# Word Sense Disambiguation with Pictures

**Kobus Barnard, Matthew Johnson**
Department of Computing Science,
721 Gould-Simpson, University of Arizona,
Tucson, Arizona, 85721-0077
{kobus, mjohnson}@cs.arizona.edu

**David Forsyth**
Computer Science Division,
University of California at Berkeley,
387 Soda Hall #1776
Berkeley California, 94720-1776
daf@cs.berkeley.edu

## Abstract

We introduce a method for using images for word sense disambiguation, either alone, or in conjunction with traditional text based methods. The approach is based in recent work on a method for predicting words for images which can be learned from image datasets with associated text. When word prediction is constrained to a narrow set of choices such as possible senses, it can be quite reliable, and we use these predictions either by themselves or to reinforce standard methods. We provide preliminary results on a subset of the Corel image database which has three to five keywords per image. The subset was automatically selected to have a greater portion of keywords with sense ambiguity and the word senses were hand labeled to provide ground truth for testing. Results on this data strongly suggest that images can help with word sense disambiguation.

## 1 Introduction

In this paper we investigate using words and pictures to disambiguate each other. Word sense disambiguation has long been studied as an important problem in natural language processing (Agirre and Rigau, 1995; Gale et al., 1992; Manning and Schütze, 1999; Mihalcea and Moldovan., 1998; Traupman and Wilensky, 2003; Yarowsky, 1995). It is illustrated in Figure 1 with the arguably overused "bank" example. A priori, the word "bank" has a number of meanings including financial institution and a step or edge as in "snow bank" or "river bank". Words which are spelt the same but have different meanings are very common, and clearly can confuse attempts to automatically deduce meaning from language. Furthermore, they cannot simply be identified as ambiguous and then ignored, as there are too many such words and they do constrain the possible meanings of a body of text.

Since the words are spelt the same, resolving what they mean requires considering context. A purely natural language based approach considers words near the one in question. Thus in the bank example, words like "financial" or "money" are strong hints that the financial institution sense is meant. Interestingly, despite much work, and a number of innovative ideas, doing significantly better than choosing the most common sense is difficult (Traupman and Wilensky, 2003).

In this work we present preliminary work on whether an associated images can help in word sense disambiguation. In the simplest application, text and images might be analyzed in conjunction; for example, a news photograph with a caption, or a larger document with illustrations.

piggy **bank** coins currency money

water grass trees **bank**s

**bank** buildings trees city

**bank** machine money currency bills

snow **bank**s hills winter

Figure 1. Word sense ambiguity in the Corel dataset.

## 2 Predicting Words from Images

To integrate image information with text data we exploit our previous work on linking images and words (Barnard et al., 2001; Barnard et al., 2003; Barnard and Forsyth, 2001; Duygulu et al., 2002). We have developed a variety of methods which can be used to predict words for image regions (region-labeling), and entire images (auto-annotation). This is achieved in practice by exploiting large image data sets with associated text. Critically, we do not require that the text be associated with the image regions, as such data is rare. Region labeling is illustrated in Figure 2. It is important to understand that we compute a posterior over the complete vocabulary for each region (and/or image), but for illustration we show the word for each region which has maximal probability.

For the results reported in this paper we use a special case of one of the models in (Barnard et al., 2003). Specifically, we model the joint probability of words and images regions as being generated by a collection of nodes, each of which has a probability distribution over both words and regions. The word probabilities are provided by simple frequency tables, and the region probability distribution are Gaussians over feature vectors. We restrict the Gaussians to have diagonal covariance (features are modeled as being independent).

Given an image region, its features imply a probability of being generated from each node. These probabilities are then used to weight the nodes for word emission. Thus words are emitted conditioned on image regions. In order to emit words for an entire image (auto-annotation), as needed for our word sense disambiguation method, we simply sum the distributions for the N largest regions. Thus each region is given equal weight, and the image words are forced to be generated through region labeling.

To be consistent with the more general models referenced above, we index the nodes by "levels", l. Given a region ("blob"), b, and a word w, we have

$$P(w \mid b) = \sum_{l} P(w \mid l) P(b \mid l) P(l) / P(b) \qquad (1)$$

where P(l) is the level prior, P(w|l) is a frequency table, and P(b|l) is a Gaussian over features. To estimate the conditional density of words given blobs for the entire image these probabilities are summed over the N largest blobs. Parameters for the conditional probabilities linking words and blobs are estimated from the word-blob co-occurrence data using Expectation Maximization (Dempster et al., 1977). For all experiments reported in this paper we use 100 nodes.

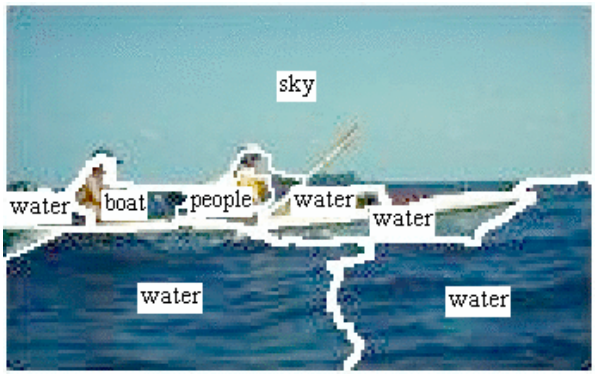



Figure 2. Illustration of labeling. Each region is labeled with the maximally probable word, but a probability distribution over all words is available for each region.

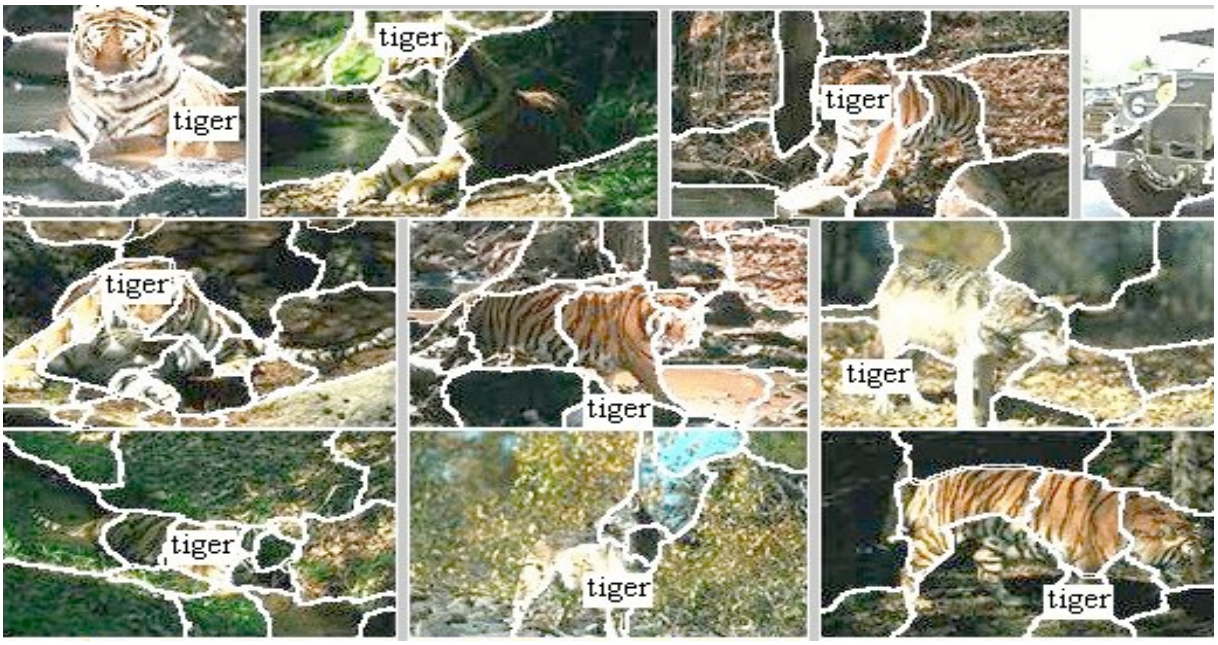
## 3 Constrained Word Prediction

In most of our applications we have studied word prediction from images in the case where prediction is applied to a completely new image with no associated words. However, if the new image has associated words, our infrastructure for image/language understanding can be exploited further. Here the main task is not word prediction, but understanding the relationship between the supplied words and the image components, and hence the meaning of both. Since the set of words of interest is known, the rest of the vocabulary can be ignored in computation. Constraining the vocabulary in this way makes a number of tasks simpler. As much noise has been removed, the relationship between the words and the image components can be established more accurately. The system is now chiefly determining the correspondence relationships between known text and image regions.

One application is automatic labeling of images for searching and browsing based on the semantics of the image parts. As shown in Figure 3, the labeling performance is much improved when we can constrain the vocabulary to largely relevant words. The presence of the words has removed ambiguity from the interpretation of the image.

The second application is, of course, the reverse—using the image to help reduce the ambiguity of the words. We assume that the system has been trained on a set of senses, S, for the vocabulary W. To clarify the notation, we may have $bank \in W$ and $bank\_1, bank\_2 \in S$. Each element of S is the sense of exactly one word in W. If we have posterior probabilities over S based on the image, then for each observed word, w, we can look at the corresponding senses for w in S, and provide the sense which has the highest posterior among the senses. More formally,


(a)




(b)

Figure 3. Illustration of the observation that our ability to predict word-region correspondences increases significantly when the words are constrained to a small set known to be relevant. We show two groups of images which have a high probability of having a region associated with the word tiger, as computed by two different processes. The region in the images with the highest posterior probability of "tiger" is labeled as such. In both cases the images shown were not used for training. In the top group (a) only image region features were used to predict words. In the bottom group (b), words associated with the images were also available to the program, and thus their the main task is to supply the correspondence between words and regions.

$$P(s \mid w,I) \quad P(s \mid I)P(s \mid w) \tag{2}$$

where

$$P(s \mid w) \quad \begin{cases} 1 & \textit{if } s \textit{ is a sense of } w \\ 0 & \textit{otherwise} \end{cases} \tag{3}$$

## 4 Word Sense from Images and Text

To integrate the above with traditional word sense disambiguation, we assume that we have a text only method which provides a better value for $P(s \mid w)$ in (2) than the trivial one defined in (3). This strategy assumes that the image provides information which is independent of the text only method, allowing the simplification from the factorization used in (2).

For this work, we do not compute a true probability for $P(s \mid w)$, but rather a score which we use as a surrogate. As it is the goal of this work to show that improvement in word sense disambiguation is possible, we use a simple word sense disambiguation method based on the work in (Barnard et al., 2001) which itself takes ideas from (Agirre and Rigau, 1995; Mihalcea and Moldovan., 1998). Specifically we use the WordNet semantic hierarchy (Miller et al., 1990) to define senses, and give higher weight to senses more closely aligned (lower in the tree) with the neighboring words. In the experiments described below, the words are part of keyword sets, and all keywords for an image are considered neighbours.

For example, if we had the keywords [lion, pride, rock] for a picture of a group of lions on a rock, the word pride presents a problem to the disambiguator. By far, pride's most common meaning is that regarding it as a deadly sin, however in this case we wish it to be disambiguated as a group of lions. Even so, one can look at the different WordNet sense hierarchies for pride and find that one, namely:

pride
=> animal group
=> biological group
=> group, grouping

contains the words animal and biological, making it a better fit for the hierarchy of lion.

With this structure in mind, our algorithm takes the set of keywords and, for each keyword in the set, performs a query such as the one shown above. Then, for each sense of the keyword, we perform the queries for the other keywords, and for each of their senses we examine the similarities between their hypernym trees. We total up these similarities (shared nodes in the tree) and for each sense of the keyword produce a subtotal for that sense. After we have performed this operation for all senses we divide the subtotal by the complete total for all senses to receive a score for that sense as the true definition of the keyword.

## 5 Experiments

For our initial test, we studied word sense disambiguation on the Corel image dataset which we have used extensively for studying word prediction from images. Each Corel image has 3-5 keywords associated with it. Unfortunately, these keywords are unusual in that they do not have much sense conflict over the data set. Put differently, although a keyword like "head" has many senses, one sense predominates in this data set.

To use the data despite this problem, we computed all the senses from WordNet (Miller et al., 1990) of all the words for an initial set of 16,000 images, together with the score for each sense using the method described above. We then applied some heuristics to create a subset of 1,800 images which had candidate sense problems. Each word was then hand labeled with the correct sense. The resulting dataset had only a handful of words with ambiguous senses present in sufficient quantity, but fortunately these were common enough such that about 1/6 of the documents had true word sense problems. The results reported below were restricted to documents that had at least one word sense ambiguity.

The data set prepared as above thus consists of a vocabulary of word-sense combinations, together with a human labeling of whether the sense was valid or not. We restrict the vocabulary to word-sense pairs which occur in at least 5 images. We represent the observed senses for a word occurring in a document as a vector over the senses for that word from the vocabulary. We give all relevant senses of the word a score of one, and incorrect senses a score of zero. We normalize this vector so that its sum is one. Although potentially a word could be ambiguous to a human examiner, typically the word sense vector would simply contain a single value of one, with the other values being zero.

We evaluate word-sense disambiguation strategies by comparing the vector of observed word-senses described above (the truth) to the vector containing estimates of the relevance of each word-sense pair corresponding to each occurring word. For example, suppose an image has the word "bank", which maps to bank_1 with hand labeling, and suppose that bank_1 and bank_3 are in the vocabulary, but no other senses of bank. Then the vector

{…,bank_1(0.7),bank_3(0.3),…}

should be ranked better than

{…,bank_1(0.3),bank_3(0.7),…}

when compared to the observed vector

{…,bank_1(1.0),bank_3(0.0),…} (hand-labeled)

To compare the vectors we simply normalize them and take the dot-product.

For the experiments we divided the data into training data (75%), and test data (25%). We averaged results

for 10 separate runs using different samples for the test and training sets. We restricted the computation of results to those documents where there was clear sense ambiguity. Because each such document typically had only one sense problem amid 3 or 4 words without sense problems, the baseline score using the measure above is greater than 0.80 because any strategy will get about 3 out or 4 correct for free. To clarify this further, we include the results of randomly chosen among senses when there is more than one available.

The results are shown in Table 1 strongly suggest that images can help disambiguate senses. The naïve method of text based disambiguation is comparable to chance, whereas adding image information substantially increased the performance.

## 6 Conclusion

These preliminary studies strongly suggest that it is worthwhile to explore combining image information with more sophisticated text based words sense disambiguation approaches. However, while the preliminary results are encouraging, it is critical that we take the next step and apply the method to a data set where there is more sense ambiguity. Possible candidates which we are actively investigating include the museum data used in (Barnard et al., 2001) and news photos with captions available on the web.

In general we have found that it is fruitful to study how image and text information can both compliment each other and disambiguate one another. Different representations of the same thing can help learn co-constructed meaning. Properties which may be implicit in one representation may be more explicit and thus more amenable for automatic extraction in another. Furthermore, relationships between the representations, which can be learnt from large corpora, can be brought to bear on the problem. In particular, in the case of disambiguating words, we have shown that images can provide a non-negligible amount of information which can be exploited by more traditional approaches.

| Word sense disambiguation strategy | Score |
|---|---|
| Naïve text based method | 0.858 (0.008) |
| Random sense choice | 0.875 (0.012) |
| Image and text method | 0.948 (0.015) |

Table 1. Word sense disambiguation results on data held out from training. The results are the average over 10 runs where a different 75% of the data was used for training and the other 25% was held out for testing. Results were computed only on images with at least one ambiguous term. Because typically only one out of four or five words was ambiguous, the baseline score is quite high as reinforced by the random result. The scoring is explained in §5. Error estimates are in parentheses.

## References


Agirre, E. and Rigau, G., 1995. A proposal for word sense disambiguation using conceptual distance, 1st International Conference on Recent Advances in Natural Language Processing, Velingrad.

Barnard, K., Duygulu, P. and Forsyth, D., 2001. Clustering Art, IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, pp. II:434-441.

Barnard, K. et al., 2003. Matching Words and Pictures. Journal of Machine Learning Research, 3: 1107-1135.

Barnard, K. and Forsyth, D., 2001. Learning the Semantics of Words and Pictures, International Conference on Computer Vision, pp. II:408-415.

Dempster, A.P., Laird, N.M. and Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 39(1): 1-38.

Duygulu, P., Barnard, K., de Freitas, J.F.G. and Forsyth, D.A., 2002. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary, The Seventh European Conference on Computer Vision, Copenhagen, Denmark, pp. IV:97-112.

Gale, W., Church, K. and Yarowsky, D., 1992. One Sense Per Discourse, DARPA Workshop on Speech and Natural Language, New York, pp. 233-237.

Manning, C. and Schütze, H., 1999. Foundations of Statistical Natural Language Processing. MIT Press. Cambridge, MA.

Mihalcea, R. and Moldovan., D., 1998. Word sense disambiguation based on semantic density, COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems, Montreal.

Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K.J., 1990. Introduction to WordNet: an on-line lexical database. International Journal of Lexicography, 3(4): 235 - 244.

Traupman, J. and Wilensky, R., 2003. Experiments in Improving Unsupervised Word Sense Disambiguation. CSD-03-1227, Computer Science Division, University of California Berkeley.

Yarowsky, D., 1995. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods, 33rd Conference on Applied Natural Language Processing. ACL, Cambridge.