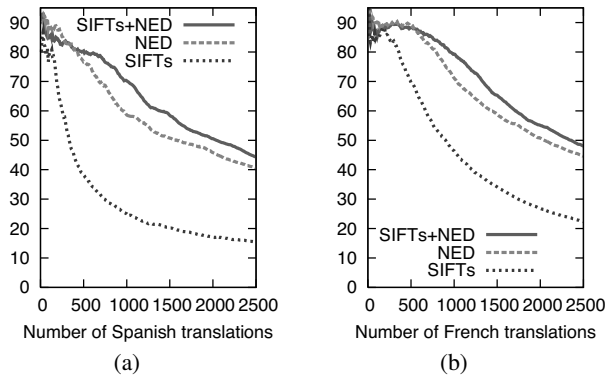

Figure 3: 20,000-word-lists experiment: Precision (%) of induced lexicons in (a) English-Spanish and (b) English-French. Adding visual similarity (SIFTs) improves over string similarity alone (NED).

In previous approaches, there would be insufficient statistical or orthographic information to enable discovery of these translation pairs. Indeed, these terms might not be found even in plentiful *parallel* text.

We also computed the matching of the English list against itself (excluding identical word matches) and present examples in the third column.[6] While visual similarity alone is rarely definitive, these results suggest that together with other indicators, visual similarity might provide helpful semantic information for detecting morphologically-related forms (e.g. *hurricanes-hurricane*), correcting spelling errors (*rosery-rosary*), and identifying semantic relationships (*fishhook* and *boathook* are taxonomic cousins: both are *hooks*).

6 Discussion and Future Work

Our work differs from approaches for *image annotation*: automatically labeling images or image parts with words or phrases [Barnard *et al.*, 2003; Lavrenko *et al.*, 2003; Feng and Lapata, 2010]. We do not analyze the image to determine applicable words; we instead rely on user-provided annotations. We focus on matching images with other images, and we use the image-image matches to link word labels. However, we can still benefit from advances in image annotation; any improvements in the monolingual word-image links will result in better image sets, and thereby better overall word-word visual similarities. In particular, advances in image annotation might allow us to do better on abstract concept words. Recent work has aimed to go beyond "*keywords*," to identify the "*attributes, relations and activities*" in images [Hodosh *et al.*, 2010]. As recognition of these improves, finding the translation of adjectives, abstract nouns and verbs could improve in tandem.

Our ultimate aim is to use visual features, along with other semantic indicators, to jointly learn bilingual correspondences and monolingual semantic relations. Beyond construction of the 20,000-word lists, our current approach does not leverage the lexico-semantic information given by

[6]Note *hurricane* matches perfectly with *hurricaine*; Google corrects the latter spelling to the former and returns identical images.

Word	Spanish	French	English
hurricane	**huracán**:0.14 **huracan**:0.08 borrasca:0.05 tsunami:0.05	**ouragan**:0.06 météorologie:0.06 tsunami:0.05 cyclone:0.05	**hurricaine**:1.00 **hurricaines**:0.28 **huricanes**:0.28 tsunami:0.05
rosary	**camándula**:0.15 puntaje:0.14 accidentalidad:0.14	**chapelet**:0.21 activité:0.15 **rosaire**:0.15 chatoiement:0.15	**rosery**:0.17 docment:0.15 precompensator:0.14 octonions:0.14
fishhook	**anzuelo**:0.13 densí-metro:0.13 chaira:0.12 pincel:0.12	**hameçon**:0.12 baton:0.11 binette:0.11 pinceau:0.11	sjambok:0.12 mangalsutra:0.12 baton:0.11 boathook:0.11

Table 3: 20,000-word-lists experiment: Examples of visually-similar words in different languages, ordered by similarity score. Correct translations in bold. Visual similarity correctly identifies translations that would be missed using string similarity (*fishhook-anzuelo*), and also finds morphologically or semantically-related words in English (*fishhook-boathook*).

frequency, contextual-similarity, etc., that was found to improve performance in previous studies. Monolingual *visual*-semantic information might also be exploited. For example, if *fishhook* and *boathook* are visually similar in English, their foreign translations should also be visually similar. Related ideas (using text) have been explored for inducing bilingual lexicons [Koehn and Knight, 2002] and building semantic taxonomies [Snow *et al.*, 2006], but not as a single combined model. Also, while large-scale efforts like ImageNet are currently linking images to words in a semantic taxonomy [Deng *et al.*, 2009], visual features have not yet been exploited to help build and extend the taxonomy itself.

7 Conclusion

We have shown that it is possible to use labeled web images to improve the performance of bilingual lexicon induction. We presented results for a number of languages and experimental settings, and investigated key parameters such as the similarity scoring function, the number of images per word, and the number of codewords in the visual codebook. Visual similarity provides substantial gains over orthographic similarity alone, even on related languages where orthographic similarity is known to be effective. On unrelated language pairs (like English-Hindi or Arabic-Chinese) the benefits of visual similarity would be even greater.

References

[Barnard *et al.*, 2003] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. *JMLR*, 3:1107–1135, 2003.

[Brants and Franz, 2009] T. Brants and A. Franz. Web 1T 5-gram, 10 European Languages, Version 1. LDC2009T25, 2009.

[Deng *et al.*, 2009] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

[Deselaers *et al.*, 2008] T. Deselaers, D. Keysers, and H. Ney. Features for image retrieval: an experimental comparison. *Information Retrieval*, 11:77–107, 2008.

[Etzioni *et al.*, 2007] O. Etzioni, K. Reiter, S. Soderland, and M. Sammer. Lexical translation with application to image search on the web. In *MT Summit XI*, 2007.

[Feng and Lapata, 2010] Y. Feng and M. Lapata. Topic models for image annotation and text illustration. In *HLT-NAACL*, 2010.

[Fergus *et al.*, 2005] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from Google's Image Search. In *ICCV*, 2005.

[Fung and Yee, 1998] P. Fung and L. Y. Yee. An IR approach for translating new words from nonparallel, comparable texts. In *COLING-ACL*, 1998.

[Haghighi *et al.*, 2008] A. Haghighi, P. Liang, T. Berg-Kirkpatrick, and D. Klein. Learning bilingual lexicons from monolingual corpora. In *ACL-08: HLT*, 2008.

[Hodosh *et al.*, 2010] M. Hodosh, P. Young, C. Rashtchian, and J. Hockenmaier. Cross-caption coreference resolution for automatic image understanding. In *CoNLL*, 2010.

[Klementiev and Roth, 2006] A. Klementiev and D. Roth. Named entity transliteration and discovery from multilingual comparable corpora. In *HLT-NAACL*, 2006.

[Koehn and Knight, 2002] P. Koehn and K. Knight. Learning a translation lexicon from monolingual corpora. In *ACL Workshop on Unsupervised Lexical Acquistion*, 2002.

[Lavrenko *et al.*, 2003] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *NIPS*, 2003.

[Lin *et al.*, 2010] D. Lin, K. Church, H. Ji, S. Sekine, D. Yarowsky, S. Bergsma, K. Patil, E. Pitler, R. Lathbury, V. Rao, K. Dalwani, and S. Narsale. New tools for web-scale N-grams. In *LREC*, 2010.

[Lowe, 2004] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60:91–110, 2004.

[Rapp, 1999] R. Rapp. Automatic identification of word translations from unrelated English and German corpora. In *ACL*, 1999.

[Schafer and Yarowsky, 2002] C. Schafer and D. Yarowsky. Inducing translation lexicons via diverse similarity measures and bridge languages. In *CoNLL*, 2002.

[Snow *et al.*, 2006] R. Snow, D. Jurafsky, and A. Y. Ng. Semantic taxonomy induction from heterogenous evidence. In *COLING-ACL*, 2006.

[Van Durme and Lall, 2010] B. Van Durme and A. Lall. Online generation of locality sensitive hash signatures. In *ACL Short Papers*, 2010.

[von Ahn and Dabbish, 2004] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *CHI*, 2004.