A Bayesian Model for Image Sense Ambiguity in Pictorial Communication Systems

Jake Rosin
Andrew Goldberg
Xiaojin Zhu
Charles Dyer

Technical Report #1692

June 2011

UNIVERSITY OF WISCONSIN MADISON

A Bayesian Model for Image Sense Ambiguity in Pictorial Communication Systems

Jake Rosin and **Andrew B. Goldberg** and **Xiaojin Zhu** and **Charles Dyer**
Computer Sciences Department
University of Wisconsin-Madison
Madison, WI 53706, USA
{rosin,goldberg,jerryzhu,dyer}@cs.wisc.edu

Abstract

Pictorial communication systems use synthesized pictures, rather than text, to communicate with users. Because such systems depend on images to convey meanings, it is critical to understand how a human user perceives the image meaning (sense). This paper offers an empirical and theoretical study of how humans perceive image senses. We conduct a user study with 113 users to elicit their perceived senses on 400 image sets, from which we discover widespread image sense ambiguities. We examine how the number of images shown relates to sense ambiguity and discover several significant patterns. We then propose a Bayesian model to explain human image perception behaviors, based on a novel random walk process on a WordNet-like sense hierarchy. Our model makes qualitative and quantitative predictions that largely agree with our observations of human perception. It can explain the "basic level" phenomenon known in psychology, and suggests a method for image sense disambiguation in pictorial communication systems.

Introduction

Pictorial communication systems aim to convey the meaning of a piece of natural language text (e.g., "The poodle runs out the door") using automatically generated pictures (Coyne and Sproat 2001; Johansson et al. 2005; Joshi, Wang, and Li 2006; Zhu et al. 2007). These systems enable novel means of communication and human-computer interaction, especially for communicative disorder patients, young students learning to read, or foreign language speakers. The success of pictorial communication systems depends critically on the comprehensibility of the generated pictures. Among various factors affecting comprehensibility, a fundamental issue is *sense ambiguity* of individual images within a picture. For concreteness, we equate senses to synsets in WordNet (Fellbaum 1998). For the example above, the picture may contain an image of a poodle. From the user perspective, however, the image may mean "poodle," or its hypernyms "dog" or "animal," because any of these senses can potentially be represented by the poodle image. Sense ambiguity is universal to pictorial communication systems that use natural images (as opposed to artificial symbols), and is unavoidable for most single images (i.e., no matter how faithful the poodle image is, such ambiguity remains).

This paper studies a basic research question: when a user is presented with one or more images and told that they represent a single sense, what sense does she perceive? The question is significant as it quantifies how precisely a concept can be conveyed via pictorial communication systems. As we show later, answers to this question can be used to disambiguate the sense of an image with additional images from the target sense.[1] For the poodle example, the system may show several other poodle images to indicate that the target sense is this particular type of dog, rather than a generic dog or animal. To our knowledge, no prior work has quantitatively addressed this question. We quantify the extent of image sense ambiguity, and the effectiveness of disambiguation, using a novel Bayesian model built upon recent work in psychology on Bayesian word learning (Xu and Tenenbaum 2007). We first describe a user study in which we collect empirical data. With the data, we then develop the Bayesian model to predict the perceived sense. Finally, we explore the degree to which the model explains the data and known aspects of human cognition.

Formally, we consider a tree $G = \{V, E\}$ where the vertices $V = \{y_1, \ldots, y_m\}$ are senses (synsets), and the directed tree edges $E = \{e_{ij}\}$ encode the hypernym (a.k.a. is-a or general-specific) relationship (Fellbaum 1998): an edge e_{ij} goes from parent y_i (e.g., "dog") to child y_j (e.g., "poodle") if y_i is the hypernym of y_j. When referring to vertices in V from here on, we use the terms "sense" and "concept" interchangeably. In addition, a special $\langle other \rangle$ sense captures all other senses not in V. We consider images $X = \{x\}$ that can each be unambiguously assigned to a unique leaf $y_x \in G$. That is, given $x \in X$ (e.g., a poodle image) and restricted to G's leaves (e.g., "poodle," "Dalmatian," "golden retriever," etc.), a user should be able to assign x to the correct leaf. In other words, these are clear, good quality images. Furthermore, we will focus on

[1] We call this strategy disambiguation-by-samples. An alternative, disambiguation-by-context, shows images commonly *associated* with the target sense, such as pink bows for poodles, or firehouses for Dalmatians. In both strategies, the user interface can indicate that these additional images are for disambiguation, and not part of the picture, e.g., by showing them in a pop-up window when the user mouses over the original poodle image. We will focus on disambiguation-by-samples, but note that the two strategies can be synergistic.

post-visual cognitive processing, by assuming that the vision task of mapping any image $x \in X$ to its unique leaf sense y_x has been accomplished (perhaps imperfectly). However, there will be ambiguity when the user is not restricted to G's leaves. For example, the image for an internal sense (e.g., "dog") can come from the many leaves under that sense (e.g., poodle, Dalmatian, or golden retriever). Therefore, a poodle image could have also come from the internal sense "dog".

With these definitions, we formulate our main question as follows: given $n \geq 1$ images $\mathbf{x}_{1\ldots n}$, what sense $y \in Y \equiv V \cup \{\langle other \rangle\}$ will a user perceive? In particular, we are interested in modeling $P(y \mid \mathbf{x}_{1\ldots n})$, which gives a distribution over possible senses given the images, providing a measure of sense ambiguity.

User Study

[Materials] In our study, we use a set of 250 common food items as V. The tree edges E (hypernym relationships) largely follow WordNet, with some modifications to fit common-sense organization of these items. The resulting tree G has 201 leaves, with a maximum depth of 6 and an average leaf depth of 4.1. For modeling purposes, we will assume that G is the tree by which the human mind organizes these items. We discuss ways to relax this assumption later.

We selected 100 test senses in G (74 leaf sense, 26 internal senses). We then manually collected 400 high-quality images, four for each test sense, with each image corresponding to the test sense (if it is a leaf) or a leaf under the test sense (if it is internal) in G. This produced 100 quadruplets $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(100)}$ of images; one such quadruplet, representing the leaf sense "Swiss cheese," is shown in Figure 1. For each leaf test sense, we construct four image sets using the first one, two, three, and all images in its quadruplet. For each internal test sense, we construct three image sets, using the first two, three and all images in its quadruplet. In the latter case, we make sure that the internal test sense is the least common ancestor of the first two images. We use $\mathbf{x}^{(i)}_{1\ldots j}$ to denote the image set consisting of $j \leq 4$ images from quadruplet i. This design permits us to probe how the human-perceived sense changes with progressively more image evidence. In total, we have 374 image sets ($i = 1 \ldots 100$, $j = 1 \ldots 4$ if $i \leq 74$, $j = 2 \ldots 4$ otherwise).

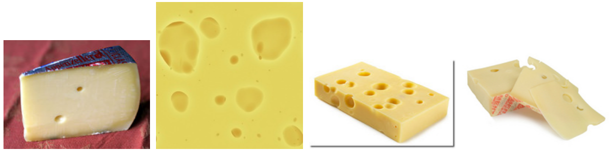

Figure 1: Four images of the leaf sense "Swiss cheese."

[Subjects and Procedure] Participants were 113 university students, participating for partial course credit. The experiment was conducted using a Web-based interface. Subjects were instructed that the goal of the study was to identify *a single food concept* found in a supermarket based on *each* image set. Subjects were also told that the image sets, as well as the corresponding concepts, were independent of each other. They were given no information about how the images were chosen, nor shown the tree G or the list of senses V. Each subject was shown 100 image sets, one set at a time, in random order. Subjects were asked to type a free-text response indicating the concept they believed was being represented by the current image set. After submitting the response, the subject was shown a new set of images on the next screen. Each subject saw exactly one set of images for every $i = 1 ... 100$. We balanced the subjects so each subset $\mathbf{x}^{(i)}_{1\ldots j}$ was seen by about 30 people. Because the subjects were not shown G or V, their free-text responses may not exactly match any synsets in V. Without considering the images shown to the subjects, a team of 3 annotators manually inspected all responses and assigned each to the closest sense in V when possible, or $\langle other \rangle$ otherwise.

[Observations] Given G, a set of images drawn from the leaves uniquely determines one *lowest common ancestor* (LCA) within the hierarchy: the most specific sense that is consistent with all examples. Under assumption that the sense hierarchy is shared by all users (that the hierarchy represents the knowledge of an "average human"), a natural *LCA hypothesis* predicts that the human response induced by a set of images will be the LCA of those images' associated leaves. As a natural consequence, one image is sufficient to induce any leaf-level test sense, and two images are sufficient to induce any internal test sense.

Our experiment demonstrates the faults in the LCA hypothesis, and the need for more advanced modeling. We make three observations:

(1) We note that sense ambiguity is common among responses and widespread across senses. When subjects are shown a single image, they name the corresponding leaf concept only 40.1% of the time. Defining an image (or image set) as ambiguous if, among all the subjects who saw it, no more than 90% (80% resp.) had the same sense response, 55 (49 resp.) of these 74 sets were ambiguous. For example, the same single "white wine" image led to 7 subjects responding "white wine," 18 subjects responding "wine," 2 subjects responding "alcoholic drink," and 1 subject responding "champagne."

Given two images drawn from different leaves, only 54.9% of responses name their LCA; by the definition of ambiguity given above, 21 (16 resp.) of these 26 sets were ambiguous. Responses other than the LCA of the image set can be categorized into two groups: *generalized* and *inconsistent* responses. Generalized responses are consistent with the images shown, but are more general than (i.e., ancestors to) the LCA of those images. Inconsistent responses name senses that are not shared by the images. As an example, consider the image set in Figure 2. When shown this set, 67.9% of subjects named the LCA sense "avocado," while 10.7% generalized to "fruit." Responses inconsistent with the images, such as "apricots," "melons," "papaya," and "squash" made up the remaining 21.5%.

(2) By the LCA hypothesis, including additional images in the set displayed will not change human responses un-

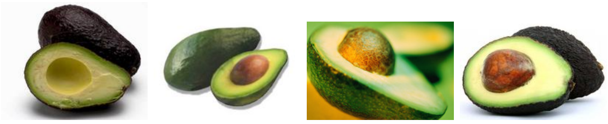

Figure 2: Four images of the leaf sense “avocado.”

less the LCA of those images is also changed. However, the data shows that including one image beyond those needed to define a given LCA (a second image for a leaf-level LCA, a third for an internal LCA) increases the proportion of responses naming the LCA. This reveals that apparently redundant information is helpful in reducing sense ambiguity.

(3) The usefulness of redundant examples appears limited, however. Including even more redundant images does not significantly change user responses. These results are shown in Table 1.

Table 1: The percentage of LCA, generalized, and inconsistent responses, for image sets of different sizes.

Images representing a leaf sense			
Images shown	LCA	Generalized	Inconsistent
$\mathbf{x}_1$	40.1	36.7	23.2
$\mathbf{x}_{1\ldots2}$	44.7	35.7	19.6
$\mathbf{x}_{1\ldots3}$	44.4	35.4	20.2
$\mathbf{x}_{1\ldots4}$	44.5	35.7	19.8

Images representing an internal sense			
Images shown	LCA	Generalized	Inconsistent
$\mathbf{x}_{1\ldots2}$	54.9	14.7	30.4
$\mathbf{x}_{1\ldots3}$	63.6	12.0	24.4
$\mathbf{x}_{1\ldots4}$	65.4	13.5	21.2

Bayesian Modeling

We now propose a novel Bayesian model to explain image sense ambiguity. Recent research has shown that Bayesian inference is preferable in modeling human word learning (Xu and Tenenbaum 2007). In this paper, we show that Bayesian inference is also a good model for image sense perception in humans. We propose to compute $P(y \mid \mathbf{x}_{1\ldots n})$ via the Bayes rule:

$$P(y \mid \mathbf{x}_{1\ldots n}) = \frac{P(y)P(\mathbf{x}_{1\ldots n} \mid y)}{\sum_{y' \in V} P(y')P(\mathbf{x}_{1\ldots n} \mid y')}. \qquad (1)$$

One major contribution of this paper is our formulation of the prior $P(y)$ and the likelihood term $P(\mathbf{x}_{1\ldots n} \mid y)$.

We estimate the prior $P(y)$ from frequency in the largest corpus available, namely the Web. For each synset y we use the singular and plural forms of its most common name, together with the word “food,” to form two search queries. We search on both singular and plural forms because some food items (e.g., “beans”) are rarely considered in individual units. Let $c_{\text{text}}(y)$ be the maximum of the number of search results found for these two queries. Our estimate is then $P(y) \propto c_{\text{text}}(y)$.

We estimate the likelihood with a random walk process. This walk can be informally understood in terms of selecting a concrete image consistent with a sense $y \in V$: any image of y is by definition generated from exactly one hyponym of y. This selection iterates until one reaches a leaf; an image of that leaf is then generated.

Formally, we first assume that the images are generated i.i.d.: $P(\mathbf{x}_{1\ldots n} \mid y) = \prod_{i=1}^{n} P(x_i \mid y)$. Then, $P(x \mid y)$ is modeled as a teleporting random walk on G, beginning at y. For simplicity, consider first the case without teleporting. We prepare G for the walk by adding absorbing nodes $a \in A$ as children to the leaf senses, one per leaf. The random walk terminates only when it reaches one of these absorbing nodes, which are included purely for computational convenience, and do not correspond to senses.

Let $\delta(y)$ be the immediate children of y in G (i.e., y's immediate hyponyms). Let $\Delta(y)$ be the subtree with root y. The random walk transitions from y to a child y' with probability proportional to the “mass” of the subtree at y':

$$P(y' \mid y) \propto \sum_{y'' \in \Delta(y')} P(y''), \quad \text{for } y' \in \delta(y). \qquad (2)$$

The mass of the subtree $\Delta(y')$ has the intuitive interpretation as the total probability of mentioning something that “is a” y'. The random walk then repeats downward from y'. For example, if y starts at “food,” it may go down to “dairy,” then to “cheese,” and finally to the leaf “Swiss cheese.” As “Swiss cheese” is a leaf-level sense in G, the next downward step is to an absorbing node and the walk terminates.

The probability $P(y_x \mid y)$, that starting from y the walk is absorbed below leaf y_x, can be easily computed in closed form (Doyle and Snell 2000) once the transition probabilities in (2) are known. Order rows and columns in a transition matrix such that the non-absorbing senses follow the absorbing nodes:

$$T = \begin{pmatrix} I & 0 \\ R & Q \end{pmatrix}. \qquad (3)$$

$R_{i,j}$ is thus the transition probability from sense $y_i \in V$ to absorbing node a_j (nonzero only when y_i the leaf to which a_j is attached), and $Q_{i,k}$ the transition probability from y_i to y_k, both $\in V$. The absorption probabilities B, with $B_{i,j}$ the probability of a walk beginning at sense y_i being absorbed by node a_j, are obtained by $B = (I - Q)^{-1}R$.

With absorbing node a_j attached to leaf y_x, we then take $P(y_x \mid y_i) = B_{i,j}$.

Once the walk terminates, the image x is generated uniformly from the available images at the leaf. Call $c_{\text{img}}(y)$ the number of images available at leaf y. As we have assumed that, from any image x, the generating leaf can be unambiguously determined, $P(x \mid y)$ is the probability that a random walk beginning at y will terminate at leaf y_x, and that image x will be generated by that leaf:

$$P(x \mid y) = P(y_x \mid y)\frac{1}{c_{\text{img}}(y_x)}. \qquad (4)$$

It is easy to show that any c_{img} is canceled out in (1), resulting in the final formulation of the random walk model:

$$P_{\text{RW}}(y|\mathbf{x}_{1\ldots n}) = \frac{P(y)\prod_{i=1}^{n} P(y_{x_i} \mid y)}{\sum_{y' \in V} P(y') \prod_{i=1}^{n} P(y_{x_i} \mid y')}. \tag{5}$$

Because the random walk process as described above only allows downward movement (generating only images consistent with the sense), it cannot model mistakes or confusion between senses. To correct this, we include a fixed probability p_t of teleporting at each step of the random walk. If teleporting occurs, a sense $y \in V$ is selected as the destination with probability proportional to $P(y)^{\tau}$. If teleporting does not occur, a sense is selected using the probabilities defined in (2). Lastly, to account for out-of-vocabulary concepts, any walk beginning at $\langle other \rangle$ will teleport as its first step with probability 1. Priors over in-vocabulary senses are normalized to accommodate $P(\langle other \rangle)$; p_t, τ and $P(\langle other \rangle)$ thus constitute our random walk model's tunable parameters. The solution for $P(y_x \mid y)$ given above holds so long as the transition matrix (3) includes the teleportation probabilities.

A Baseline Model for Comparison

For comparison we investigate an alternative model, based on the LCA hypothesis described earlier. Under the lowest common ancestor model, once a human user has assigned all images $\mathbf{x}_{1\ldots n}$ to leaf-level senses, the sense LCA($\mathbf{x}_{1\ldots n}$) is the perceived sense.

The probabilistic interpretation of LCA can be written as $P_{\text{LCA}}(y \mid \mathbf{x}_{1\ldots n}) = 1$ if $y = \text{LCA}(\mathbf{x}_{1\ldots n})$; 0 otherwise. Because it assigns a probability of zero to most concepts in G, it cannot account for sense ambiguities. To compensate, we perform smoothing by interpolating the LCA model with three simple models. The "prior model" uses the prior defined above: $P_{\text{prior}}(y \mid \mathbf{x}_{1\ldots n}) = P(y)$, and the "uniform model" is a uniform distribution over V: $P_{\text{unif}}(y \mid \mathbf{x}_{1\ldots n}) = 1/|V|$. All these models assign zero probability to $y = \langle other \rangle$. Finally, the "other model" captures out-of-vocabulary senses: $P_{\text{other}}(y \mid \mathbf{x}_{1\ldots n}) = 1$, if $y = \langle other \rangle$; 0 otherwise. The final smoothed LCA (SLCA) model is

$$P_{\text{SLCA}}(y \mid \mathbf{x}_{1\ldots n}) = \sum_{j} \lambda_j P_j(y \mid \mathbf{x}_{1\ldots n}) \tag{6}$$

where $j \in \{\text{LCA, prior, unif, other}\}$, and the interpolation weights λ's are non-negative and sum to 1. $\lambda_{\text{LCA}}, \lambda_{\text{prior}}$ and λ_{unif} thus constitute the tunable parameters of the SLCA model; λ_{other} is fully determined by these three parameters.

Model Behavior

We examine our models by considering their aggregated predictions over the experimental image sets, in a format analogous to Table 1. Given a set of examples $\mathbf{x}_{1\ldots n}$, both the random walk and SLCA models produce a distribution over senses. We use a Gibbs classifier that produces a random output sense according to this distribution (as opposed to the Bayes classifier that always outputs the sense with maximum probability). This corresponds to the well-known Luce choice rule, a model of human choice probabilities (Luce 1963). We then aggregate the output senses. We fit both models' parameters by minimizing the sum-squared-difference between model prediction and the first cell of both sections of Table 1. The parameters found are given in Table 2.

Table 2: Parameter values

Random walk		Smoothed LCA	
p_t	0.37	λ_{LCA}	0.56
τ	2.0	λ_{prior}	0.44
$P(\langle other \rangle)$	0	λ_{unif}	0

Using these parameters, the results for the random walk model are given in Table 3; results for SLCA are given in Table 4. It is clear from these results that both RW and SLCA can fit observation (1): sense ambiguity is common when no redundant images are presented.

Table 3: Percentages of LCA, generalized, and inconsistent responses as predicted by the RW model.

Images representing a leaf sense			
Images shown	LCA	Generalized	Inconsistent
$\mathbf{x}_1$	40.6	35.1	24.3
$\mathbf{x}_{1\ldots2}$	75.4	21.6	3.0
$\mathbf{x}_{1\ldots3}$	92.0	7.5	0.5
$\mathbf{x}_{1\ldots4}$	96.8	3.1	0.1
Images representing an internal sense			
Images shown	LCA	Generalized	Inconsistent
$\mathbf{x}_{1\ldots2}$	58.5	10.3	31.3
$\mathbf{x}_{1\ldots3}$	75.9	7.6	16.6
$\mathbf{x}_{1\ldots4}$	85.9	5.0	9.2

Table 4: Percentages of LCA, generalized, and inconsistent responses as predicted by the SLCA model.

Images representing a leaf sense			
Images shown	LCA	Generalized	Inconsistent
$\mathbf{x}_1$	55.9	5.1	39.0
$\mathbf{x}_{1\ldots2}$	55.9	5.1	39.0
$\mathbf{x}_{1\ldots3}$	55.9	5.1	39.0
$\mathbf{x}_{1\ldots4}$	55.9	5.1	39.0
Images representing an internal sense			
Images shown	LCA	Generalized	Inconsistent
$\mathbf{x}_{1\ldots2}$	56.4	4.3	39.3
$\mathbf{x}_{1\ldots3}$	56.4	4.3	39.3
$\mathbf{x}_{1\ldots4}$	56.4	4.3	39.3

Only the RW model predictions, however, conform to observation (2): adding a redundant image increases the posterior likelihood (and thus the response proportion) of the LCA. The reasons for this are explored in the Discussion section below. The SLCA model, by contrast, ignores redundant images (as did the original LCA hypothesis) and so cannot explain the observed effect.

Lastly, observation (3) is not explained by either model. For the responses collected in our experiment, redundant images after the first had greatly diminishing returns. The RW model predicts that LCA responses move asymptotically towards 100% as more redundant images are included; the

SLCA model does not predict any effect for redundant images.

Discussion

We have shown that the random walk model of sense ambiguity explains two of three observations of human behavior. It also provides an explanation for the psychological phenomenon of *basic-level senses*. Basic-level senses are a set of senses that are common, located in the middle of a sense hierarchy, and learned relatively earlier in life (Rosch et al. 1976). For example, "beans" is a basic-level sense, while its hyponyms "black beans," "kidney beans," "pinto beans," etc., are not. Similarly, "lettuce" is a basic-level sense, while its hyponyms "iceberg lettuce," "romaine lettuce," etc., are not. The generalized responses collected in our experiment show the effect of basic-level bias: we notice that when presented with an image from a non-basic-level leaf sense (e.g., "romaine lettuce"), people tend to generalize to its basic-level hypernym (e.g., "lettuce"). This is consistent with the psychology literature.

The random walk model predicts a bias towards generalized responses with high prior probability. By the nature of the random walk, the closer a sense y is to an image x's associated leaf y_x, the higher the likelihood $P(x|y)$, maximized at $y = y_x$. However, the *posterior probability* of a hypernym y given x will still be greater if $P(y) \gg P(y_x)$. As sense priors are determined by sense frequency, this is exactly the nature of basic-level senses—common, interior senses preferred as explanations over their hyponyms.

On the other hand, as more *redundant* images are introduced, the likelihood of the images becomes the dominant factor in the posterior. This manifests in a tendency for the posterior probability assigned to the LCA by the random walk model to approach 1 as more images are included. The assumption that sense ambiguity is minimized by displaying many highly varied example images is used by (Li et al. 2008) in the design of a system for word-representation as image sets; our random walk model provides a formal explanation.

For an example of this tendency, see Table 5, regarding user responses and model predictions for the image set shown in Figure 1. The RW model successfully predicts that initially, the basic-level response "cheese" dominates when there is only one swiss cheese image. The RW model also qualitatively predicts the trend away from basic-level and toward more specific responses as more swiss cheese images are given.

On this example, the RW model is quantitatively over-specializing. When all four images are displayed, the RW model assigns "Swiss cheese" a probability of 0.94; however, only 28.5% of experimental subjects responded with that sense. This discrepancy can perhaps in part be attributed to poor prior estimates, which influences the speed of specializing. If the true prior for "Swiss cheese" were very low, the slow response specialization in the human experimental results would have be predicted.

Perhaps more importantly, we assume the sense hierarchy G is universal. Our experimental participants were university students and almost certainly possessed differing degrees of culinary knowledge. It is entirely possible that some portion of the subjects simply could not recognize certain leaf senses in G as distinct from their immediate hypernym: no participants gave a response of "iceberg lettuce," for example, even when presented with four images of it, while 82.1% gave the more general response "lettuce." The parameters of the random walk model cannot be adjusted to account for this; no matter how G and the transition probabilities are altered, the posterior of some unique sense in V will always approach 1 as more redundant examples are shown.

Table 5: RW model probabilities and subject response counts for the Swiss cheese images shown in Figure 1.

images		food	dairy	cheese	Swiss
$\mathbf{x}_1$	$P(y\|\mathbf{x})$	0.08	0.05	0.54	0.03
	#humans	0	0	27	2
$\mathbf{x}_{1\ldots2}$	$P(y\|\mathbf{x})$	0.02	0.01	0.72	0.27
	#humans	0	0	23	4
$\mathbf{x}_{1\ldots3}$	$P(y\|\mathbf{x})$	0.00	0.00	0.28	0.72
	#humans	0	0	25	3
$\mathbf{x}_{1\ldots4}$	$P(y\|\mathbf{x})$	0.00	0.00	0.06	0.94
	#humans	0	0	20	8

Informal examination of the experimental responses showed two common types of inconsistencies. Some responses were inconsistent with a small number of example images, but consistent with the rest. This usually occurred when an image was somewhat vague, or the prior of the associated leaf sense was low. As an example, consider the top row of images shown in Figure 3. Although the three images showing poultry are easily recognizable as such, the image of beef shows several cuts of meat that cannot be trivially identified as belonging to a particular type of animal. Only 39.3% of responses named the LCA "meat"; the rest named senses within the subtree rooted at "poultry." For 60.7% of subjects, either the image of beef was not comprehensible, or it was regarded as irrelevant.

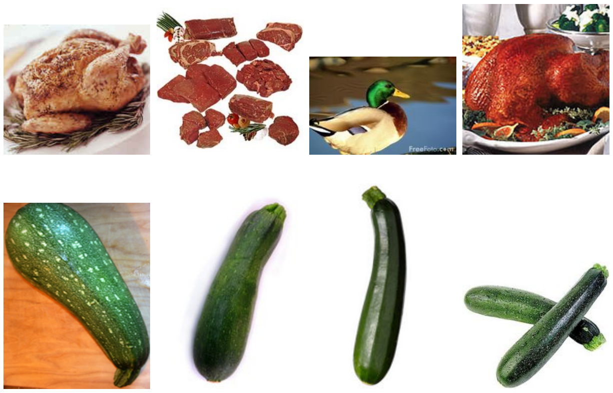

Figure 3: Two image sets for which a majority of responses were inconsistent. Above: "meat." Below: "zucchini."

A second type of inconsistency involved visual confusion between concepts. The bottom row of images in Figure 3 shows four images of zucchini. 32.1% responses named this leaf sense, while 57.1% misidentified the sense as “cucumber.” The mistake is understandable, as zucchinis and cucumbers are visually similar. Context, including visual similarity, is ignored when teleporting in the random walk model, and so neither type of inconsistency is specifically predicted.

Opportunities for Future Work

The random walk model construction and the data collected in our experiment help provide insight into human behavior and perceived sense ambiguity. The model’s deficiencies in fully explaining observed behavior are largely attributable to the assumptions and estimations used. Relaxing these assumptions, and improving estimation, will form the basis of future work.

We assume unambiguous assignment of images to leaves. Images containing a clear and unambiguous example of exactly one object type are rare, forcing careful selection before images are displayed. Image-to-leaf uncertainty can be handled by allowing multiple leaves to contribute nonzero probability to (4) (this would require estimating $c_{\text{img}}(y)$, a step the current formulation makes unnecessary).

Sense frequencies c_{text} provide the basis for the sense prior and transition probabilities (2), and in turn result in basic-level behavior matching that known to occur in humans. Because these frequencies are so essential for the model, they should be estimated as accurately as possible; exploring other techniques for this estimation is an obvious direction for future work. Existing research regarding basic-level concepts could allow priors to be based more directly on human cognition.

We use a predetermined, fixed sense hierarchy G for all subjects. In the spirit of Bayesian modeling, the graph G itself can be made a random variable with an appropriate prior, so that different subjects may have different instances of the sense hierarchy to match their preconceptions. A less obvious advantage of treating G as a random variable is to bound the posterior of a sense by the likelihood that the random variable G contains that sense, intuitively corresponding to the notion that no number of examples will prompt a sense response unknown to the user.

Visual confusion between senses is unexplained by our model, but there are two obvious approaches to its inclusion. Firstly, visual similarity between leaf senses could be considered as image-to-leaf uncertainty, as described above. Secondly, the random walk process could be modified to make teleporting between visually similar concepts more likely. Either approach requires a measure of visual similarity. Similarity can be empirically estimated using image sets corresponding to the senses in V; alternatively, unsupervised techniques such as that in (Sivic et al. 2008) allow the discovery of visual object hierarchies without predetermined interior senses. The resulting hierarchy can be included as a possible value for random variable G, allowing visual information to influence the sense posterior without changing the particulars of the random walk itself.

Concluding Remarks

We have proposed a Bayesian model for human sense perception, which predicts qualitative features of collected human data and human cognition, including the well-known basic-level effect. Bayesian inference over a sense hierarchy has been applied in recent psychology research (Xu and Tenenbaum 2007); our model presents novel formulations for the sense prior $P(y)$ and likelihood $P(\mathbf{x}_{1\ldots n} \mid y)$.

Our experimental data demonstrates the difficulty of disambiguation by example. Effective pictorial communication systems require unambiguous picture generation: a model of sense ambiguity as perceived by the user is vital to the success of such systems. This work demonstrates a model that captures many aspects of sense ambiguity and provides several opportunities for further development.

Acknowledgment This material is based upon work supported by the National Science Foundation under Grant No. 0711887, and by the Wisconsin Alumni Research Foundation. We thank Josh Tenenbaum for helpful discussions.

References

[Coyne and Sproat 2001] Coyne, B., and Sproat, R. 2001. WordsEye: An automatic text-to-scene conversion system. In *Proc. SIGGRAPH 2001*, 487–496.

[Doyle and Snell 2000] Doyle, P. G., and Snell, J. L. 2000. Random walks and electric networks.

[Fellbaum 1998] Fellbaum, C., ed. 1998. *Wordnet: An Electronic Lexical Database*. Bradford Books.

[Johansson et al. 2005] Johansson, R.; Berglund, A.; Danielsson, M.; and Nugues, P. 2005. Automatic text-to-scene conversion in the traffic accident domain. In *Proc. 19th Int. Joint Conf. Artificial Intelligence*, 1073–1078.

[Joshi, Wang, and Li 2006] Joshi, D.; Wang, J. Z.; and Li, J. 2006. The story picturing engine – A system for automatic text illustration. *ACM Trans. Multimedia Computing, Communications and Applications* 2(1):68 – 89.

[Li et al. 2008] Li, H.; Tang, J.; Li, G.; and Chua, T.-S. 2008. Word2image: towards visual interpreting of words. In *Proceeding of the 16th ACM international conference on Multimedia*, 813–816. ACM.

[Luce 1963] Luce, R. D. 1963. Detection and recognition. In Luce, R. D.; Bush, R. R.; and Galanter, E., eds., *Handbook of Mathematical Psychology*, volume 1. New York and London: John Wiley and Sons, Inc. 103–190.

[Rosch et al. 1976] Rosch, E.; Mervis, C. B.; Gray, W. D.; Johnson, D. M.; and Braem, P. B. 1976. Basic objects in natural categories. *Cognitive Psychology* 8(3):382–439.

[Sivic et al. 2008] Sivic, J.; Russell, B. C.; Zisserman, A.; Freeman, W. T.; and Efros, A. A. 2008. Unsupervised discovery of visual object class hierarchies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[Xu and Tenenbaum 2007] Xu, F., and Tenenbaum, J. B. 2007. Word learning as Bayesian inference. *Psychological review* 114(2).

[Zhu et al. 2007] Zhu, X.; Goldberg, A.; Eldawy, M.; Dyer, C.; and Strock, B. 2007. A Text-to-Picture synthesis system for augmenting communication. In *The 22nd AAAI Conference on Artificial Intelligence*.