# Unsupervised Disambiguation of Image Captions

**Wesley May, Sanja Fidler, Afsaneh Fazly, Sven Dickinson, and Suzanne Stevenson**
Department of Computer Science
University of Toronto
Toronto, Ontario, Canada, M5S 3G4
{wesley,fidler,afsaneh,sven,suzanne}@cs.toronto.edu

## Abstract

Given a set of images with related captions, our goal is to show how visual features can improve the accuracy of unsupervised word sense disambiguation when the textual context is very small, as this sort of data is common in news and social media. We extend previous work in unsupervised text-only disambiguation with methods that integrate text and images. We construct a corpus by using Amazon Mechanical Turk to caption sense-tagged images gathered from ImageNet. Using a Yarowsky-inspired algorithm, we show that gains can be made over text-only disambiguation, as well as multimodal approaches such as Latent Dirichlet Allocation.

## 1 Introduction

We examine the problem of performing unsupervised word sense disambiguation (WSD) in situations with little text, but where additional information is available in the form of an image. Such situations include captioned newswire photos, and pictures in social media where the textual context is often no larger than a tweet.

Unsupervised WSD has been shown to work very well when the target word is embedded in a large quantity of text (Yarowsky, 1995). However, if the only available text is "The crane was so massive it blocked the sun" (see Fig. 1), then text-only disambiguation becomes much more difficult; a human could do little more than guess. But if an image is available, the intended sense is much clearer. We develop an unsupervised WSD algorithm based on Yarowsky's that uses words in a short caption along with "visual words" from the captioned image to choose the best of two possible senses of an ambiguous keyword describing the content of the image.


Figure 1: "The crane was so massive it blocked the sun." Which sense of crane? With images the answer is clear.

Language-vision integration is a quickly developing field, and a number of researchers have explored the possibility of combining text and visual features in various multimodal tasks. Leong and Mihalcea (2011) explored semantic relatedness between words and images to better exploit multimodal content. Jamieson et al. (2009) and Feng and Lapata (2010) combined text and vision to perform effective image annotation. Barnard and colleagues (2003; 2005) showed that supervised WSD by could be improved with visual features. Here we show that unsupervised WSD can similarly be improved. Loeff, Alm and Forsyth (2006) and Saenko and Darrell (2008) combined visual and textual information to solve a related task, image sense disambiguation, in

We thank NSERC and U. Toronto for financial support. Fidler and Dickinson were sponsored by the Army Research Laboratory and this research was accomplished in part under Cooperative Agreement Number W911NF-10-2-0060. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either express or implied, of the Army Research Laboratory or the U.S. Government.

an unsupervised fashion. In Loeff et al.'s work, little gain was realized when visual features were added to a great deal of text. We show that these features have more utility with small textual contexts, and that, when little text is available, our method is more suitable than Saenko and Darrell's.

## 2 Our Algorithm

We model our algorithm after Yarowsky's (1995) algorithm for unsupervised WSD: Given a set of documents that contain a certain ambiguous word, the goal is to label each instance of that word as some particular sense. A seed set of collocations that strongly indicate one of the senses is initially used to label a subset of the data. Yarowsky then finds new collocations in the labelled data that are strongly associated with one of the current labels and applies these to unlabelled data. This process repeats iteratively, building a decision list of collocations that indicate a particular sense with a certain confidence.

In our algorithm (Algorithm 1), we have a document collection $D$ of images relevant to an ambiguous keyword $k$ with senses $s_1$ and $s_2$ (though the algorithm is extensible to more than two senses). Such a collection might result from an internet image search using an ambiguous word such as "mouse".

Each $D_i$ is an image–caption pair repsented as a bag-of-words that includes both lexical words from the caption, and "visual words" from the image. A visual word is simply an abstract representation that describes a small portion of an image, such that similar portions in other images are represented by the same visual word (see Section 3.2 for details). Our seed sets consist of the words in the definitions of $s_1$ and $s_2$ from WordNet (Fellbaum, 1998). Any document whose caption contains more words from one sense definition than the other is initially labelled with that sense. We then iterate between two steps that (i) find additional words associated with $s_1$ or $s_2$ in currently labelled data, and (ii) relabel all data using the word sense associations discovered so far.

We let $V$ be the entire vocabulary of words across all documents. We run experiements both with and without visual words, but when we use visual words, they are included in $V$. In the first step, we compute a confidence $C_i$ for each word $V_i$. This confidence is a log-ratio of the probability of seeing $V_i$ in documents labelled as $s_1$ as opposed to documents labelled as $s_2$. That is, a positive $C_i$ indicates greater association with $s_1$, and vice versa. In the second step we find, for each document $D_j$, the word $V_i \in D_j$ with the highest magnitude of $C_i$. If the magnitude of $C_i$ is above a labelling threshold $\tau_c$, then we label this document as $s_1$ or $s_2$ depending on the sign of $C_i$. Note that all old labels are discarded before this step, so labelled documents may become unlabelled, or even differently labelled, as the algorithm progresses.

**Algorithm 1** Proposed Algorithm

$D$: set of documents $D_1$ ... $D_d$
$V$: set of lexical and visual words $V_1$ ... $V_v$ in $D$
$C_i$: log-confidence $V_i$ is sense 1 vs. sense 2
$S_1$ and $S_2$: bag of dictionary words for each sense
$L_1$ and $L_2$: documents labelled as sense 1 or 2

**for all** $D_i$ **do** ▷ Initial labelling using seed set
  **if** $|D_i \cap S_1| > |D_i \cap S_2|$ **then**
    $L_1 \leftarrow L_1 \cup \{D_i\}$
  **else if** $|D_i \cap S_1| < |D_i \cap S_2|$ **then**
    $L_2 \leftarrow L_2 \cup \{D_i\}$
  **end if**
**end for**

**repeat**
  **for all** $i \in 1..v$ **do** ▷ Update word conf.
    $C_i \leftarrow log\left(\frac{P(V_i|L_1)}{P(V_i|L_2)}\right)$
  **end for**

  $L_1 \leftarrow \emptyset$, $L_2 \leftarrow \emptyset$ ▷ Update document conf.
  **for all** $D_i$ **do**
    ▷ Find word with highest confidence
    $m \leftarrow \underset{j \in 1..v, V_j \in D_i}{\arg\max} |C_j|$
    **if** $C_m > \tau_c$ **then**
      $L_1 \leftarrow L_1 \cup \{D_i\}$
    **else if** $C_m < -\tau_c$ **then**
      $L_2 \leftarrow L_2 \cup \{D_i\}$
    **end if**
  **end for**
**until** no change to $L_1$ or $L_2$

## 3 Creation of the Dataset

We require a collection of images with associated captions. We also require sense annotations for the keyword for each image to use for evaluation. Barnard and Johnson (2005) developed the

"Music is an important means of expression for many teens."

"Keeping your office supplies organized is easy, with the right tools."

"The internet has opened up the world to people of all nationalities."

"When there is no cheese I will take over the world."

Figure 2: Example image-caption pairs from our dataset, for "band" (top) and "mouse" (bottom).

ImCor dataset by associating images from the Corel database with text from the SemCor corpus (Miller et al., 1993). Loeff et al. (2006) and Saenko and Darrell (2008) used Yahoo!'s image search to gather images with their associated web pages. While these datasets contain images paired with text, the textual contexts are much larger than typical captions.

### 3.1 Captioning Images

To develop a large set of sense-annotated image–caption pairs with a focus on caption-sized text, we turned to ImageNet (Deng et al., 2009). ImageNet is a database of images that are each associated with a synset from WordNet. Hundreds of images are available for each of a number of senses of a wide variety of common nouns. To gather captions, we used Amazon Mechanical Turk to collect five sentences for each image. We chose two word senses for each of 20 polysemous nouns and for each sense we collected captions for 50 representative images. For each image we gathered five captions, for a total of 10,000 captions. As we have five captions for each image, we split our data into five sets. Each set has the same images, but each image is paired with a different caption in each set.

We specified to the Turkers that the sentences should be relevant to, but should not talk directly about, the image, as in "In this picture there is a blue fish", as such captions are very unnatural. True captions generally offer orthogonal information that is not readily apparent from the image. The keyword for each image (as specified by ImageNet) was not presented to the Turkers, so the captions do not necessarily contain it. Knowledge of the keyword is presumed to be available to the algorithm in the form of an image tag, or filename, or the like. We found that forcing a certain word to be included in the caption also led to sentences that described the picture very directly. Sentences were required to be a least ten words long, and have acceptable grammar and spelling. We remove stop words from the captions and lemmatize the remaining words. See Figure 2 for some examples.

### 3.2 Computing the Visual Words

We compute visual words for each image with ImageNet's feature extractor. This extractor lays down a grid of overlapping squares onto the image and computes a SIFT descriptor (Lowe, 2004) for each square. Each descriptor is a vector that encodes the edge orientation information in a given square. The descriptors are computed at three scales: 1x, 0.5x and 0.25x the original side lengths. These vectors are clustered with k-means into 1000 clusters, and the labels of these clusters (arbitrary integers from 1 to 1000) serve as our visual words.

It is common for each image to have a "vocabulary" of over 300 distinct visual words, many of which only occur once. To denoise the visual data, we use only those visual words which account for at least 1% of the total visual words for that image.

## 4 Experiments and Results

To show that the addition of visual features improves the accuracy of sense disambiguation for image–caption pairs, we run our algorithm both with and without the visual features. We also compare our results to three different baseline methods: K-means (K-M), Latent Dirichlet Allocation (LDA) (Blei et al., 2003), and an unsupervised WSD algorithm (PBP) explained below. We use accuracy to measure performance as it is commonly used by the WSD community (See Table 1).

For K-means, we set $k = 2$ as we have two senses, and represent each document with a $V$-dimensional

Table 1: Results (Average accuracy across all five sets of data). Bold indicates best performance for that word.

| | Ours text | Ours w/vis | K-M text | K-M w/vis | LDA text | LDA w/vis | PBP text |
|---|---|---|---|---|---|---|---|
| band | .80 | **.82** | .66 | .65 | .64 | .56 | .73 |
| bank | .77 | **.78** | .71 | .59 | .52 | .67 | .62 |
| bass | **.94** | **.94** | .90 | .88 | .61 | .62 | .49 |
| chip | **.90** | **.90** | .73 | .58 | .57 | .66 | .75 |
| clip | .70 | **.79** | .65 | .58 | .48 | .53 | .65 |
| club | .80 | **.84** | .80 | .81 | .61 | .73 | .63 |
| court | .79 | .79 | .61 | .53 | .62 | **.82** | .57 |
| crane | .62 | .67 | **.76** | **.76** | .52 | .54 | .66 |
| game | **.78** | **.78** | .60 | .66 | .60 | .66 | .70 |
| hood | **.74** | .73 | .73 | .70 | .51 | .45 | .55 |
| jack | **.76** | .74 | .62 | .53 | .58 | .66 | .47 |
| key | .81 | **.92** | .79 | .54 | .57 | .70 | .50 |
| mold | .67 | **.68** | .59 | .67 | .57 | .66 | .54 |
| mouse | **.84** | **.84** | .71 | .62 | .62 | .69 | .68 |
| plant | .54 | .54 | .56 | .53 | .52 | .50 | **.72** |
| press | .60 | .59 | .60 | .54 | .58 | **.62** | .48 |
| seal | .70 | **.80** | .61 | .67 | .55 | .53 | .62 |
| speaker | **.70** | .69 | .57 | .53 | .55 | .62 | .63 |
| squash | .89 | **.95** | .84 | .92 | .55 | .67 | .79 |
| track | .78 | **.85** | .71 | .66 | .51 | .54 | .69 |
| **avg.** | .76 | **.78** | .69 | .65 | .56 | .63 | .62 |

vector, where the $i$th element is the proportion of word $V_i$ in the document. We run K-means both with and without visual features.

For LDA, we use the dictionary sense model from Saenko and Darrell (2008). A topic model is learned where the relatedness of a topic to a sense is based on the probabilities of that topic generating the seed words from its dictionary definitions. Analogously to k-means, we learn a model for text alone, and a model for text augmented with visual information.

For unsupervised WSD (applied to text only), we use WordNet::SenseRelate::TargetWord, hereafter PBP (Patwardhan et al., 2007), the highest scoring unsupervised lexical sample word sense disambiguation algorithm at SemEval07 (Pradhan et al., 2007). PBP treats the nearby words around the target word as a bag, and uses the WordNet hierarchy to assign a similarity score between the possible senses of words in the context, and possible senses of the target word. As our captions are fairly short, we use the entire caption as context.

The most important result is the gain in accuracy after adding visual features. While the average gain across all words is slight, it is significant at $p < 0.02$ (using a paired t-test). For 12 of the 20 words, the visual features improve performance, and in 6 of those, the improvement is 5–11%.

For some words there is no significant improvement in accuracy, or even a slight decrease. With words like "bass" or "chip" there is little room to improve upon the text-only result. For words like "plant" or "press" it seems the text-only result is not strong enough to help bootstrap the visual features in any useful way. In other cases where little improvement is seen, the problem may lie with high intra-class variation, as our visual words are not very robust features, or with a lack of orthogonality between the lexical and visual information.

Our algorithm also performs significantly better than the baseline measurements. K-means performs surprisingly well compared to the other baselines, but seems unable to make much sense of the visual information present. Saenko and Darrell's (2008) LDA model makes substansial gains by using visual features, but does not perform as well on this task. We suspect that a strict adherence to the seed words may be to blame: while both this LDA model and our algorithm use the same seed definitions initially, our algorithm is free to change its mind about the usefulness of the words in the definitions as it progresses, whereas the LDA model has no such capacity. Indeed, words that are intuitively non-discriminative, such as "carry", "lack", or "late", are not uncommon in the definitions we use.

## 5 Conclusion and Future Work

We present an approach to unsupervised WSD that works jointly with the visual and textual domains. We showed that this multimodal approach makes gains over text-only disambiguation, and outperforms previous approaches for WSD (both text-only, and multimodal), when textual contexts are limited.

This project is still in progress, and there are many avenues for further study. We do not currently exploit collocations between lexical and visual information. Also, the bag-of-SIFT visual features that we use, while effective, have little semantic content. More structured representations over segmented image regions offer greater potential for encoding semantic content (Duygulu et al., 2002).

## References


Kobus Barnard and Matthew Johnson. 2005. Word sense disambiguation with pictures. In *Artificial Intelligence*, volume 167, pages 13–130.

Kobus Barnard, Matthew Johnson, and David Forsyth. 2003. Word sense disambiguation with pictures. In *Workshop on Learning Word Meaning from Non-Linguistic Data*, Edmonton, Canada.

David M. Blei, Andrew Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. In *JMLR*, volume 3, pages 993–1022.

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

Pinar Duygulu, Kobus Barnard, Nando de Freitas, and David Forsyth. 2002. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *European Conference on Computer Vision*, Copenhagen, Denmark.

Christiane Fellbaum. 1998. Wordnet: An electronic lexical database. In *Bradford Books*.

Yansong Feng and Mirella Lapata. 2010. Topic models for image annotation and text illustration. In *Annual Conference of the North American Chapter of the ACL*, pages 831–839, Los Angeles, California.

Michael Jamieson, Afsaneh Fazly, Suzanne Stevenson, Sven Dickinson, and Sven Wachsmuth. 2009. Using language to learn structured appearance models for image annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1):148–164.

Chee Wee Leong and Rada Mihalcea. 2011. Measuring the semantic relatedness between words and images. In *International Conference on Semantic Computing*, Oxford, UK.

Nicolas Loeff, Cecilia Ovesdotter Alm, and David Forsyth. 2006. Discriminating image senses by clustering with multimodal features. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 547–554, Sydney, Australia.

David Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110.

George Miller, Claudia Leacock, Randee Tengi, and Ross Bunker. 1993. A semantic concordance. In *Proceedings of the 3rd DARPA Workshop on Human Language Technology*, pages 303–308.

Siddharth Patwardhan, Satanjeev Banerjee, and Ted Pedersen. 2007. UMND1: Unsupervised word sense disambiguation using contextual semantic relatedness. In *Proceedings of SemEval-2007*, pages 390–393, Prague, Czech Republic.

Sameer Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. Task 17: English lexical sample, SRL and all words. In *Proceedings of SemEval-2007*, pages 87–92, Prague, Czech Republic.

Kate Saenko and Trevor Darrell. 2008. Unsupervised learning of visual sense models for polysemous words. In *Proceedings of Neural Information Processing Systems*, Vancouver, Canada.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the ACL*, pages 189–196, Cambridge, Massachusetts.