Incremental Models of Natural Language Category Acquisition

Trevor Fountain (t.fountain@sms.ed.ac.uk)
Mirella Lapata (mlap@inf.ed.ac.uk)
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB, UK

Abstract

Learning categories from examples is a fundamental problem faced by the human cognitive system, and a long-standing topic of investigation in psychology. In this work we focus on the acquisition of natural language categories and examine how the statistics of the linguistic environment influence category formation. We present two incremental models of category acquisition — one probabilistic, one graph-based — which encode different assumptions about how concepts are represented (i.e., as a set of topics or nodes in a graph). Evaluation against gold-standard clusters and human performance in a category acquisition task suggests that the graph-based approach is better suited at modeling the acquisition of natural language categories.

Introduction

The task of *categorization*, in which people cluster stimuli into categories and then use those categories to make inferences about novel stimuli, has long been a core problem within cognitive science. Understanding the mechanisms involved in categorization is essential, as the ability to generalize from experience underlies a variety of common mental tasks, including perception, learning, and the use of language. As a result, category learning has been one of the most extensively studied aspects in human cognition, with computational models that range from strict *prototypes* (categories are represented by a single idealized member which embodies their core properties; e.g., Reed 1972) to full exemplar models (categories are represented by a list of previously encountered members; e.g., Nosofsky 1988) or combinations of the two (e.g., Griffiths et al. 2007a).

Historically, the stimuli involved in such studies tend to be either concrete objects with an unbounded number of features (e.g., physical objects; Bornstein and Mash 2010) or highly abstract, with a small number of manually specified features (e.g, binary strings, colored shapes; Medin and Schaffer 1978, Kruschke 1993). Furthermore, most existing models focus on adult categorization, i.e., it is assumed that a large number of categories have already been learned. A notable exception is Anderson's (1991) rational model of categorization (see also Griffiths et al. 2007a) where it is assumed that the learner starts without any predefined categories and stimuli are clustered into groups as they come along. When a new stimulus is observed, it can either be assigned to one of the pre-existing clusters, or to a new cluster of its own.

In this work, we concentrate on the task of acquiring natural language (semantic) categories and examine how the statistics of the linguistic environment as approximated by large corpora influences category learning. Evidently, categories are learned not only from exposure to the linguistic environment but also from our interaction with the physical world. Perhaps unsurprisingly, words that refer to concrete entities and actions are among the first words being learned as these are directly observable in the environment (Bornstein et al. 2004). Experimental evidence also shows that children respond to categories on the basis of visual features, e.g., they generalize object names to new objects often on the basis of similarity in shape and texture (Landau et al. 1998, Jones et al. 1991). Nevertheless, we focus on the acquisition of semantic categories from large text corpora based on the hypothesis that simple co-occurrence statistics can be used to capture word meaning quantitatively. The corpus-based approach is attractive for modeling the development of linguistic categories. If simple distributional information really does form the basis of a word's cognitive representation, this implies that learners are sensitive to the structure of the environment during language development. As experience with a word accumulates, more information about its contexts of use becomes encoded, with a corresponding increase in the ability of the language learner to use the word appropriately and make inferences about novel words of the same category.

The process of learning semantic categories is necessarily incremental. Human language acquisition is bounded by memory and procecessing limitations, and it is implausible that children process large amounts of linguistic input at once and induce an optimal set of categories. An incremental model learns as it is applied, meaning it does not require separate training and testing phases. Behavioral evidence (Bornstein and Mash 2010) suggests that this scenario more closely mirrors the process by which infants acquire categories. Having this in mind, we formulate two incremental categorization models, each differing in the way they represent categories. Both models follow the exemplar tradition — categories are denoted by a list of stored exemplars and inclusion of an unknown item in a category is determined by some notion of similarity between the item and the category exemplars. Previous work (Voorspoels et al. 2008, Storms et al. 2000, Fountain and Lapata 2010) indicates that exemplar models perform consistently better across a broad range of natural language

Algorithm 1: Batch Chinese Whispers

```
initialize;
for node_i ∈ Nodes do
    class (node) = i;
end
while changes do
    for node ∈ Nodes (in random order) do
        class (target) = class (nearest neighbor)
    end
end
```

Algorithm 2: Incremental Chinese Whispers

```
initialize;
for node_i ∈ Nodes do
    class (node) = i
end
for target,context ∈ Documents do
    update target representation given context;
    class (target) = nearest neighbor
end
```

categorization tasks. This finding is also in line with studies involving artificial stimuli (e.g., Nosofsky 1988). While these studies focus on natural language categories they tend not to specifically address the task of language acquisition; Storms et al. (2000) compare various categorization models in a natural language context, Voorspoels et al. (2008) model typicality ratings for natural language concepts, and Fountain and Lapata (2010) explore a number of corpus-based representations for linguistic exemplars.

Our first model is reminiscent of semantic networks (Collins and Loftus 1975). In this framework, concepts are represented as nodes in a graph and edges represent relationships between such concepts. Although semantic networks are traditionally hand-coded by modelers, we learn them from naturally occurring data. In our model, nodes in the graph correspond to words and weighted edges indicate distributional similarity rather than semantic or syntactic relationships. Categories arise naturally in such a representation as densely connected regions or subgraphs. While most research on semantic networks focuses on their use within a larger model of spreading activation (Anderson 1983), they have also been used to gain insight into performance deficits in patients with psychological impairments (Tyler et al. 2000) and to draw comparisons between internet search and memory access (Griffiths et al. 2007b). Our second model follows a probabilistic approach where categories correspond to topics in a generative model (Griffiths et al. 2007c). Topics themselves are modeled as probability distributions over words, and can be thought of as a "soft" list of exemplars belonging to that category. In order to obtain a hard clustering of words into categories we need only compute the most likely category for each word. Topic models have been successful at modeling a wide range of cognitive phenomenal including lexical priming, word association, synonym selection, and reading times (see Griffiths et al. 2007c).

Category Acquisition Models

Any model of human category acquisition should demonstrate two important features: (1) the input should be processed as it arrives, i.e., the set of clusters is incrementally updated and (2) the set of clusters should not be fixed in advance, but rather determined by the characteristics of the input data. In what follows, we present two models that satisfy both constraints.

Semantic Networks The standard conception of a semantic network is a graph with edges between word nodes. Such a graph is *unipartite*: there is only one type of node, and those nodes can be interconnected freely. While traditional research using semantic networks has focused on performing inference using fully-formed networks, we are argue that they are also well suited to modeling acquisition, as updating the graph to reflect newly acquired information is a straightforward procedure. Furthermore, meaningful categories can be extracted from such a representation by identifying well-structured subgraphs within the network.

The task of extracting such subgraphs is generally viewed as a graph clustering problem; Chinese Whispers (CW, Biemann 2006) is one such randomized graph-clustering algorithm that takes as input a graph with weighted edges and produces a hard clustering over the nodes in the graph. It has several desirable properties, including a tendency to converge rapidly and the ability to infer the number of output clusters. The CW algorithm consists of two steps: initialization and iteration. In the initialization step, each node in the graph is assigned a unique class. In the iterative step, each node in the graph (in random order) adopts the highest ranked class in its neighborhood (i.e., the set of nodes with which it shares an edge). Algorithm 1 shows this procedure in pseudocode. CW is in general not guaranteed to converge; in particular, a node with two equally-distant nearest neighbors may flip between the classes of those neighbors indefinitely. In practice, however, it tends to reach 'almost-convergence' quite rapidly.

Vanilla CW requires that the entire graph be known before it can be applied, and thus makes no provision for graphs which change over time, as would be expected in an acquisition task. Modifying the algorithm for use in an incremental setting is straightforward: we need only to update the edges of the graph with newly-encountered input before each iteration and to run the algorithm until there is no more input to process rather than until convergence (see Algorithm 2).

While applying the incremental CW algorithm to the task of acquiring semantic categories from text, we maintain a weighted, undirected graph in which each node represents a target word and edges between nodes are weighted according to the similarity between words. To compute this similarity, the implementation maintains a running co-occurrence matrix in which each row corresponds to a target word and each column to a possible context word. Similarity between words is computed as the cosine distance between the corresponding rows. Matrix cells are transformed into (positive) pointwise mutual information values (Bullinaria and Levy 2007). Our experiments used a context window centered around a target word, however non-symmetric contexts are also possible; target representations are updated according to the context words appearing in the window.

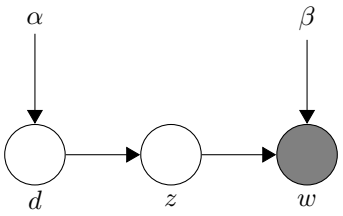

Figure 1: The Latent Dirichlet Allocation model (Griffiths et al. 2007c). d is the distribution of topics within a single document; z is the distribution over observable words w for a topic. α and β function as smoothing parameters for d and w, respectively.

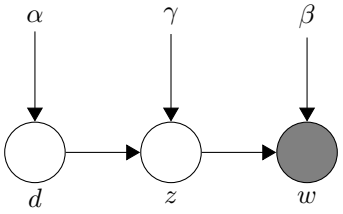

Figure 2: A nonparametric topic model which infers the number of topics during training. γ indicates the amount of probability mass reserved for unseen categories (analogous to Anderson's (1990) coupling probability.

Topic Models A great deal of work in recent years has focused on the idea of topic models, in which the meaning of a particular document or word is encapsulated by the latent topics it contains or from which it is generated. Conceptually such models seem appropriate for categorization tasks, as the notions of "topic" and "category" have much in common.

One particular topic model which has seen wide success is Latent Dirichlet Allocation (LDA, Blei et al. 2003, Griffiths et al. 2007c), which provides a probabilistic model of document generation. In LDA, a document is modeled as a probability distribution over a set of latent topics; similarly, a topic is modeled as a distribution over words. The actual words composing a document are supposed to have been generated by a process of repeatedly sampling first a topic from the document distribution, then a single word from the selected topic. LDA (and generally topic models) can be viewed as a form of a *bipartite* graph consisting of two types of nodes, i.e., words and topics and connections between them.

One drawback to LDA is that it requires the number of topics to be known in advance. As this assumption clearly does not hold in the case of category acquisition, we developed a nonparametric, incremental topic model which is similar in spirit to LDA. This model maintains the generative assumptions of LDA, and much of the same graphical structure; it differs in the addition of a coupling probability (Anderson 1990) used to infer the number of categories during training. Additionally, it performs no final re-estimation of probabilities (as in standard LDA) in order to maintain incrementality.

In terms of graphical structure our topic model differs from standard LDA (Figure 1) by the addition of a third parameter, γ, on the topic distribution. The γ parameter indicates the proportion of probability mass to reserve for a new, previously unseen topic; as additional topics are created the probability of assigning a word to a new topic decreases in relation to γ, α, and β act as invisible counts for each topic in a document and each word in a topic, respectively. Combining these parameters with the graphical model in Figure 2 yields the following probabilistic model:

$$
\begin{aligned}
P(w|z) &= \frac{\eta_w^z+\beta}{\sum_x^W(\eta_x^z+\beta)} \\
P(z|d) &= \frac{(\eta_z^d+\alpha+|W|\beta)(1-\gamma)}{(\sum_y^Z(\eta_y^d+\alpha+|W|\beta)(1-\gamma))+(\alpha+|W|\beta)\gamma} \\
P(z'|d) &= \frac{(\alpha+|W|\beta)\gamma}{(\sum_y^Z(\eta_y^d+\alpha+|W|\beta)(1-\gamma))+(\alpha+|W|\beta)\gamma} \\
P(d) &= \frac{\sum_y^{Z+z'}(\eta_y^d+\alpha)}{Z+z'} \\
&\quad \sum_e^D \sum_y (\eta_y^e+\alpha)
\end{aligned}
$$

where w, z and d represent a word, topic (category), or document, respectively. z' represents a previously unseen topic; a word w assigned to z' is instead assigned to a newly created category initialized to a uniform distribution. The notation η_w^z signifies the number of times word w has appeared in topic z, while η_z^d similarly indicates the count of occurrences of z within document d.

To maintain incrementality, the model performs no re-estimation of probabilities; instead, as each item w of input is encountered it is assigned to a sampled topic z. The relevant document and topic distributions are then updated in accordance with the sampled topic. While these individual predictions are not revised (as in LDA) by subsequent re-samplings, predicted topics for subsequent encounters of w change based on the distribution of words and topics; the equations for $P(w|z)$ and $P(z|d)$ are thus analogous to those used during Gibbs sampling in LDA. With additional documents these distributions converge to (hopefully) meaningful topics.

Experiment 1

Our first goal was to compare our two categorization models and establish their performance on a large corpus. To do this, we trained both on the British National Corpus (BNC) and compared each model's resulting clustering against a human-produced gold standard. In the following we describe how this gold standard was created, discuss how the model parameters were estimated, and explain how the model output was evaluated.

Method In order to train our models, the BNC was preprocessed so as to remove stopwords and highly infrequent words. Target words corresponded to frequently-used nouns, however this is not a limitation of our models which could be also applied to verbs or adjectives. The topic model has three free parameters, i.e., α (the prior observation count for the number of times a topic is sampled in a document), β (the prior observation count on the number of times words are

REPTILE
salamander, iguana, frog, alligator, rattlesnake, tortoise, crocodile, turtle, toad
FURNITURE
chair, stool, rocker, sofa, cabinet, desk, bookcase, mirror, shelves, bed, drapes, clock, table, bathtub, bureau, cupboard, dresser, fence, cushion, bench, bayonet, armour
FRUIT
peach, yam, nectarine, banana, cantaloupe, apple, plum, raspberry, pear, grape, blueberry, raisin, pineapple, prune, rhubarb, strawberry, lemon, honeydew, orange, tomato, lime, cherry, coconut, olive, grapefruit, tangerine, avocado, pumpkin, cranberry, mandarin

Table 1: Example gold standard categories with their exemplars from Fountain and Lapata (2010).

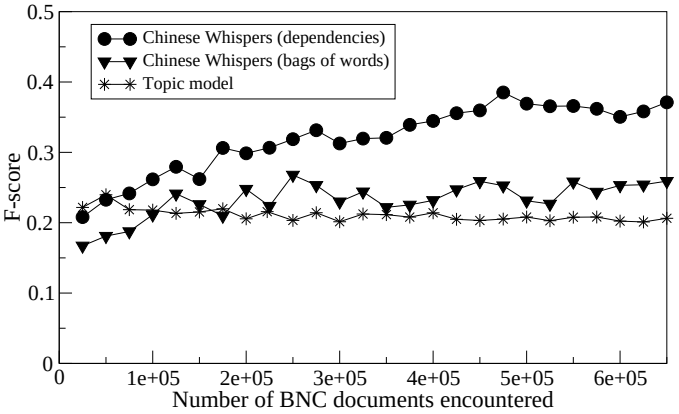

Figure 3: Performance of the topic model and Chinese Whispers using dependencies and a bag of words context window.

sampled from a topic), and γ (the probability mass reserved for new topics). For α and β we chose values in accordance with the literature on LDA (Teh et al. 2006); these parameters were set to 1.2 and 0.1, respectively. The γ parameter was tuned on a development corpus (10% of the BNC), with the final value of 0.10. Because of this tuning procedure, all scores reported are from application on the remaining 90% of the BNC not used for development.

Note that the output of the topic model is a set of probability distributions rather than a hard clustering over words. We can nevertheless coerce the model to produce such a clustering by assigning each word to the category (topic) which maximizes its likelihood:

$$category(w) = \operatorname*{argmax}_z P(z|w) \quad (1)$$

The incremental CW model was trained on noun-centered context windows of ± 5, which were extracted from the BNC. As the output of CW is a hard clustering over nodes in the graph, no additional post-processing is required. One obvious question that arises in the context of this experiment is whether using a richer contextual representation yields more accurate categories; we examined this hypothesis by applying the incremental CW algorithm to a dependency-parsed version of the BNC.[1] Specifically, we obtained dependency information from the output of MINIPAR, a broad coverage parser (Lin 2001). To minimize noise, this output was restricted to a small set of lexicalized dependency relations: subject, object, and conjunction.

Both models were evaluated based on their clustering of words into semantic categories and their output was compared against similar clusters elicited from human participants. In particular, we used the data from Fountain and Lapata's (2010) category naming study as a gold standard.[2] The aim of their experiment was to augument McRae et al.'s (2005) semantic feature norms with category information. These norms consist of 541 basic-level concepts (e.g., DOG and CHAIR) with features collected in multiple studies over several years. Fountain and Lapata obtained category labels for 517 of these concepts. They presented participants with a number of nouns chosen at random from McRae et al.'s norms, and asked them to name the category to which each noun belonged. Participants responded in freeform strings, i.e., they were not provided with a list of possible categories. After adjusting for differences in spelling and conflating synonyms, these responses were used to determine the most "correct" category label for each of the 517 nouns.

Because the norms were originally drawn from a limited number of concepts many of the nouns were labeled with the same category label; we were exploited this overlap in order to construct a clustering over McRae et al.'s norms in which each cluster corresponds to a subset of nouns assigned the same category label in Fountain and Lapata (2010). Overall, we obtained 32 categories averaging approximately 16 nouns apiece. Examples of the clusters used in our experiments are shown in Table 1.

Each model produced a clustering over the nouns taken from the McRae et al. norms which we compared against the human-produced gold standard clustering described above; to evaluate cluster quality we computed the F-score measure described in Agirre and Soroa (2007). Under their evaluation scheme, the gold standard is partitioned into a test and training corpus. The latter is used to derive a mapping of the induced clusters to the gold standard labels. This mapping is then used to calculate the system's F-score on the test corpus. We calculated F-score as the harmonic mean of precision and recall defined as the number of correct members of a cluster divided by the number of items in the cluster and the number of items in the gold-standard class, respectively.

Results CW and the topic model produced clusters for 517 nouns. As both models are non-parametric, they induce the number of clusters (i.e., categories) from the data as well as which nouns belong to these clusters. The topic model partitioned the target nouns into 167 clusters and CW into 35.

Compared to the gold-standard clustering, the topic model achieved an F-score of 0.179; CW obtained an F-score of 0.212 when using a bag of words context window. The

[1] Incorporating syntactic information into an incremental topic model is less straightforward, although extensions of the basic LDA model have been proposed that take syntax into account (e.g., Boyd-Graber and Blei 2009).

[2] Available from `http://bit.ly/categorization`.

The **fendle** is the very dense region consisting of nucleons (**daxs** and **tomas**) at the center of a **gazzer**. Almost all of the mass in a **gazzer** is made up from the **daxs** and **tomas** in the **fendle**, with a very small contribution from the orbiting **wugs**. The diameter of the **fendle** is in the range of 1.5fm (1.75×10-15m) for **tulver** to about 15fm for the heaviest **gazzers** such as **tupa**.

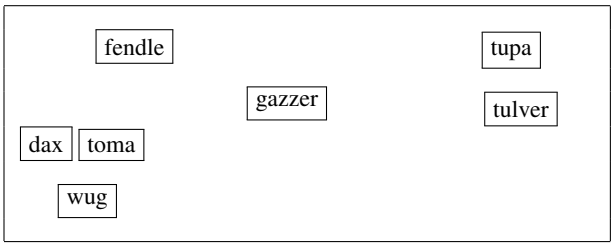

Figure 4: The incremental categorization task as seen by participants. Each trial consisted of a series of paragraphs from the same source document; the words to be clustered (shown in boldface) remained constant, with participants asked to update their clustering after each trial.

model's performance improved to an F-score of 0.371 when dependency relations were used. To put these numbers into perspective, we also implemented a baseline algorithm that groups nouns into clusters randomly, which achieved an inferior F-Score of 0.135. Overall, our results indicate that more fine-grained linguistic information beyond simple co-occurrence is beneficial for categorization. Figure 3 shows how performance on the category acquisition task varies over time (i.e., over the course of encountering all documents in the training set). As can be seen, the quality of clusters produced by CW increases with additional data, i.e., the algorithm's performance improves with more iterations.

Experiment 2

While the previous experiment explored how effectively the two models capture large-scale category information it did not assess the effect of incrementality. The difficulty in performing such an evaluation is that it requires a snapshot of category structure throughout the process of category acquisition. Getting such snapshots from children would be ideal, however a longitudinal study of category acquisition would be a major undertaking spanning several years. Getting such snapshots from adults is also problematic, as they clearly possess a great deal of world knowledge about the target words used in a hypothetical experiment. To rectify this, we conducted a study in which participants were given a series of paragraphs containing nonsense words and asked, after having read each paragraph, to group the nonsense words into categories. The hope was that the results from such a study would illuminate the kinds of interim categories the mind might construct when presented with minimal information about a set of novel stimuli.

Method Thirteen source documents were compiled from Wikipedia articles on various technical domains, including medicine, physics, biology, and mixology[3]. Each document consisted of 3–5 paragraphs, each containing between 4–6 sentences in which a small number of re-occurring content words were replaced with nonce words (nine on average per document). The study was completed by 250 participants, mostly undergraduates.

One serious concern in conducting a study like this is ensuring that participants do not actually perform a separate, but related task, instead determining the mapping between nonsense words and their meaningful equivalents. We mitigated this problem by extracting the text from highly technical documents, the subject matter of which would almost certainly be unfamiliar to participants and thus limiting the amount of world knowledge they could bring to bear. Also of concern was avoiding priming subjects with the number of categories; to avoid such influence, participants were asked to group target words into clusters by dragging items together on a virtual canvas, rather than by assigning labels or placing items into pre-specified bins. A snapshot of the experimental interface our participants saw is given in Figure 4.

The topic model and CW were trained on the same set of paragraphs, and the interim clustering produced after processing each document saved in order to investigate how well the models captured the interim categories formed during incremental learning. Note that both models were trained from a blank state, reflecting a lack of pre-existing world knowledge. Again, we used a bag-of-words representation for CW as the prevalence of nonsense words in the data resulted in many parsing mistakes. Following on Experiment 1, we then applied the topic model and CW to the same set of paragraphs and evaluated the resulting categories against those produced by participants, again using F-score (Agirre and Soroa 2007).

Results Firstly, we assessed how well our participants agreed on the category acquisition task.[4] We computed the F-score of a single participant's clustering for each trial as the average F-score between it and each of the other participants' clusterings for that trial; and then calculated the mean reliability as the average F-score of all trials for all participants. On the category acquisition experiment, participants achieved a mean reliability of 0.694. CW achieved a comparable F-score of 0.656, followed by the topic model with an F-score of 0.634. These F-scores were computed by a procedure similar to the human reliability described above. The model was treated as a single participant and the F-score for each stage was computed as the average F-score between the model's clustering in that stage and each participant's clustering, with the individual stage scores averaged to produce the final score. Figure 5 shows the F-scores achieved by the two models for each trial against the human upper bound. It is interesting to note that both models are close to human performance, with Chinese Whispers having mostly the lead over the topic model. Counterintuitively, performance of both models and human participants declines over time; this is primarily an effect of increasing disagreement between participants when exposed to additional observations.

[3]Molecular Mixology is the term applied to the process of creating cocktails using the scientific equipment and techniques of molecular gastronomy.

[4]Subject data for Experiment 3 is available from `http://bit.ly/categorization`.

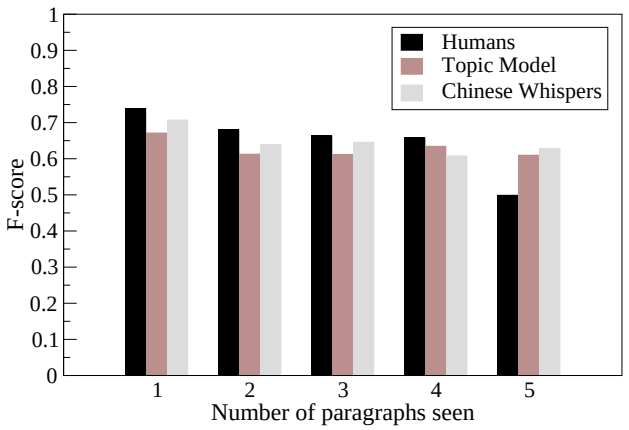

Figure 5: Model performance and human upperbound after each trial.

General Discussion

At first glance the scores on the large-scale task (Experiment 1) for both models appear quite low. Our aim in this first experiment, however, was merely to establish a comparison between the two approaches on a clustering task. This is challenging considering that the models are expected to assign 500+ words into an unspecified number of well-defined semantic categories from word co-occurrence information alone. Humans acquire semantic categories from a richer environment based on their sensorimotor experiences in addition to linguistic input.

Regardless, a strict comparison of results shows that CW outperformed the topic model on this large-scale category experiment. Manual inspection of the clusters output by the topic model suggests an explanation: the learned topics, while clearly capturing some notion of semantic relatedness between words, rarely correspond to the desired semantic categories. Instead they cut across categories, collating words that share a theme or context rather than words belonging to a common category. The clusters output by CW, conversely, capture more of the semantic category information but tend to do so at a higher level (e.g. conflating FRUIT, VEGETABLE, and FOOD into a single meta-category).

This is particularly interesting in light of the differences between the two models; CW is a simpler model, both in terms of the way it represents and forms categories. Recall that the algorithm creates a unipartite graph with one type of nodes (i.e., words) which can be interconnected freely. In the topic model, semantic information is organized in a bipartite graph consisting of words, topics, and their interconnections. This more structured representation does not seem appropriate for the category acquisition task. In particular, the notion of topic as it is used in the context of the topic model is not equivalent to that of a semantic category. The relative success of CW, combined with its simplicity and plausibility, suggests that such comparatively simple models can often provide a better approach for modeling low-level cognitive tasks, such as predicting category-specific deficits in patients with cognitive impairments (Tyler et al. 2000).

The results of the second experiment show that CW (and the topic model to a lesser extent) produce categories *incrementally* that are both meaningful and cognitively plausible. Interestingly, in this experiment the upper bound (i.e., inter-annotator agreement) is high despite the seeming difficulty of the task. This suggests that people are quite consistent in the types of categories they form even when those categories are based on only one or two pieces of information, and enforces the idea that, in the absence of real-world knowledge, people learn categories in an incremental fashion (Lamberts and Shapiro 2002).[5]

An important direction for future work is to model the hierarchical structure of categories. Inspection of the clusters produced in Experiment 2 reveals that participants tend to organize words into hierarchies rather than flat categories.

References

Agirre, E. and Soroa, A. (2007). Semeval-2007 task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 7–12, Prague, Czech Republic.

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, 22:261–295.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, 98:409–429.

Biemann, C. (2006). Chinese whispers - an efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of TextGraphs: the 1st Workshop on Graph Based Methods for Natural Language Processing*, pages 73–80, New York City.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Bornstein, M. H., Cote, L. R., Maital, S., Painter, K., Park, S.-Y., and Pascual, L. (2004). Cross-linguistic analysis of vocabulary in young children: Spanish, Dutch, French, Hebrew, Italian, Korean, and American English. *Child Development*, 75(4):1115–1139.

Bornstein, M. H. and Mash, C. (2010). Experience-based and on-line categorization of objects in early infancy. *Child Development*, 81(3):884–897.

Boyd-Graber, J. L. and Blei, D. (2009). Syntactic topic models. In *Advances in Neural Information Processing Systems 21*, pages 185–192.

Bullinaria, J. and Levy, J. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39:510–526.

Collins, A. M. and Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, 82(6):407–428.

Fountain, T. and Lapata, M. (2010). Meaning representation in natural language categorization. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, pages 1916–1921, Amsterdam, The Netherlands.

Griffiths, T. L., Canini, K. R., Sanborn, A. N., and Navarro, D. J. (2007a). Unifying rational models of categorization via the heriarchical Dirichlet process. In *Proceedings of the 29th Annual Conference of the Cognitive Science Society*, pages 323–328, Nashville, Tennessee.

Griffiths, T. L., Steyvers, M., and Firl, A. (2007b). Google and the mind: Predicting fluency with pagerank. *Psychological Science*, 18(12):1069–1076.

Griffiths, T. L., Tenenbaum, J. B., and Steyvers, M. (2007c). Topics in semantic representation. *Psychological Review*, 114:2007.

Jones, S. S., Smith, L. B., and Landau, B. (1991). Object properties and knowledge in early lexical learning. *Child Development*, (62):499–516.

Kruschke, J. K. (1993). Human category learning: Implications for backpropagation models. *Connection Science*, 5:3–36.

Lamberts, K. and Shapiro, L. (2002). Exemplar models and category-specific deficits. *Behavioral and Brain Sciences*, 24(3):484–485.

Landau, B., Smith, L., and Jones, S. (1998). Object perception and object naming in early development. *Trends in Cognitive Science*, 27:19–24.

Lin, D. (2001). LaTaT: Language and text analysis tools. In *Proceedings of the 1st Human Language Technology Conference*, pages 222–227, San Francisco, CA.

McRae, K., Cree, G. S., Seidenberg, M. S., and McNorgan, C. (2005). Semantic feature production norms for a large set of living and non-living things. *Behavioral Research Methods Instruments & Computers*, 37(4):547–559.

Medin, D. L. and Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, 85(3):207–238.

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14:700–708.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, 3(3):382–407.

Storms, G., Boeck, P. D., and Ruts, W. (2000). Prototype and exemplar-based information in natural language categories. *Journal of Memory and Language*, 42:51–73.

Teh, Y., Jordan, M., Beal, M., and Ble, D. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Tyler, L. K., Moss, H. E., Durrant-Peatfield, M. R., and Levy, J. P. (2000). Conceptual structure and the structure of concepts: A distributed account of category-specific deficits. *Brain and Language*, 75:195–231.

Voorspoels, W., Vanpaemel, W., and Storms, G. (2008). Exemplars and prototypes in natural language concepts: A typicality-based evaluation. *Psychonomic Bulletin & Review*, 15(3):630–637.

[5] While conceptually unsurprising, we nevertheless found this result somewhat unexpected given the number of complaints from participants regarding the difficulty of the task.