

A Rose is a Roos is a Ruusu: Querying Translations for Web Image Search

Janara Christensen Mausam Oren Etzioni

Turing Center

Dept. of Computer Science and Engineering

University of Washington, Seattle, WA 98105 USA

{janara, mausam, etzioni} @cs.washington.edu

Abstract

We query Web Image search engines with words (*e.g.*, spring) but need images that correspond to particular senses of the word (*e.g.*, flexible coil). Querying with polysemous words often yields unsatisfactory results from engines such as Google Images. We build an image search engine, IDIOM, which improves the quality of returned images by focusing search on the desired *sense*. Our algorithm, instead of searching for the original query, searches for multiple, automatically chosen translations of the sense in several languages. Experimental results show that IDIOM outperforms Google Images and other competing algorithms returning 22% more relevant images.

1 Introduction

One out of five Web searches is an image search (Basu, 2009). A large subset of these searches is subjective in nature, where the user is looking for different images for a single concept (Linsley, 2009). However, it is a common user experience that the images returned are not relevant to the intended concept. Typical reasons include (1) existence of homographs (other words that share the same spelling, possibly in another language), and (2) polysemy, several meanings of the query word, which get merged in the results.

For example, the English word 'spring' has several senses – (1) the season, (2) the water body, (3) spring coil, and (4) to jump. Ten out of the first fifteen Google images for *spring* relate to the season sense, three to water body, one to coil and none to the jumping sense. Simple modifications to query do not always work. Searching for *spring water* results in many images of bottles of spring water and searching for *spring jump* returns only three images (out of fifteen) of someone jumping.

Polysemous words are common in English. It is estimated that average polysemy of English is more than 2 and average polysemy of common English words is much higher (around 4). Thus, it is not surprising that polysemy presents a significant limitation in the context of Web Search. This is especially pronounced for image search where query modification by adding related words may not help, since, even though the new words might be present on the page, they may not be all associated with an image.

Recently Etzioni *et al.* (2007) introduced PAN-IMAGES, a novel approach to image search, which presents the user with a set of translations. *E.g.*, it returns 38 translations for the coil sense of spring. The user can query one or more translations to get the relevant images. However, this method puts the onus of choosing a translation on the user. A typical user is unaware of most properties of languages and has no idea whether a translation will make a good query. This results in an added burden on the user to try different translations before finding the one that returns the relevant images.

Our novel system, IDIOM, removes this additional burden. Given a desired sense it *automatically* picks the good translations, searches for associated images and presents the final images to the user. For example, it automatically queries the French *ressort* when looking for images of spring coil. We make the following contributions:

- We automatically learn a predictor for "good" translations to query given a desired sense. A good translation is one that is monosemous and is in a major language, *i.e.*, is expected to yield a large number of images.
- Given a sense we run our predictor on all its translations to shortlist a set of three translations to query.
- We evaluate our predictor by comparing the images that its shortlists return against the

images that several competing methods return. Our evaluation demonstrates that IDIOM returns at least one good image for 35% more senses (than closest competitor) and overall returns 22% better images.

2 Background

IDIOM makes heavy use of a sense disambiguated, vastly multilingual dictionary called PANDICTIONARY (Mausam et al., 2009). PANDICTIONARY is automatically constructed by probabilistic inference over a graph of translations, which is compiled from a large number of multilingual and bilingual dictionaries. For each sense PANDICTIONARY provides us with a set of translations in several languages. Since it is generated by inference, some of the asserted translations may be incorrect – it additionally associates a probability score with each translation. For our work we choose a probability threshold such that the overall precision of the dictionary is 0.9 (evaluated based on a random sample). PANDICTIONARY has about 80,000 senses and about 1.8 million translations at precision 0.9.

We use Google Image Search as our underlying image search engine, but our methods are independent of the underlying search engine used.

3 The IDIOM Algorithm

At the highest level IDIOM operates in three main steps: (1) Given a new query q it looks up its various senses in PANDICTIONARY. It displays these senses and asks the user to select the intended sense, s_q . (2) It runs Algorithm 1 to shortlist three translations of s_q that are expected to return high quality images. (3) It queries Google Images using the three shortlisted translations and displays the images. In this fashion IDIOM searches for images that are relevant to the intended concept as opposed to using a possibly ambiguous query.

The key technical component is the second step – shortlisting the translations. We first use PANDICTIONARY to acquire a set of high probability translations of s_q . We run each of these translations through a learned classifier, which predicts whether it will make a good query, *i.e.*, whether we can expect images relevant to this sense if queried using this translation. The classifier additionally outputs a confidence score, which we use to rank the various translations. We pick the top three translations, as long as they are above a

minimum confidence score, and return those as the shortlisted queries. Algorithm 1 describes this as a pseudo-code.

Algorithm 1 findGoodTranslationsToQuery(s_q)

```

1: translations = translations of  $s_q$  in PANDICTIONARY
2: for all  $w \in translations$  do
3:    $pd = \text{getPanDictionaryFeatures}(w, s_q)$ 
4:    $g = \text{getGoogleFeatures}(w, s_q)$ 
5:    $conf[w] = \text{confidence in } \text{Learner.classify}(pd, g)$ 
6: sort all words  $w$  in decreasing order of  $conf$  scores
7: return top three  $w$  from the sorted list

```

3.1 Features for Classifier

What makes a translation w good to query? A desired translation is one that (1) is in a high-coverage language, so that the number of images returned is large, (2) monosemously expresses the intended sense s_q , or at least has this sense as its dominant sense, and (3) does not have homographs in other languages. Such a translation is expected to yield images relevant to only the intended sense. We construct several features that provide us evidence for these desired characteristics. Our features are automatically extracted from PANDICTIONARY and Google.

For the first criterion we restrict the translations to a set of high-coverage languages including English, French, German, Spanish, Chinese, Japanese, Arabic, Russian, Korean, Italian, and Portuguese. Additionally, we include the *language* as well as *number of documents returned by Google search of w* as features for the classifier.

To detect if w is monosemous we add a feature reflecting the degree of polysemy of w : the *number of PANDICTIONARY senses that w belongs to*. The higher this number the more polysemous w is expected to be. We also include the *number of languages that have w in their vocabulary*, thus, adding a feature for the degree of homography.

PANDICTIONARY is arranged such that each sense has an English source word. If the source word is part of many senses but s_q is much more popular than others or s_q is ordered before the other senses then we can expect s_q to be the dominant sense for this word. We include features like *size of the sense* and *order of the sense*.

Part of speech of s_q is another feature. Finally we also add the *probability score* that w is a translation of s_q in our feature set.

3.2 Training the Classifier

To train our classifier we used Weka (Witten and Frank, 2005) on a hand labeled dataset of 767 ran-

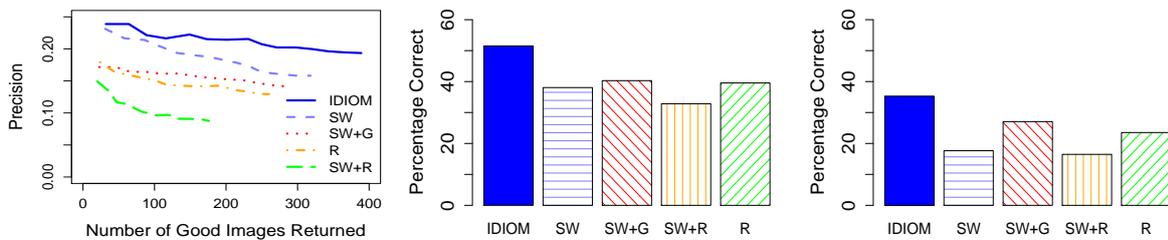


Figure 1: (a): Precision of images vs. the number of relevant images returned. IDIOM covers the maximum area. (b,c) The percentage of senses for which at least one relevant result was returned, for (b) all senses and (c) for minor senses of the queries.

domly chosen word sense pairs (e.g., pair of ‘primavera,’ and ‘the season spring’). We labeled a pair as positive if googling the word returns at least one good image for the sense in the top three. We compared performance among a number of machine learning algorithms and found that Random Forests (Breiman, 2001) performed the best overall with 69% classification accuracy using ten fold cross validation versus 63% for Naive Bayes and 62% for SVMs. This high performance of Random Forests mirrors other past experiments (Caruana and Niculescu-Mizil, 2006).

Because of the ensemble nature of Random Forests it is difficult to inspect the learned classifier for analysis. Still, anecdotal evidence suggests that the classifier is able to learn an effective model of good translations. We observe that it favors English whenever the English word is part of one or few senses – it picks out *auction* when the query is ‘sale’ in the sense of “act of putting up for auction to highest bidder”. In cases where English is more ambiguous it chooses a relatively less ambiguous word in another language. It chooses the French word *ressort* for finding ‘spring’ in the sense of coil. For the query ‘gift’ we notice that it does not choose the original query. This matches our intuition, since gift has many homographs – the German word ‘Gift’ means poison or venom.

4 Experiments

Can querying translations instead of the original query improve the quality of image search? If so, then how much does our classifier help compared to querying random translations? We also analyze our results and study the variation of image quality along various dimensions, like part of speech, abstractness/concreteness of the sense, and ambiguity of the original query.

As a comparison, we are interested in how IDIOM performs in relation to other methods for querying Google Images. We compare IDIOM to several methods. (1) *Source Word (SW)*: Querying with only the source word. This comparison func-

tions as our baseline. (2) *Source Word + Gloss (SW+G)*: Querying with the source word and the gloss for the sense¹. This method is one way to focus the source word towards the desired sense. (3) *Source Word + Random (SW+R)*: Querying with three pairs of source word and a random translation. This is another natural way to extend the baseline for the intended sense. (4) *Random (R)*: Querying with three random translations. This tests the extent to which our classifier improves our results compared to randomly choosing translations shown to the user in PANIMAGES.

We randomly select fifty English queries from PANDICTIONARY and look up all senses containing these in PANDICTIONARY, resulting in a total of 134 senses. These queries include short word sequences (e.g., ‘open sea’), mildly polysemous queries like ‘pan’ (means Greek God and cooking vessel) and highly polysemous ones like ‘light’.

For each sense of each word, we query Google Images with the query terms suggested by each method and evaluate the top fifteen results. For methods in which we have three queries, we evaluate the top five results for each query. We evaluate a total of fifteen results because Google Images fits fifteen images on each page for our screen size.

Figure 1(a) compares the precision of the five methods with the number of good images returned. We vary the number of images in consideration from 1 to 15 to generate various points in the graph. IDIOM outperforms the others by wide margins overall producing a larger number of good images and at higher precision. Surprisingly, the closest competitor is the baseline method as opposed to other methods that try to focus the search towards the intended sense. This is probably because the additional words in the query (either from gloss or a random translation) confuse Google Images rather than focusing the search. IDIOM covers 41% more area than SW. Overall

¹PANDICTIONARY provides a gloss (short explanation) for each sense. E.g., a gloss for ‘hero’ is ‘role model.’

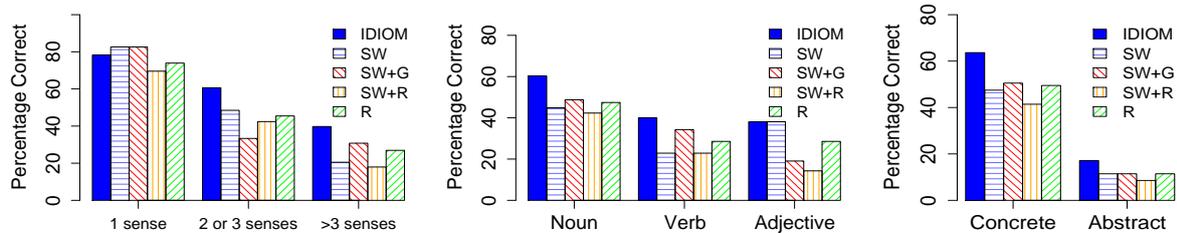


Figure 2: The percentage of senses for which at least one relevant result was returned varied along several dimensions: (a) polysemy of original query, and (b) part of speech of the sense, (c) abstractness/concreteness of the sense.

IDIOM produces 22% better images compared to SW (389 vs 318).

We also observe that random translations return much worse images than IDIOM suggesting that a classifier is essential for high quality images.

Figure 1(b) compares the percentage of senses for which at least one good result was returned in the fifteen. Here IDIOM performs the best at 51%. Each other method performs at about 40%. The results are statistically highly significant ($p < 0.01$).

Figure 1(c) compares the performance just on the subset of the non-dominant senses of the query words. All methods perform worse than in Figure 1(b) but IDIOM outperforms the others.

We also analyze our results across several dimensions. Figure 2(a) compares the performance as a function of polysemy of the original query. As expected, the disparity in methods is much more for high polysemy queries. Most methods perform well for the easy case of unambiguous queries.

Figure 2(b) compares along the different parts of speech. For nouns and verbs, IDIOM returns the best results. For adjectives, IDIOM and SW perform the best. Overall, nouns are the easiest for finding images and we did not find much difference between verbs and adjectives.

Finally, Figure 2(c) reports how the methods perform on abstract versus concrete queries. We define a sense as abstract if it does not have a natural physical manifestation. For example, we classify ‘nest’ (a bird built structure) as concrete, and ‘confirm’ (to strengthen) as abstract. IDIOM performs better than the other methods, but the results vary massively between the two categories.

Overall, we find that our new system consistently produces better results across the several dimensions and various metrics.

5 Related Work and Conclusions

Related Work: The popular paradigm for image search is keyword-based, but it suffers due to polysemy and homography. An alternative paradigm is content based (Datta et al., 2008), which is very

slow and works on simpler images. The field of cross-lingual information retrieval (Ballesteros and Croft, 1996) often performs translation-based search. Other than PANIMAGES (which we outperform), no one to our knowledge has used this for image search.

Conclusions: The recent development of PANDICTIONARY (Mausam et al., 2009), a sense-distinguished, massively multilingual dictionary, enables a novel image search engine called IDIOM. We show that querying unambiguous translations of a sense produces images for 35% more concepts compared to querying just the English source word. In the process we learn a classifier that predicts whether a given translation is a good query for the intended sense or not. We plan to release an image search website based on IDIOM. In the future we wish to incorporate knowledge from WordNet and cross-lingual links in Wikipedia to increase IDIOM’s coverage beyond the senses from PANDICTIONARY.

References

- L. Ballesteros and B. Croft. 1996. Dictionary methods for cross-lingual information retrieval. In *DEXA Conference on Database and Expert Systems Applications*.
- Dev Basu. 2009. How To Leverage Rich Media SEO for Small Businesses. In *Search Engine Journal*. <http://www.searchenginejournal.com/rich-media-small-business-seo/9580>.
- L. Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.
- R. Caruana and A. Niculescu-Mizil. 2006. An empirical comparison of supervised learning algorithms. In *ICML’06*, pages 161–168.
- R. Datta, D. Joshi, J. Li, and J. Wang. 2008. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2):1–60.
- O. Etzioni, K. Reiter, S. Soderland, and M. Sammer. 2007. Lexical translation with application to image search on the Web. In *Machine Translation Summit XI*.
- Peter Linsley. 2009. Google Image Search. In *SMX West*.
- Mausam, S. Soderland, O. Etzioni, D. Weld, M. Skinner, and J. Bilmes. 2009. Compiling a massive, multilingual dictionary via probabilistic inference. In *ACL’09*.
- I. Witten and E. Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.